POLYTECHNIQUE MONTRÉAL

Exploring corporate websites to assess enterprise innovation – Measuring and validating the signal from textual content

MIKAËL HÉROUX-VAILLANCOURT

Département de mathématiques et de génie industriel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*

Génie industriel

Mai 2023

© Mikaël Héroux-Vaillancourt, 2023.

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Cette thèse intitulée :

Exploring corporate websites to assess enterprise innovation – Measuring and validating the signal from textual content

présenté par **Mikaël HÉROUX-VAILLANCOURT**

en vue de l'obtention du diplôme de *Philosophiæ Doctor*

a été dûment accepté par le jury d'examen constitué de :

Nathalie DE MARCELLIS-WARIN, présidente
Catherine BEAUDRY, membre et directrice de recherche
Pierre MOHNEN, membre externe
Philip SHAPIRA, membre externe

DEDICATION

À mes parents.

ACKNOWLEDGEMENTS

When studying innovation management, one quickly realises the undeniable contribution of the role of the environment as a success factor. I am extremely grateful to have been born into a developed society that allows anyone to achieve their full potential. I consider myself fortunate to have received the education to get to where I am today.

To my parents, thank you for stimulating my curiosity from a young age, for always pushing me to improve and most importantly, for always supporting and loving me unconditionally. Thank you to my family and friends for your encouragement and support.

I would also like to thank Catherine Beaudry for the trust she has placed in me throughout my young academic journey, in particular by giving me the opportunity to teach the course IND6117-Innovation Management. I am grateful for her unwavering support, her many sound advice, her innovative ideas and her exemplary rigour. Thank you for being such an inspiring model of excellence and determination.

Thanks also to Carl St-Pierre for his teachings, advice and suggestions that allowed us to make relevant and robust statistical analyses and ensure the credibility of our work. I am grateful for the many hours in his presence that were generously granted to me.

I could also not fail to mention the contribution of Constant Rietsch who was the first to explore corporate websites as a source of data during his master's degree and who generously continued his collaboration, notably for the collection of data that were required for the first paper presented entitled « *Using web content analysis to create innovation indicators—What do we really measure?* ».

I continue these acknowledgements with Davide Pulizzotto to whom I owe most of my knowledge in natural language processing as well as the data obtained for papers two and three presented in this thesis. I am particularly grateful for his generous contribution in creating an efficient and rigorous web mining process from The Wayback Machine, as well as for supporting me in processing and cleaning unstructured data.

I would also like to thank the entire team at The Evidence Network (TEN) for the trust and support you have given us in providing access to your survey data. As these data are used in the article two and three of this thesis, this input has been critical to the completion of much of the work presented

in this thesis. Special thanks to Margaret Dalziel for initiating this collaboration and I have also benefited from her support in writing articles two and three.

I would also like to thank all the Polytechnique employees who facilitate our work every day. Finally, I have a special thought for my colleagues of the Chair in Management and Economics of Innovation at Polytechnique. Thank you for your suggestions, ideas, advice and support. For those who have not yet done it, your turn will come soon.

REMERCIEMENTS

En étudiant la gestion de l'innovation, on se rend rapidement compte de l'apport indéniable du rôle de l'environnement comme facteur de succès. Je suis extrêmement reconnaissant d'avoir eu la chance d'être né dans une société développée permettant à quiconque de se réaliser pleinement. Je me considère choyé d'avoir reçu l'éducation me permettant d'en arriver où j'en suis aujourd'hui.

À mes parents, merci d'avoir stimuler ma curiosité dès le plus jeune âge, de m'avoir toujours poussé à m'améliorer et surtout et plus important, de m'avoir toujours supporté et aimé inconditionnellement. Merci à mes proches pour vos encouragements et votre support toujours bien senti.

Je remercie également Catherine Beaudry pour la confiance qu'elle m'a accordée tout le long de mon parcours académique, notamment en me donnant l'opportunité d'enseigner le cours IND6117- Gestion de l'innovation. Je suis reconnaissant pour son support inébranlable, ses nombreux conseils toujours judicieux, ses idées novatrices et sa rigueur exemplaire. Merci d'être un modèle aussi inspirant d'excellence et de détermination.

Merci également à Carl St-Pierre pour ses enseignements, ses conseils et ses suggestions qui ont permis de faire des analyses statistiques pertinentes, robustes et qui assurent la crédibilité de nos travaux. Je suis reconnaissant pour les nombreuses heures en sa présence qui m'ont été généreusement accordées.

Je ne pourrais également passer sous silence la contribution de Constant Rietsch qui a été le premier porteur de l'exploration des sites internet corporatifs comme source de données lors de sa maitrise et qui a généreusement continué sa collaboration, notamment pour la collecte des données qui ont été requises pour premier article présenté qui s'intitule « *Using web content analysis to create innovation indicators—What do we really measure?* ».

Je poursuis ces remerciements avec Davide Pulizzotto à qui je dois la plupart de mes connaissances en traitement du langage naturel ainsi que des données obtenues pour les articles deux et trois présentés dans cette thèse. Je suis particulièrement reconnaissant pour sa généreuse contribution qui nous a permis de créer un processus efficace et rigoureux de moissonnage du web à partir de The Wayback Machine, ainsi que de me supporter dans le traitement et le nettoyage de données non structurées.

Je voudrais également remercier toute l'équipe de The Evidence Network (TEN) pour la confiance et le soutien qu'elle nous a accordés en nous donnant accès à ses données d'enquête. Comme ces données sont utilisées dans les articles deux et trois de cette thèse, cette contribution a été essentielle à la réalisation d'une grande partie du travail présenté dans cette thèse. Nous remercions tout particulièrement Margaret Dalziel pour avoir initié cette collaboration et j'ai également bénéficié de son support lors de la rédaction des articles deux et trois.

Je voudrais aussi remercier tous les employés de Polytechnique qui facilitent notre travail chaque jour. Finalement, j'ai une pensée spéciale pour mes collègues de la Chaire en gestion et économie de l'innovation de Polytechnique. Merci pour vos suggestions, idées, conseils et support. Pour ceux que ce n'est pas encore fait, votre tour viendra bientôt.

RÉSUMÉ

La majorité des recherches quantitatives en gestion de l'innovation et de la technologie s'appuient aujourd'hui sur des bases de données publiques et des enquêtes par questionnaire pour évaluer les stratégies industrielles et les activités d'innovation. Cependant, les bases de données publiques sont souvent incomplètes ou de nature trop générale. Bien que les questionnaires restent des instruments précis, le processus de conception, de test et d'administration des enquêtes peut être particulièrement long et coûteux pour les chercheurs. En outre, les enquêtes par questionnaire souffrent de taux de réponse de plus en plus faibles (moins de 10 %) et menacent donc la validité de ces études. Or, la plupart des entreprises entreprenant des activités d'innovation disposent d'un site internet pour informer leurs parties prenantes de leur mission, de leurs activités, de ce qu'elles offrent et ce qu'elles recherchent. Cette source d'information abondante et gratuite semble être une source de données intéressantes pour les études quantitatives. Cependant, les entreprises divulguent stratégiquement seulement les informations qu'elles croient les avantagent. Malgré la présence de ce biais de désirabilité sociale très fort, le site internet corporatif reste une source de données intéressante à explorer.

Une étude antérieure a réalisé une analyse par la fréquence des mots-clés pour évaluer les activités de R-D des PME basées au Royaume-Uni. Cependant, l'indicateur web qui en a résulté ne corrélait pas significativement avec des indicateurs traditionnels liés à la R-D, suggérant que les deux indicateurs représentent des éléments différents. La nature équivoque entourant la signification de ce que cet indicateur provenant du web mesure véritablement représente un objet d'étude fascinant qui a inspiré les travaux de recherche présentés dans cette thèse. Le but premier de la thèse est de comprendre la signification de la mesure d'un indicateur provenant du site internet corporatif et de déterminer sa valeur épistémique afin de potentiellement pouvoir l'utiliser dans des études quantitatives. La thèse répond à trois grands objectifs. Le premier est de trouver une correspondance entre des indicateurs tirés du site internet d'entreprises et des mesures issues d'une source de données classique telle qu'un questionnaire. Le deuxième est de proposer une ou des interprétations tirées des résultats empiriques provenant de nos travaux. Le troisième est de déterminer les bénéfices et les limitations de ces indicateurs afin de comprendre comment les utiliser.

La thèse répond aux objectifs avec l'aide de deux sources de données principales. La première est une base de données de 79 entreprises œuvrant en nanotechnologies et matériaux avancés obtenus avec une enquête par questionnaire à l'automne 2015 suivi par le moissonnage du web des sites internet correspondants en février 2016. La deuxième source de donnée nous provient de « The Evidence Network », une entreprise canadienne de consultation spécialisée dans l'évaluation de l'impact d'intermédiaires d'innovation. Les données correspondent à 2,379 réponses d'entreprises canadiennes et finlandaises qui ont été à la fois collectées par voie de questionnaires entre 2010 et 2016 par les représentants de TEN et par le moissonnage du site internet de ces mêmes entreprises à partir de « The Wayback Machine » avec un scripte que nous avons développé à cette fin. La thèse répond aux objectifs à l'aide de trois articles qui se complètent et qui évaluent notre question de recherche avec des angles différents.

Le premier article publié a trouvé que plusieurs mesures du questionnaire liées à un facteur d'innovation corrélaient fortement avec l'indicateur provenant du web correspondant à ce même facteur, démontrant ainsi qu'il n'y avait pas de validité convergente et discriminante. Par contre, en combinant plusieurs différentes mesures issues du questionnaire en un indice global d'un facteur, nous obtenons une forte correspondance avec l'indicateur du web correspondant au même facteur, et ce avec un résultat satisfaisant en termes de validité convergente et discriminante du construit. Cela suggère que la mesure du web représente un indice global, soit l'importance relative qu'une firme accorde à un facteur d'innovation. Le deuxième article présenté dans cette thèse prend une perspective de signalement du marché pour établir que les entreprises qui ont les signaux les plus forts liés à l'innovation sont fortement associées à l'obtention de financement externe, et non pas à de l'évidence de résultats d'innovation mesurée par voie de questionnaire. Ce résultat suggère que le signal lié à l'innovation pourrait permettre de cibler des entreprises qui nécessitent des ressources externes pour innover. Le troisième article montre que les entreprises qui signalent sont généralement celles qui démontrent avoir un peu d'évidence d'innovation, une intensité de connaissance élevée, et elles opèrent dans un secteur que l'on peut s'attendre à voir plus d'activités d'innovation. Par contre, on semble voir que le signal attendu est sous-évalué dans les entreprises manufacturières et possiblement surévalué dans les entreprises liées des activités avec un plus haut degré de numérisation. Les résultats permettent également d'évaluer la performance des indicateurs pour chaque facteur d'innovation par rapport à ce qui était attendu, et donc d'établir que les champs lexicaux les plus consensuels entre la théorie et la pratique sont ceux qui sont les plus performants.

ABSTRACT

Most quantitative research in the field of innovation and technology management relies on questionnaire-based surveys and public or financial databases to assess industrial strategies and innovation activities. However, public databases are often incomplete or too general in nature. Although questionnaires remain accurate instruments, the process of designing, testing and administering surveys can be particularly time-consuming and expensive for researchers. Moreover, questionnaires are plagued by progressively diminished participation levels (below 10%), thereby posing a significant threat to the validity of such investigations.

Meanwhile, most innovation-active companies have a website to inform their stakeholders about their mission, their activities, what they offer and what they are looking for. This abundant and free source of information seems to be an interesting source of data for quantitative studies. However, companies strategically disclose only the information they believe to be beneficial. Despite the presence of this strong social desirability bias, the corporate website remains an interesting source of data to investigate.

A previous study conducted a keyword frequency analysis to assess the R&D activities of UK-based SMEs. However, the resulting web-based indicator did not correlate significantly with traditional indicators of R&D, suggesting the indicators represent different concepts. The conundrum surrounding the meaning of what this web-based indicator actually measures represents a fascinating object of study that has inspired the research presented in this thesis. The primary aim of the thesis is to understand the meaning behind the measurement of a web-based indicator and to determine its epistemic value so that it can potentially be used in quantitative studies. The thesis has three main objectives. The first is to find a correspondence between web-based indicators and measures from a traditional data source such as a questionnaire. The second is to propose plausible interpretations from the empirical evidence from our research. The third is to determine the benefits and limitations of these indicators in order to understand how to use them.

The thesis addresses the objectives with the help of two main data sources. The first is a database of 79 nanotechnology and advanced materials firms obtained through a questionnaire-based survey in autumn 2015 followed by web mining of the corresponding websites in February 2016. The second source of data comes from "The Evidence Network", a Canadian consulting firm specialising in the evaluation of the impact of innovation intermediaries. The data corresponds to

2,379 responses from Canadian and Finnish companies that were collected through questionnaires between 2010 and 2016 by TEN representatives and then, through our web mining script to collect data from The Wayback Machine. The thesis addresses the objectives with three complementary papers that assess our research question from different angles.

The first published paper found that several questionnaire-based measures related to an innovation factor correlated strongly with the web-based indicator corresponding to the same factor, demonstrating that there was no convergent and discriminant validity. On the other hand, by combining several different questionnaire-based measures into a global index corresponding to a factor, we obtain a strong correspondence with the web-based indicator corresponding to the same factor, with a satisfactory result in terms of convergent and discriminant construct validity. This suggests that the web measure represents a broad index, which can be interpreted as the relative importance that a firm allocates to an innovation factor.

The second paper in this thesis takes a market signalling perspective to establish that firms with the strongest innovation-related signals are strongly associated with obtaining external financing, not with evidence of innovation results measured by questionnaire. This result suggests that the innovation signal may enable targeting firms that require external resources to innovate.

The third paper shows that the firms with the strongest signals are generally those that show some evidence of innovation, high knowledge intensity, and operate in a sector that one would expect to see more innovation activities. On the other hand, it seems the expected signal is underestimated in manufacturing companies and possibly overestimated in companies related to activities with a higher degree of digitalisation. The results also allow us to evaluate the performance of the indicators for each innovation factor in relation to what was expected, and thus establish the lexical fields that are more consensual between theory and practice are the ones with the higher performance.

TABLE OF CONTENTS

DEDICATION III

ACKNOWLEDGEMENTS IV

REMERCIEMENTS VI

RÉSUMÉ VIII

ABSTRACT X

TABLE OF CONTENTS XII

LIST OF TABLES XVII

LIST OF FIGURES XXII

LIST OF SYMBOLS AND ABBREVIATIONS XXIII

LIST OF APPENDICES XXVI

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 LITERATURE REVIEW 7

2.1 Innovation at the firm level 7

2.1.1 Fundamentals 7

2.1.2 Evolution of the innovation process 10

2.1.3 Knowledge and appropriation 14

2.1.4 Sectorial perspectives 17

2.1.5 Innovation culture 18

2.1.6 Financing innovation 21

2.1.7 Size and age 22

2.1.8 Measuring innovation 23

2.2 Web-based studies 26

2.2.1 Webometrics 26

2.2.2 Web-based innovation studies 28
2.3 Assessing corporate website as an object of study 44
2.3.1 Semiotics 44
2.3.2 Signalling theories 46
2.3.3 Signalling from language 48
2.3.4 Assessing the self-reported nature of corporate website 50
CHAPTER 3 RESEARCH DESIGN 54
3.1 Research question and objectives 54
3.2 Propositions 55
3.3 Data and variable-related methodologies 64
3.3.1 Building indices 64
3.3.2 Keywords analysis 67
3.3.3 Data and treatment 70
3.4 Methods 104
3.4.1 Correlation analysis and MTMM matrix 104
3.4.2 Regression methods 105
3.4.3 One-way ANOVA analysis 108
CHAPTER 4 ARTICLE 1: USING WEB CONTENT ANALYSIS TO CREATE INNOVATION INDICATORS – WHAT DO WE REALLY MEASURE? 110
4.1 Abstract 110
4.2 Introduction 110
4.3 Theoretical development 113
4.3.1 Using the web as a data source for social sciences 113
4.3.2 Construct definition 115

4.4 Data and Methodology 118
4.4.1 Questionnaire-Based Data Collection 118
4.4.2 Web-Based Data Collection 124
4.4.3 The Multi-Trait Multi Method (MTMM) matrix technique 128
4.5 Results 131
4.5.1 Keywords analysis 131
4.5.2 Testing proposition 1.1 133
4.5.3 Testing proposition 1.2 140
4.5.4 Post hoc analysis: MTMM confirmatory analysis 143
4.6 Discussion and implication for research 146
4.7 Limitations and future research 149
4.8 Author contributions 150
CHAPTER 5 ARTICLE 2: IT PAYS TO SIGNAL – A COMPARISON OF WEB-BASED AND SURVEY-BASED INNOVATION INDICATORS WITH FIRMS' INVESTMENT 151
5.1 Abstract 151
5.2 Introduction 152
5.3 Conceptual framework 153
5.3.1 Measuring innovation at the firm level 153
5.3.2 Factors that contribute to innovation 154
5.3.3 Web signalling, investment and innovation evidence 156
5.4 Methodology 159
5.4.1 Data 159
5.4.2 Variable construction 163
5.4.3 Regression model 166

5.5 Descriptive statistics 167
5.6 Results 168
5.6.1 3SLS regression results 168
5.6.2 Marginal effect analysis 173
5.7 Post-hoc analysis 177
5.8 Discussion 179
5.9 Conclusion 181
5.10 Author contributions 183
CHAPTER 6 ARTICLE 3: WHERE THERE IS SMOKE, IS THERE FIRE? EXPLORING TO WHAT EXTENT INNOVATION SIGNALS FROM CORPORATE WEBSITES MEET EXPECTATIONS 184
6.1 Abstract 184
6.2 Introduction 184
6.3 Theoretical development 187
6.3.1 Innovation signal and innovation 187
6.3.2 Innovation signal and the innovation literature 191
6.4 Methodology 196
6.4.1 Data 196
6.4.2 Dependent variables 200
6.4.3 Independent variables 201
6.4.4 Model 203
6.5 Results 206
6.5.1 Descriptive statistics 206
6.5.2 Regressions results 208
6.6 Discussion 223

6.7 Conclusion 225
6.8 Author contributions 226
CHAPTER 7 GENERAL DISCUSSION 228
7.1 Web-based indicators vs questionnaire-based indicators on innovation factors 228
7.2 Web-based indicators as evidence of innovation 230
7.3 Innovation signaling vs the innovation literature 235
7.4 A note about trustworthiness 237
CHAPTER 8 CONCLUSION AND RECOMMENDATIONS 238
8.1 Research limitations 240
8.2 Future research 241
REFERENCES 243
APPENDIX A – SUPPLEMENTARY MATERIALS ARTICLE 1 278
APPENDIX B – EXAMPLES OF TEN'S QUESTIONS 286
APPENDIX C – KEYWORDS AND STATISTICS ARTICLE 2 AND 3 290
APPENDIX D – SUPPLEMENTARY MATERIALS ARTICLE 2 299
APPENDIX E – SUPPLEMENTARY MATERIALS ARTICLE 3 322

LIST OF TABLES

Table 3-1. All variables used to test the propositions 92

Table 4-1. Estimated response rate 120

Table 4-2. Questionnaire-based survey construct 123

Table 4-3. The web mining construct 127

Table 4-4. The MTMM matrix technique 129

Table 4-5. Keywords Frequencies 131

Table 4-6. Validation construct for testing proposition 1.1 133

Table 4-7. MTMM matrix output example 135

Table 4-8. Level of importance of contracting R&D with the number of IP mechanisms used 136

Table 4-9. Level of importance of contracting R&D with the number of patents 136

Table 4-10. Level of importance of providing R&D with the number of IP mechanisms used . 137

Table 4-11. Level of importance of providing R&D with the number of patents 137

Table 4-12. Proportion of employees assigned primarily in R&D with the number of IP mechanisms used 138

Table 4-13. Proportion of employees assigned primarily in R&D with the number of patents . 138

Table 4-14. Summary of construct validity results 139

Table 4-15. The validation construct for proposition 1.2 testing 142

Table 4-16. MTMM matrix for R&D_INDEX and IP_INDEX 143

Table 4-17. MTMM CFA results 145

Table 4-18. MTMM CFA correlations 145

Table 5-1. 3SLS regression results 170

Table 5-2. Marginal increase in investment by marginal increase in web-based signalling 173

Table 5-3. Marginal increase in investment by marginal increase in web-based signalling 174

Table 5-4. Investment received by web-based & evidence of innovation and sectors 178

Table 6-1. Marginal effects related to innovation evidence 209

Table 6-2. Marginal effects related to time to market 210

Table 6-3. Marginal effects related to knowledge intensity 211

Table 6-4. Marginal effects related to sector 213

Table 6-5. Marginal effects related to sector x Knowledge intensity 214

Table 6-6. Marginal effects related to sectorial category 221

Table 6-7. All results summary 222

Table A-1. Correlations table 281

Table A-2. Descriptive statistics 283

Table A-3. Correlations between individual questionnaire-based and web-based indicators 284

Table A-4. PLS regression results 285

Table C-1. The web keyword analysis construct 290

Table C-2. Document keywords frequencies by unique website-questionnaire year pair 292

Table C-3. Standardised document keyword frequencies by unique website-questionnaire year pair 293

Table C-4. Standardised document keywords frequencies by concept by unique website-questionnaire year pair 295

Table C-5. Proportion of observations without keyword occurrences 297

Table D-1. Sampling summary 299

Table D-2. Questionnaire-based survey variables 301

Table D-3. Web-based variables construct 302

Table D-4 Descriptive statistics 304

Table D-5 Pearson's correlation table 305

Table D-6. OLS regression results 308

Table D-7. OLS regression results – *post-hoc* 311

Table D-8. First order - Combined loadings and cross-loadings - first iteration 314

Table D-9. First Order - indicator weights - first iteration 315

Table D-10. First order - Combined loadings and cross-loadings – Final results 316

Table D-11. First Order - indicator weights - final results 317

Table D-12. Average variances extracted 318

Table D-13. Full Collinearity VIFs 318

Table D-14. Second order - Combined loadings and cross-loadings 318

Table D-15. Second order - First order - Indicator weights 318

Table D-16. 3SLS regression results - PLS web-based construct 319

Table E-1. Sampling summary 322

Table E-2. Average of sentences obtained by sector and country 324

Table E-3. Dependent variables 325

Table E-4. Independent and control variables 325

Table E-5. Descriptive statistics 329

Table E-6. Pearson's correlation table 332

Table E-7. One-way ANOVA - Innovation signalling and innovation evidence 339

Table E-8. One-way ANOVA - Innovation signalling and time to market 341

Table E-9. One-way ANOVA - Innovation signalling and Knowledge intensity 343

Table E-10. One-way ANOVA - Innovation signalling and sector 345

Table E-11. One-way ANOVA - Innovation signalling and sectorial category 348

Table E-12. One-way ANOVA - Innovation signalling and size 350

Table E-13. One-way ANOVA - Innovation signalling and age 352

Table E-14. One-way ANOVA - Innovation signalling and country 353

Table E-15. Zero-inflated negative binomial with countAll, countInnov and countCollab – Model a and b 357

Table E-16. Zero-inflated negative binomial with countOpen, countRD and countIP – Model a and b..... 366

Table E-17. Zero-inflated negative binomial regressions – Model c..... 375

Table E-18. Count overall innovation signal marginal effects – Model 1a 381

Table E-19. Count overall innovation signal marginal effects – Model 1b 383

Table E-20. Count innovation culture signal marginal effects - Model 2a..... 386

Table E-21. Count innovation culture signal marginal effects – Model 2b 388

Table E-22. Count collaboration signal marginal effects - Model 3a..... 391

Table E-23. Count collaboration signal marginal effects – Model 3b 393

Table E-24. Count OI signal marginal effects - Model 4a..... 396

Table E-25. Count OI signal marginal effects – Model 4b 398

Table E-26. Count RD signal marginal effects - Model 5a 401

Table E-27. Count RD signal marginal effects – Model 5b..... 403

Table E-28. Count IP signal marginal effects – Model 6a..... 406

Table E-29. Count IP signal marginal effects – Model 6b..... 408

Table E-30. Count overall innovation signal marginal effects – Model 1c 411

Table E-31. Count innovation culture signal marginal effects – Model 2c 412

Table E-32. Count collaboration signal marginal effects –Model 3c 413

Table E-33. Count OI signal marginal effects – Model 4c 414

Table E-34. Count R&D signal marginal effects – Model 5c..... 415

Table E-35. Count IP signal marginal effects –Model 6c..... 416

Table E-36. Results summary – Overall innovation..... 417

Table E-37. Results summary – Innovation culture..... 419

Table E-38. Results summary – Collaboration 421

Table E-39. Results summary – OI 423

Table E-40. Results summary – R&D 425

Table E-41. Results summary – IP 427

LIST OF FIGURES

Figure 4-1. Data mining and treatment process. 126

Figure 4-2. MTMM CFA with formative indexes 144

Figure 5-1. Web text content acquisition process 162

Figure 5-2. Investment received by the sum of standardised keyword frequencies by concept – short scale 176

Figure 5-3. Investment received by the sum of standardised keyword frequencies by concept – long scale 177

Figure 6-1. Text content collection process 200

LIST OF SYMBOLS AND ABBREVIATIONS

AC	Absorptive capacity
AI	Artificial intelligence
ANOVA	Analysis of Variance
API	Application Programming Interface
CIS	European Community Innovation Survey
CFA	Confirmatory factor analysis
cld2	Google's Compact Language Detector 2
EFA	Exploratory factor analysis
EU	European Union
FA	Factor analysis
GLS	Generalised least squares
HTML	Hypertext Markup Language
ID	Identification
ICT	Information and communication technologies
inv	inverse function
IPO	Initial public offering
IP	Intellectual property
ISED	Innovation, Science and Economic Development Canada
IV	Instrumental variable
KIBS	Knowledge intensive business services
KMO	Kaiser-Meyer-Olkin Measure of Sampling Adequacy
LDA	Latent Dirichlet allocation
LL	Likelihood

LLM	Large language model
ln	Natural logarithm
MTMM	Multi-trait multi-method
NAICS	North American Industry Classification System code
NB	Negative binomial
NLP	Natural Language Processing
OCP	Open COVID Pledge
OECD	Organisation for Economic Co-operation and Development
OEM	Original Equipment Manufacturer
OI	Open innovation
OLS	Ordinary least squares regression
PCA	Principal component analysis
Ph.D	Doctor of Philosophy
PLS	Partial Least Squares regression
PDF	Portable Document Format
PR	Public relation
R&D	Research and development
Regex	Regular expression
RFID	Radio frequency identification
SEM	Structural Equation Modelling
SME	Small and medium enterprise
STEM	Science, Technology, Engineering and Mathematics
URL	Uniform Resource Locator
UK	United Kingdom

US	United States of America
USD	US dollar
USPTO	United States Patent and Trademark Office
TEN	The Evidence Network
TF-IDF	Term Frequency-Inverse Document Frequency
UTF-8	Unicode Transformation Format-8
VC	Venture capital
VIF	Variance Inflation Factor
ZINB	Zero-Inflated Negative Binomial
ZIP	Zero-Inflated Poisson
2SLS	Two stage least square regression
3SLS	Three stages least square regression

LIST OF APPENDICES

APPENDIX A – Supplementary materials Article 1 278

APPENDIX B – Examples of TEN's questions 286

APPENDIX C – Keywords and statistics Article 2 and 3 290

APPENDIX D – Supplementary materials Article 2 299

APPENDIX E – Supplementary materials Article 3 322

CHAPTER 1 INTRODUCTION

Innovation has been the driving force behind economic growth, societal progress, and improving the well-being of the population (Grossman & Helpman, 1993; OECD & Eurostat, 2018; Scherer, 1986). Firms are one of the primary drivers of innovation (Chesbrough, 2003a; Christensen, 1997; Schumpeter, 1942). As they conduct research and development (R&D), they explore new applied knowledge base (Chesbrough, 2003a; W. M. Cohen & Levinthal, 1990; Liyanage et al., 1999; Rosenberg, 1990), bring new technologies and know-how in the form of products, processes and services to society, which lead to increased productivity and improved standards of living (OECD & Eurostat, 2018). Since the rise of the globalised economy, firms must compete with the best, not only from their local or even regional environment, but from all over the world. As high value-added products and services yield the best returns, the competition for those is intensifying bringing breakthrough technologies and practices of yesterday into the commodities of tomorrow (O'Connor & DeMartino, 2006). To thrive in today's business environment, firms must consistently provide value propositions that are superior to those of their competitors (Migol et al., 2018). Therefore, firms must continuously innovate and adapt to stay ahead of the curve, as complacency and a lack of innovation can quickly lead to obsolescence and declining profitability (Christensen, 1997). As a result, the firm's innovation process must be as effective as possible.

Meanwhile, the world is facing a perfect storm of crises: global warming, the collapse of biodiversity, the sanitary crisis and its aftermath and countless social injustices that are being exposed every day (Cowie et al., 2022; Levidow & Papaioannou, 2018; Mazzucato, 2021). To this context, we witness the rise of one of the most disruptive enabling technologies the world has yet seen: artificial intelligence (D. E. Bock et al., 2020; Puaschunder, 2019). Addressing these global challenges will require the development and implementation of breakthrough technologies, cultural shifts, and effective policy-making (Diercks et al., 2019; Dörr et al., 2022; Mazzucato, 2021). Thus, firms are increasingly expected to be both implementing innovative solutions in order to stay ahead of the competition and seemingly contribute to a more sustainable and equitable future (Stubbs, 2019). As never before, firms are operating in a very challenging and complex socio-economic environment.

Moreover, the state of innovation in Canada's recent years has been concerning. Indeed, the country is among the lowest growth rates among other developed countries and despite having a

highly educated population with strengths in important areas of research and science, the evolution of Canadian R&D is worrisome on many levels such as a declining financial support for R&D in all sectors, including firms investments (Council of Canadian Academies, 2018). The fact there are strong science and research performance while witnessing a lack of actual innovation results is a puzzling situation. Thus, we need to properly assess the causes that limits the innovative capacity. As such, innovation indicators are undeniable decision-making tools that give us a quantifiable signal of a specific element related to the innovation process (OECD & Eurostat, 2018). In doing so, they allow us to evaluate, compare and see the evolution of measurable dimensions of the innovation process at scale. As a result, they serve as a solid basis for making analyses that are as objective as possible. However, any indicator only reveals a tiny part of a bigger story which may reveal useful patterns at scale, and most traditional indicators tend to lack the nuances needed to properly assess a given situation.

This was further demonstrated with Canada's Superclusters initiative. As an attempt to redress Canada's innovation performance, the Superclusters initiative was put together in 2017 by the federal government to foster innovation within five regions to cultivate a critical mass of expertise in key enabling technologies, elaborating new ties between firms of all sizes, academic institutions, and non-profits, while promoting a Silicon Valley type of entrepreneurial spirit (Beaudry & Solar-Pelletier, 2019). However, how to properly evaluate the impact of such a complex program is not trivial. Nevertheless, it was decided to use broad generic performance indicators (e.g., the number of collaborative projects, the number of firms in collaborative projects or the number of jobs created) which does not lead to any meaningful contextual insights needed to fully assess the performance of the initiative. Beaudry & Solar-Pelletier (2019) even proposed that the Supercluster initiative framework is an opportunity to experiment with novel indicators that would better reflect the complexity of the innovation ecosystem perspective. Indeed, better knowledge about the innovation ecosystem should improve the collective choice-making process which should allow better support to all innovators in the country. That being said, we realise more and more the challenge that represents the attempt to accurately and empirically assess innovation performance across complex systems (Kinne & Axenbeck, 2020; Nagaoka et al., 2010; OECD & Eurostat, 2018; Squicciarini et al., 2013).

Additionally, even at the "simpler" firm's level, a growing number of researchers acknowledge the lack of contextual information in quantitative studies that restrict our understanding of the

factors that influence the adoption and effectiveness of practices such as those associated with open innovation (OI) for instance (Bogers et al., 2017). Of course, qualitative research may help to bridge that gap by allowing researchers to gather in-depth knowledge especially rich in context and nuances on a few subjects which is an essential part for building new theories. However, due to their small sample size and the absence of measurable quantities, qualitative studies cannot be used to generalise research results to a population of subjects, which hinders their ability to accurately gauge the state of innovation on a wider scale.

On the other hand, quantitative study researchers can measure key aspects of the innovation process gathered empirically with ideally a statistically representative sample of the population which allows the generalisation of the results. One of the main challenges faced by quantitative researchers is access to appropriate data. This is why many quantitative studies rely heavily on publicly available data, such as those provided by the United States Patent and Trademark Office (USPTO), Statistics Canada, and financial statement data as well as questionnaire-based surveys, which can be tailored to specific research questions. These two traditional data sources have been prevalent in past decades of quantitative research, but they also have limitations that need to be considered (Becheikh et al., 2006). Publicly available data sources may provide a broad range of information, but they may not always include the necessary data to accurately assess the specific innovation elements being examined in a research study. On the other hand, questionnaire-based surveys are precise instruments, but there is an increasing challenge in convincing respondents to participate in innovation studies as business leaders have time constraints, leading to lower response rates which can threaten the validity of the studies, as it may introduce significant nonresponse biases. Additionally, the process of designing, testing, and administering surveys is time-consuming and costly, making it a risky endeavour for researchers. Thus, researchers are exploring ways to reduce surveys size while seeking alternative sources of complementary data to gain a deeper understanding of how innovation is influenced within firms (Al Baghal et al., 2021; Blazquez & Domenech, 2018; Bottai et al., 2022; Callegaro & Yang, 2018; Eck et al., 2021). This opens the door to exploring new sources of data on innovation.

The rise of internet enabled firms to channel their official communication online, on very few media. Indeed, most, if not all, innovation-active firms have a website and may even have a presence on social media such as YouTube, Twitter, Facebook, LinkedIn, Instagram, and so on. However, the firm's website is arguably the most important public relations (PR) and

communication platform that it can use to effectively transmit its core messaging to achieve strategic goals (Cornelissen, 2020). The corporate website serves as the primary window for the firm to showcase its products, services, and values to the world. It is the first impression that a firm makes on its stakeholders and can shape the way they perceive the company. Therefore, a well-designed website that effectively communicates the firm's value proposition, brand, credibility, and unique selling points is crucial for a firm's success. If a firm is engaging in innovative-related activities or has an innovative culture, it is likely it would communicate these on its website. In this sense, the firm's website can not only reveal what a firm does, but it can serve as a window into its innovation culture (Herzog, 2011; Schein, 2004). If we can correctly identify these innovation-related signals, and to properly quantify them into valid indicators, this could represent a golden opportunity to bring much needed nuance to quantitative models. This could help better understand a firm's innovation profile and performance, and provide valuable insights into the broader innovation landscape, by complementing traditional data sources (Ritala et al., 2020). Not only quantifying these signals could help identify key innovation attributes across firms, industries, and countries (i.e., providing a more comprehensive view of the innovation ecosystem), but it may also significantly reduce the lag that currently exists between the data used in quantitative investigations and the present moment which could enable real-time monitoring (Dörr et al., 2022).

Although using corporate websites as a source of data for research offers potential benefits, it also comes with its own set of epistemic limitations and methodological challenges. First, firms self-report information on their websites in a way that aims to compel their stakeholders or incite them to take some form of action, a practice known as signalling (Heil & Robertson, 1991; Sebeok, 2001). Indeed, firms selectively disclose information on their websites, showcasing what they consider important while omitting other details whether deemed unimportant, strategic, or confidential (Song Shin, 2003; Verrecchia, 1983). These are non-disclosure and social desirability biases which both have for consequence to potentially risk inducing a partial, potentially distorted, vision of reality. As a result of the market social pressures discussed earlier, firms may engage in practices such as “greenwashing”, where they prominently promote their environmental commitments in secondary activities while failing to disclose the detrimental impacts of their primary activities on the climate and biodiversity (Dunlap & McCright, 2011; Guix et al., 2022; Siano et al., 2017). Furthermore, as innovation is a buzzword used by firms, innovation-related

signalling may not be a reliable measure to account for real innovativeness (Bensaude Vincent, 2014; Bergfeld & Weber, 2011; Bock B. B., 2012; Brem, 2009; Kuratko et al., 2014; Mirończuk & Protasiewicz, 2020). This put the overall level of trustworthiness of firms' official PR communication into question. As a result, the information collected on websites is certainly incomplete, subjective and may contain a certain level of noise. These considerations put to question the epistemic value of corporate website as a data source. At the same time, there is increased evidence in the branding literature that shows how authenticity enables better performance (Chen et al., 2022; Fritz et al., 2017). As such, it makes the firm's innovation-related signalling a fascinating and complex phenomenon that may reveal useful information about the firm's real innovative attributes.

Second, corporate websites are highly diverse, with substantial differences in web frameworks, content structure, size, and design (see Kinne & Axenbeck, 2020). As a result, many types of data are accessible on a website in the form of media such as images and video, written content, the hyperlinks and metadata (Axenbeck & Breithaupt, 2021). Technically, the primary difficulties stem from collecting and cleaning the data, followed by treating and making sense of highly unstructured information. By focusing on the text content, it is possible to mitigate the challenges associated with unstructured information through the use of web mining, natural language processing and text analysis techniques, which provide standardised methods for structuring the data needed to generate web-based indicators (Herrouz et al., 2013; Jenifer et al., 2016; Kayser & Blind, 2017; Xu et al., 2011). In the context of this thesis, we are interested in analysing specifically innovation-related signalling which we can retrieve from textual data with a classic content analysis technique: the keyword frequency analysis, which may be a proxy for real offline practices and activities (Thelwall, 2009).

In the beginning of the 2010s, some innovation studies started to incorporate information found on corporate websites as ways to enrich their quantitative models with specific contextualised information (S. K. Arora et al., 2013; Libaers et al., 2016; Youtie et al., 2012). One of these studies attempted to measure real innovation-related metrics from the information found on the corporate websites, more specifically by comparing a R&D web-based indicator built from a keyword frequency method from other traditional R&D indicators to see how they would correspond (Gök et al., 2015). While this study was not conclusive, it opened the door to further investigations on this topic. Is there really a lack of correspondence between what a firm repeatedly states on its

website and objective measures of the corresponding activities? What do these web-based indicators really measure? What is the true meaning behind the words? Are firms honest enough for researchers to consider their website as a reliable source of information? What are the characteristics of the firms that are mentioning more innovation-related terms?

The thesis, as a whole, assesses the epistemic value of web-based innovation indicators based on corporate innovation-related signalling and brings significant advances in the measurement of innovation, the use and development of state-of-the-art methods to acquire and process large-scale textual data, to build and validate entirely new indicators, and to better understand the use of language related to innovation as well as its impact on the market. It contributes to enriching the field of Innovation management with rigour and diversity of methods as called by prominent journals (N. Anderson et al., 2004; Antons et al., 2020; Faems, 2020; Ritala et al., 2020; Shafique, 2013).

The first article assesses the construct validity of web-based indicators through correlation analysis, Multi-trait multi-methods (MTMM) matrix and confirmatory factor analysis (CFA). The paper showed evidence that web-based indicators for R&D and Intellectual property (IP) protection do measure the same concepts as the ones built with broader indices from questionnaire data. Taking a signalling perspective, the second article showed that the web-based indicators explain better the investment received than the real evidence of innovation results that we assessed with survey data. It is suggested that corporate websites may serve as a valuable source for understanding the external resources that firms require based on the signals they emit. The third article continues the investigation with the signalling perspective by assessing how the web-based indicators compared with what would be expected from a literature's perspective on a wide range of different innovation-related external and internal factors.

The remainder of the thesis is organised as follows: Chapter 2 provides an overview of the related literature; Chapter 3 outlines our research objectives and hypotheses; Chapter 4, Chapter 5, and Chapter 6 present the articles; Chapter 7 discusses the thesis results; and, lastly, we conclude in Chapter 8.

CHAPTER 2 LITERATURE REVIEW

2.1 Innovation at the firm level

What follows is a review of all the conceptual elements that were covered from the innovation key literature that is relevant for this research. Furthermore, this review is especially relevant for innovation in technological small and medium enterprises (SMEs). It begins with the fundamental notions related to firms' innovation. Then it describes the evolution of the innovation process including perspectives about R&D, collaboration and OI. It follows with subsections about knowledge and appropriation, sectorial perspectives, innovation culture, financing, and other contextual determinants of innovation such as firms' sizes and ages. It finishes with innovation measurements.

2.1.1 Fundamentals

The concept of innovation is dynamic and constantly evolving over time. However, there are core elements that remain stable. Let's consider this definition from the Organization for Economic Cooperation and Development (OECD)'s third edition of the Oslo manual:

> Innovation is the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organisational method inbusiness practices, workplace organisation or external relations. (OECD & Eurostat, 2005, p.46)

This definition provides a good foundation to understand what is meant by innovation more generally. Already, the reader may appreciate two key characteristics inherently of any innovation: novelty and implementation. However, the definition has the default of his qualities: because of its ability to more or less fit most use cases, it was considered rather confusing to be operationalised by practitioners. This led to the latest edition of the Oslo Manual where the concept of innovation can either be interpreted as an outcome or a process. In the context of innovation in firms, an outcome is referred as a business innovation, which can be defined as:

> […] a new or improved product or business process (or combination thereof) that differs significantly from the firm's previous products or business processes and that has been introduced on the market or brought into use by the firm. (OECD & Eurostat, 2018, p.33)

While the implementation criteria did not change, the scope for novelty has been reduced to the firm's perspective. The interpretation of firms' innovation as a process led to this definition:

> Innovation activities include all developmental, financial and commercial activities undertaken by a firm that are intended to result in an innovation for the firm. (OECD & Eurostat, 2018, p.33)

How well a firm performs in these activities should be a determining factor on innovation outcomes. Thus, the innovation process may be broken down into different activities, from the generation of ideas to the commercialisation of goods and services or the implementation of new commercial or business practices. One of those activities that is central to technological innovation is research and development, which is defined in the Frascati Manual as:

> Research and experimental development (R&D) comprise creative and systematic work undertaken in order to increase the stock of knowledge – including knowledge of humankind, culture and society – and to devise new applications of available knowledge. (OECD, 2015, p. 44)

Therefore, innovation relies on the ability to translate pieces of knowledge into practical applications. A key aspect of this process involves possessing enough knowledge to modify, supplement, and integrate previously acquired knowledge from a fresh perspective, making it useful within the context of potential implementation (Zahra & George, 2002). This transformation necessitates bissociation i.e., combining or recombining two seemingly unrelated ideas or pieces of knowledge to generate a fundamentally innovative third element, a process called creativity (Koestler, 1964). Therefore, the combination of creativity and knowledge allows the R&D function to generate new outputs (Amabile, 1996, 1997; Amabile et al., 1996; Worren et al., 2002) which are called inventions (Burke, 1998). These inventions are intellectual assets, and if those are functionalities, they may be protected from being copied by patents. Going back to the definition, only implemented inventions are considered innovation.

The latest edition of the Oslo Manual also distinguishes different typologies of firms based on their innovation profiles. During a given time period, the innovative firms are defined as those who report one or more innovations, while innovation-active firms are those that perform one or more innovation activities without necessarily having registered any innovation results yet (OECD & Eurostat, 2018). Furthermore, the two main types of innovations, product and business process innovations are also important to distinguish and are described as follows:

> A product innovation is a new or improved good or service that differs significantly from the firm's previous goods or services and that has been introduced on the market. (OECD & Eurostat, 2018, p.34)
>
> A business process innovation is a new or improved business process for one or more business functions that differ significantly from the firm's previous business processes and that has been brought into use by the firm. (OECD & Eurostat, 2018, p.34)

These two types of innovations were mostly considered from a technological perspective, until the latest Oslo Manual edition which now encompasses all different forms of processes (OECD & Eurostat, 2005, 2018; OECD, 1997). This distinction between technological product and process innovation was first introduced with Schumpeter early work (Becker & Knudsen, 2002). The product innovation provides value in terms of increased functionality/differentiation/quality, while the process innovation enables cost reduction, improving flexibility and lead time (Damanpour & Aravind, 2006). Utterback & Abernathy (1975) showed the close sequential relationship between product and process innovation whereas a completely new product is introduced for the first time in an industry and see a high rate of product innovation until there is a market consensus on a set of core functionalities and form factor, also called a dominant design (P. Anderson & Tushman, 1990), which then lead to a higher rate of process innovations.

Many other typologies of innovation were proposed over time. There were technological (products, processes and services), marketing (the "4Ps", i.e., product, price, place and promotion) and administrative (strategy, structure, systems in place and human resources) innovations (Afuah, 1998; OECD & Eurostat, 2005). The main distinction between a technological product innovation and a marketing product innovation is that the former introduces a new functionality, while the latter is materialised when there is significant novelty introduced in product design or packaging (OECD & Eurostat, 2005).

Innovations can also be classified according to their degree of novelty. An innovation is said to be incremental when there is an improvement on existing products, services or processes, whereas the product of the innovation process is radical when the innovation introduced is completely new with a degree of radicalness that fluctuates according to whether the novelty is at the scale of the firm or at the scale of the world (Ettlie et al., 1984). A radical innovation changes the nature of an industry is called a disruptive innovation (Christensen, 1997) as these innovations create new value

chains that outcompete and replace the older ones in a process also called creative destruction (Schumpeter, 1942). Within the disruptive innovation theory, sustaining innovations follow the introduction of a disruptive innovation which describes incremental product innovations, which is also followed with efficiency innovations that describes process innovations (Christensen, 2013; Christensen & Raynor, 2013).

While other types of innovation exist (e.g., social innovation, responsible innovation, sustainable innovation, value innovation, etc.), this thesis focus on technological innovation such as product, process and services.

2.1.2 Evolution of the innovation process

While its role evolved over time, R&D has always been instrumental to innovation results. In the post-war period, the innovation process was essentially closed and linear (Bush, 1945). The process began broadly with a period of basic research that could lead to discoveries and then led to ideas which turned into inventions (Kuznets, 1962). When an idea became interesting to explore for commercial purposes, an applied research process began. Then, when the idea of a concrete product was defined, the development period began and finally ended with the commercialisation of the finished product. R&D was therefore done in a vacuum and was considered a major asset for companies, especially for the larger ones. These large companies maintained and improved their R&D by recruiting and retaining highly qualified employees. The firm would then invest in its own R&D projects to develop new products and services. This ability to perform R&D allowed these large firms to generate enormous economies of scale, allowing them to maintain a clear competitive advantage (Chandler, 1990). The commercialisation of these new products generated new revenues allowing the company to invest in new R&D projects (Schumpeter, 1942). This innovation process model was defined as science-push, while scientists were left to do creative and more basic research (Burns & Stalker, 1961). It was followed by more applied approaches such as technology-push that describe the outgoing flows from the R&D department towards the market and of market-pull flows which orient the R&D process to meet a specific market demand (Mowery & Rosenberg, 1979). Applied research has thus replaced basic research as the focus of R&D. Gradually, there is the insertion of a multidisciplinary approach and management of specific projects, controlled and chosen by the organisation according to their commercial potential (Griliches, 1994). This led to a third model which introduced the concept of portfolio management

and integration with the firm's other business activities. The objective was to better plan R&D in large firms by allowing for better interaction between the R&D department and the decision-makers who deal with strategic issues (Roussel et al., 1991). Then a fourth model proposes to include internal and external knowledge management with research and collaboration networks, since it recognised that partners in an R&D consortium have shorter time-to-market, reduced costs and are less exposed to risks (Liyanage et al., 1999). This implied that innovation results were improved through optimised R&D processes by opening its boundaries for knowledge and technologies acquisition and active collaboration with external organisations.

The positive relationship between collaboration and innovation is now well established as different actors come together with different complementary expertise and perspective to cooperate by solving a problem more efficiently (Cropper et al., 2008; Schrage, 1990; Wood & Gray, 1991). Innovative activities require a high level of collaboration among individuals whether these activities are solely within an organisation or across organisations (Maccoby, 2006). Upstream of the innovation process, collaboration may enhance the processes of idea generation, evaluation, and validation (Koen et al., 2014). Firms benefit from cooperating on the advancement and implementation of complex and emerging technologies, as it allows them to share R&D expenses, reduce risks, and improve their overall performance (Johnson & Filippini, 2009; Jordan & Lowe, 2004; Parker, 2000). As companies within a value chain have different profiles, strengths and weaknesses their motivations to enter strategic alliances varies, although there is always something to be gained whether the reasons are due to access to economies of scale, risk management, market considerations, product synergies, access to complementary resources and skills or just as preemptive strategies (Tian Xie & Johnston, 2004). Furthermore, there is empirical evidence that collaboration enhance a firm's innovation performance (Faems et al., 2005; Hagedoorn et al., 2000; Roja & Nastase, 2013). It has also been shown to increase indicators such as patent numbers, sales growth, and return on innovation sales (Arvanitis, 2012; Belderbos et al., 2004; Carboni, 2013). The Triple Helix (Etzkowitz & Leydesdorff, 2000) also describes collaborative relationships between industrial, academic and governmental entities and has shown to yield a positive impact on firms' performance (Armellini et al., 2014, 2017; Kim et al., 2014; Li et al., 2018; McNeil et al., 2007).

Several socio-economic factors explain this shift towards openness. These factors described in Chesbrough (2003) are essentially the advent of the internet and telecommunications, the

emergence of globalisation, which had the main effect of promoting the free movement of individuals and capital, and the rise of venture capital. The advent of information and communication technologies (ICT) initially facilitated the codification and digitisation of knowledge and some easier to codify know-how. With the arrival of the Internet, the large-scale dissemination of this codified knowledge has been made possible, which has greatly facilitated access to an international pool of knowledge. This pool represents the accumulation of positive externalities (spillovers) from companies, universities, governmental institutes and others. This new supply of free and accessible information were initially considered as a threat for those companies that relied heavily on knowledge asymmetry as a competitive advantage. This represents an opportunity for potential new entrants that know how to adapt to this new reality. With the free movement of people, it is becoming increasingly difficult for companies to retain the most skilled workers as it is now easy for these workers to be offered better jobs by higher bidding foreign companies. In doing so, the company loses the means to protect and retain the ideas and tacit knowledge within the company. With the rise of the free movement of capital, the venture capital (VC) market has also expanded. Indeed, VC firms have never been as numerous as at the dawn of the 21st century, which has generated a very strong competition for the investment of innovative projects. As a result, it is now very attractive for employees of a large company to want to launch their own high-growth technology company (also called "startup") and to commercially develop themselves the fruit of the research previously done within the company, a process called a “spin off”.

As a result, the Open innovation (OI) concept proposes that today's innovators need to go beyond the confines of their own walls in order to maximise their R&D process and successfully launch new products, services, develop new processes or other organisational practices (Chesbrough, 2003a, 2003b; Enkel et al., 2009). As it was shown before, OI does not really refer to new practices as successful companies were engaged in them for many years (Oakey, 2013; Trott & Hartmann, 2009). However, the processes and practices prescribed by OI provide a useful framework on how a firm should manage its innovation process in nowadays business environment as these are now fully integrated in the latest Oslo Manual edition.

Although several definitions have been attempted since 2003, this definition proposed a few years ago is considered:

> Open innovation is a distributed innovation process based on purposively managed knowledge flows across organisational boundaries, using pecuniary and non-pecuniary mechanisms in line with each organisation's business model. (Chesbrough & Bogers, 2014, p. 27)

OI is therefore an innovation process that offers a new way of organising work for the purpose of innovation. OI can be executed through three key groups of activities: 1-) inbound activities to channel flows of intellectual assets (knowledge, ideas, know-how, inventions, or technologies) from the external environment into the firm (also called "outside-in"), 2-outbound activities to channel flows of intellectual assets from the firm to the external environment (also called "inside-out"), and 3- coupled activities that are a combination of the first two groups which happens during a collaboration between two or more firms with complementary intellectual assets (Gassmann & Enkel, 2004).

As mentioned in the definition above, OI is manifested through both pecuniary and non-pecuniary mechanisms. Pecuniary mechanisms include acquisitions for inbound OI activities, sales for outbound activities and are executed in the form of monetary transactions while nonpecuniary mechanisms include technology scouting for inbound activities and disclosures for outbound activities (Dahlander & Gann, 2010). Non-monetary mechanisms offer indirect benefits such as access to new ideas and knowledge (Laursen & Salter, 2006b) and the opportunity to solve problems from radically different perspectives (Lakhani, 2008). In the case of disclosure, the benefits are more subtle and strategic, such as gaining technical assistance from an external community (Henkel, 2006), gaining credibility from the external environment (Nuvolari, 2004) as well as at the societal level, allowing for the stimulation of incremental and cumulative innovations (Levin et al., 1987).

Inbound activities are by far the most documented in the literature (Bogers et al., 2017). The most common activities are technology scouting, licensing of IP, research programs with universities, external spin-in or investment, or incubation of startups, collaboration with intermediaries, with suppliers and customers, use of non-disclosure agreements, open outsourcing, innovation competitions, and access to online communities (Bogers et al., 2017; Chesbrough & Bogers, 2014).

The idea with outbound is to offer to the innovation market, technologies and knowledge not used internally (Chesbrough & Schwartz, 2007). The main idea behind outbound OI activities is to be

able to convert the uncontrolled externalities of the firm into business opportunities through the use of an appropriate business model (Chesbrough, 2006a). Moreover, a major paradigm shift is taking place in the development of business models designed specifically to take advantage of technologies, as opposed to the development of technologies that serve the dominant business model of the company (Chesbrough, 2006b), which used to put firms into the innovator's dilemma position (Christensen, 1997). As a result, large firms have advantage to spin out the development of radical technologies (Chesbrough & Rosenbloom, 2002; Christensen, 1997). Activities listed in the literature include licensing IP, donating IP, spinning off, selling or spinning off from within the firm, corporate venture capital, corporate incubators, creating joint ventures and strategic alliances.

Coupled activities correspond to an exchange of flows across the organisational boundary, which result in collaborative innovation creation and/or commercialisation projects in the form of strategic alliances, consortia, joint ventures, formal or informal collaboration, and the establishment of innovation networks. Coupled activities can be bidirectional, in the form of exchanges, or interactive where a true co-creation exercise can take place (Piller & West, 2014). The respective goals of the stakeholders must be jointly aligned for the collaboration to be productive as it was exposed by the Symbian fiasco (West & Wood, 2014). The same applies to consortia, strategic alliances, and joint ventures.

The determinants favourable to greater levels of OI practices are: 1) high product modularity, 2) high industry speed, 3) industry has a high level of knowledge, 4) interfaces are highly complex and 5) a portion of the value of positive externalities is capturable by the source; whereas the determinants that favour a rather closed innovation strategy are (1) low product modularity, (2) low industry speed, (3) the industry has a low level of knowledge, (4) the interfaces are low in complexity and (5) a portion of the value of positive externalities is very difficult to capture by the source (Gassmann & Enkel, 2004).

2.1.3 Knowledge and appropriation

Although organisations can easily tap into a vast pool of external knowledge, it does not imply that they should stop investing in R&D. Rosenberg (1990) has shown that one of the reasons firms invest in basic research is to develop their ability to grasp the significance of applied science coming from within or elsewhere. Cohen & Levinthal (1990) went further by visiting the

behavioural and cognitive science literature promoting learning at the individual level which allowed them to introduce the concept of absorptive capacity (AC) which is defined as:

> […] the ability of a firm to recognise the value of new, external information, assimilate it, and apply it to commercial ends [...] is largely a function of the firm's level of prior related knowledge. (Cohen and Levinthal, 1990, p.128)

There are two important elements that emerge from the definition: 1) it corresponds to the organisation's ability to recognise the usefulness of external information and use it to achieve innovation goals, and 2) its effectiveness depends almost entirely on the firm's current level of knowledge. AC is therefore a by-product of R&D efforts. When a firm invests in its R&D, it also invests in its ability to analyse and integrate knowledge into its innovation process. Zahra & George (2002) describe a formal AC process as:

> […] a set of organisational routines and processes by which firms acquire, assimilate, transform and exploit knowledge to produce a dynamic organisational capability." (Zahra and George, 2002, p.186).

The benefits of AC on innovation (Tsai, 2001), performance in joint venture (Lane et al., 2001), intra-organisational (Szulanski, 1996) and inter-organisational (Lane & Lubatkin, 1998) knowledge transfers, and the performance of international firms (D. R. Adams et al., 2015) have been demonstrated in the literature.

The main sources of knowledge used for R&D are universities and public research institutes, clients, suppliers, competitors, knowledge-intensive service providers, strategic alliances and manufacturing teams (W. M. Cohen et al., 2002; OECD & Eurostat, 2018).

Foray (2004) describes two main types of knowledge: codified and tacit. Tacit knowledge is related to a person’s direct experience under specific context that allows expertise to emerge. Learning tacit knowledge requires hands-on training, experimentation and learning by doing. It is therefore more difficult to formalise, to replace and to obtain as it requires exposure to individuals with the relevant experiences and communication abilities to share them. There are also often described in terms of know-how. On the other hand, codified knowledge can be communicated directly through language and recorded via technological means such as paper or information technologies.

Therefore, it is related to facts or ideas or norms. It can be easily formalised, reproduced and transferred. As a result codified knowledge improves the efficiency of knowledge market.

Furthermore, this codification of knowledge enables stronger mechanisms to protect IP. This capacity to protect invention from being copied through the use of IP mechanisms is called appropriability, which enable firms to profit from their inventions by obtaining exclusive rights (Levin et al., 1987; Teece, 1986). The motivation to apply for a patent is anchored in a strategy which can be used to block alternative designs or to encourage competition to negotiate, either by adopting the patented standard or to access their technologies (W. M. Cohen et al., 2000a). Stronger appropriability also improves knowledge market efficiency.

Potential buyers and sellers of intellectual assets are entangled into a fundamental information asymmetry problem. The buyer wants as much information as possible to ensure that the asset is worth what seller says it is. The seller, on the other hand, wants to give as little information as possible, as he does not want to be bypassed by disclosing too much information that would allow the second party to recreate the technology. This is called the information paradox (Arrow, 1962). A formal IP protection such as a patent contributes greatly to solve this issue as the information is made public while ensuring the firms to retain control over the ability to capture the value of the innovation. This makes strong appropriability an enabling factor for OI practices allowing firms to sell their unused inventions (Chesbrough, 2003b; Rivette & Kline, 2000). Furthermore, formal IP ensures a climate of trust when exploring new collaboration projects (W. M. Cohen et al., 2000a).

Laursen & Salter (2014) compared the extent of external sources of information in innovation activities and the depth of collaboration with the degree of appropriation of 2931 firms from 23 different industries. They found positive and significant concave relationships between the degree of appropriation, the breadth of external sources and the depth of collaboration. This means that after a certain point there was a slight diminishing return. Similar results were also observed between appropriation and innovation performance (Laursen & Salter, 2006a). In light of these two studies, the more firms use IP protection mechanisms, the more they seem to be relying on external sources of information in their innovation activities until a certain point where it appears to become less necessary. In addition, the more firms use IP protection mechanisms, the more they seem to carry out formal collaboration up to a certain point where a decrease in activity is observed.

2.1.4 Sectorial perspectives

In the literature, industrial sectors are analysed on the dynamics and relationships between firms within a specific sector as well as the interactions and linkages between firms across different sectors.

Industrial sectors can be analysed through the lenses of technological regimes. These are defined in terms of opportunities, properties of the knowledge base of the sector, cumulativeness and appropriability conditions (Breschi et al., 2000; Breschi & Malerba, 1997). The characteristics of a firm's knowledge varies significantly across different technologies, exhibiting varying levels of specificity, tacitness, complexity, and independence (Winter, 1997). Technological opportunities can be understood as higher returns from the investment made in R&D. Firms operating in sectors with higher opportunities should yield more innovation results. The cumulative nature of technological advancements is described as a process where new technologies are developed, and build upon previous knowledge and innovations, leading to a continuous cycle of progress and improvement. Every innovation sets off a sequence of subsequent innovations, which either enhance the original innovation incrementally or introduce fresh knowledge that can be used in related fields. Environments with high levels of cumulativeness are often characterised by a steady flow of innovative activities and increasing returns, where each new advancement builds upon the knowledge gained from past innovations. Strong appropriability is a cutting-edge sword, as it gives an incentive for innovation, but it may also limit the pace of cumulativeness (Scotchmer, 1991).

Pavitt (1984) sectoral taxonomy enables the differentiation of sectors based on their position along the innovation value chain, with four categories: supplier-dominated, scale-intensive, specialised-suppliers, and science-based industries. The supplier-dominated sectors correspond mainly with SMEs with low internal innovation capacity, focusing on process innovation to achieve cost reduction. The scale-intensive category is associated with large firms with high capacity for process innovation and vertical technological diversification, aiming to optimise for economies of scale. Specialised-suppliers refer mainly to small firms with high innovation capacity that focus on very specific product innovation. Finally, science-based industries are also classified as large firms with high internal innovation capacity, emphasising both product and process innovation.

Statistics Canada has adapted Pavitt's taxonomy by creating a dichotomy based on the level of wages, adding value-added knowledge-based industries (combining specialised-suppliers and

science-based categories) and scale-based industries that provide higher-paying jobs compared to the supplier-dominated category, which was further divided into resource-based and labour-intensive industries to better reflect the Canadian economy's characteristics (Baldwin, 1995; Baldwin & Rafiquzzaman, 1995). These categories not only differentiate industries with higher-paying jobs but also distinguish innovative industries from less innovative ones. Castellacci (2008) also adapted the taxonomy by distinguishing advanced knowledge providers (value-added knowledge-based) from mass production goods (scale-based and science-based) and another category called "Supporting infrastructure services" and the supplier-dominated goods and services. With data from the 2002-2004 European Community Innovation Survey (CIS), Castellacci (2008) showed that advanced knowledge providers, represented by knowledge-intensive business services and specialised providers, had a higher level of innovativeness than supporting infrastructure services and supplier-dominated goods and services and a similar level of innovativeness as mass production goods. Furthermore, there were also meaningful differences between advanced knowledge providers, with Knowledge intensive business services (KIBS) showing higher levels of opportunities, higher acquisition of external knowledge, lower appropriability through patents, and closer ties with universities. Finally, science-based firms shown to have significantly higher levels of innovativeness, cumulativeness condition, appropriability through industrial design, process-product orientation, marketing innovation and intramural R&D (Castellacci, 2008).

2.1.5 Innovation culture

Any group of individuals who share common elements and have a common history has a culture (Schein, 2004). As organisations are made up of individuals, they also develop of a common culture. Two perspectives are used to study organisational cultures: either organisation possesses a culture or an organisation is defined by its culture (Ernst, 2001; Herzog, 2011). For the purposes of this analysis, we will follow the trend in the management literature that considers organisational culture as an element that the organisation possesses. Therefore, this definition is considered:

> The culture of a group can now be defined as a pattern of shared basic assumptions that was learned by a group as it solved its problems of external adaptation and internal integration, that has worked well enough to be considered valid and, therefore, to be taught

> to new members as the correct way to perceive, think, and feel in relation to those problems. (Schein, 2004, p.17)

Corporate culture is therefore dynamic, iteratively shaped by all the accumulated organisational experiences. It is transmitted through verbal and non-verbal communication among group members and is greatly influenced by the group's leaders. Although it is dynamic, it can be difficult to change (Barney, 1986). The larger and more established the organisation, the more it will struggle with the structural inertia that makes change particularly difficult (Hannan & Freeman, 1984).

Corporate culture is a complex phenomenon whose some characteristics may be easily spotted by an external observer while other more subtle manifestations only occur at the unconsciousness level of the group members. For this reason, Schein (2004) define three levels that allow the characteristics of organisational culture to be classified according to different levels of clarity: 1-practices which includes artefacts and behaviours, 2-norms and 3- shared basic values.

The practices are the visible processes and organisational structures of the firm as well as the architecture of the physical environment, the language used, the technologies and products used, the artistic creations, the various styles of dress, the decorum and emotional restraint, the rituals, ceremonies and rites, the stories and myths told about the organisation, etc. (Schein, 2004). Within the practices, there are behaviours which are the elements that an external observer will notice about the way people act and interact within a group. The other aspect, the artefacts, are everything else noticeable to an outsider. Although many elements can be appreciated and understood from an external observer, Schein (2004) gives the examples of the pyramids of Egypt and the Mayans to convey how it can be difficult to decipher the real symbolism behind these grandiose monuments. The norms constitute what is expected of individuals by the group in terms of behaviours, social principles, shared goals, philosophies, attitudes, and adopted social standards. Shared basic values are the deepest manifestations of culture, encompassing unconscious believes that determine what is right or wrong, normal or abnormal. Any transgression by a member of the group would be seen as inconceivable.

Burns & Stalker (1961) analysed both organisational structure and culture to obtain two types of organisations: mechanistic versus organic. The mechanistic structure is fundamentally hierarchical and structured, whereas the organic structure has a flat structure and allows for more flexible

organisation. Cameron et al., (1991) added the dimensions "internal maintenance" versus "external positioning" to the organic and mechanistic structures, we refine the characterisation according to the quality of the competitive environment deriving four types of culture: clan, adhocracy, hierarchy and market.

Different orientations allow companies to differentiate themselves in terms of their innovation philosophy. These orientations represent beliefs, attitudes and actions on the part of companies, which allows us to trace a direct link with the second level of clarity associated with organisational culture, i.e., norms. An analysis of the literature on market orientations (the market pull), technological orientations (the technological push), learning orientations and entrepreneurial orientations, allowed to gather the characteristics inherent to an innovation culture:

> […] (1) organisation-wide shared basic values that support innovation,
>
> (2) organisation-wide norms for innovation, and
>
> (3) perceptible innovation-oriented practices (artefacts and behaviours)
>
> constituting a firm's or business unit's internal environment that, for example, encourages risk taking, supports openness to new ideas, tolerates failures, fosters learning, and promotes constructive dissent. (Herzog, 2011, p.81).

Openness to new ideas, which may also be seen as fostering a creative culture, ranked as a primary aspect of innovation culture as reported by several authors (P. Berthon et al., 1999; Hurley & Hult, 1998; Kleinschmidt et al., 2007; Zaltman et al., 1973).

Risk taking and acceptance of the possibility of failure are two almost inseparable characteristics, i.e., one cannot exist without the other. Without risk taking and acceptance of the possibility of failure, no company would bring innovative new products, especially radical ones, to market. Therefore, an innovation culture must encourage risk-taking (Amabile, 1997; Cooper & Kleinschmidt, 1995; Jaworski & Kohli, 1993; Sundgren et al., 2005) and must have the attitude that failure is a source of learning (Botcheva et al., 2002; Cummings & Teng, 2003; Day & Schoemaker, 2000; Sundgren et al., 2005).

Firms that encourage communication between scientific and technological personnel within the firm are very important, as Capon et al. (1992) has shown empirically. This is related to the intrinsic quality of an organic structure that allows interdisciplinary communication to flourish.

Beyond the intensity of communication, a climate that allows the clash of ideas is a very important dimension. Indeed, groupthink is a psychological phenomenon that describes the propensity of individuals within a group to want to preserve a harmonious work climate or one that is consistent with the group's thinking (Janis, 1971). Freedom of expression (Horibe, 2016; Siguaw et al., 2006; Sundgren et al., 2005) and a climate that fosters the emancipation of divergent opinions (Hauschildt & Salomo, 2007; Van de Ven & Chu, 1989) is also necessary. In a nutshell, a culture of innovation must foster multi-directional communication as well as create a climate that minimises, if not eliminates, the possible effects of groupthink.

2.1.6 Financing innovation

OECD & Eurostat (2018) describes 11 sources of funding which are pre-existing internal funds, transfers from affiliations, customer orders which includes public procurement, shareholder loans, public loans, loans from multinationals, private equity, grants or subsidies, bonds and obligations and other sources such as crowdfunding. The unpredictability of R&D outcomes and the long-term horizon for returns makes it challenging to secure sufficient financing. As a result, firms may struggle to undertake and sustain R&D activities (K. D. Miller & Bromiley, 1990). For young firms operating in cutting edge immature industries need external funding to survive the so-called valley of death (Kalil, 2005; McNeil et al., 2007). External financing can provide firms with the resources needed to undertake R&D activities, allowing them to develop and implement innovative products, processes, and services, while financing the commercialisation phase enables firms to expand their operations and market their innovations more effectively (Kor & Mahoney, 2005; Santarelli, 1991). Furthermore, researchers have consistently found that investment in R&D is associated with increased innovation performance, which in turn can drive firms' growth and competitiveness (Cornaggia et al., 2015; Hsu et al., 2014; Kerr & Nanda, 2015; Wellalage & Fernandez, 2019). With publicly traded firms, Aghion et al. (2004) demonstrated the relationship also works in the opposite direction, with firms signalling higher R&D intensity receiving more investment as it proved to be often a marker for upcoming growth. Indeed, younger smaller firms have limited internal resources, and receiving external funding enables more product and process innovations (Wellalage & Fernandez, 2019). Furthermore, private equity investors have shown to provide firms with the resources and expertise needed to undertake innovative projects and develop new products or services, thus driving R&D productivity (Brown et al., 2009). However, firms

adopting highly innovative approaches face increased uncertainty, resulting in a decreased probability of obtaining initial external funding compared to firms with incremental strategies, who take advantage of well-defined market niches and more evident success criteria (Burton et al., 2002).

Public financing resources are also very valuable for innovators. Public procurement of innovations has also been identified as a potent demand-side innovation policy instrument creating new firms and collaborative ties in the economy (Edler & Georghiou, 2007; Geroski, 1990). Furthermore, Government support for research and development (R&D) activities can provide firms with the resources needed to undertake innovative projects that may not be financially viable otherwise, particularly when private capital is not available because the technology is not deemed market-ready enough (Beugelsdijk & Cornet, 2002; Keizer et al., 2002). Some argued that governments are even better risk takers than private entities when it comes to funding potentially radical innovation (Mazzucato, 2011). Another benefit of public funding it that it can signal to private investors that a firm's R&D activities have promised and that it is a worthwhile investment opportunity, potentially attracting additional private financing (Meuleman & De Maeseneire, 2012).

2.1.7 Size and age

Larger firms generally have greater resources to allocate towards R&D and are less susceptible to market fluctuations to better withstand the risks associated with developing and introducing new products and services (Van Dijk et al., 1997). The main distinction between large firms and small firms' strengths are generally thought in terms of resource accessibility versus behavioural characteristics (Vossen, 1998). Although agility can be observed in both big and small companies, smaller organisations frequently demonstrate a greater capacity to promptly adapt to market demands and evolving circumstances compared to larger entities (A. Arora & Gambardella, 1994; Rothwell, 1989). There is an extensive literature on tech SMEs to be important contributors with important new ideas and technologies to the market (Acs & Audretsch, 1990; Baumol, 2005; Hicks & Hegde, 2005). This is further supported by how much important the impact of SMEs is with introducing higher performing innovations (Breitzman & Hicks, 2008; Utterback & Abernathy, 1975).

In addition, large enterprises typically operate within mechanistic structures that are fundamentally hierarchical and organized, while smaller firms tend to operate within more organic structures with less hierarchy (Burns & Stalker, 1961; Herzog, 2011). While collaboration is proven to be quite impactful for SMEs (Ebersberger & Herstad, 2012; Hanna & Walsh, 2002; Li et al., 2018; Torkkeli et al., 2007), they tend to collaborate with few external partners, often clients or providers, as they lack the resources to entertain more ties (Kaminski et al., 2008). Larger firms can also have more ease to carry on OI initiatives as seen with the cases of large companies such as P&G, Xerox, IBM and Intel (Chesbrough, 2003a). In fact, 125 large companies responded to an OI survey showing an overall average of 78% adoption in 2012 (Chesbrough & Brunswicker, 2014). While OI should be especially valuable for SMEs (Brunswicker & Van de Vrande, 2014), the use of inter-organisational linkages is especially difficult to leverage (S. Lee et al., 2010). Larger firms are also better equipped when it comes to IP protection since it entails high costs (Acs & Preston, 1997; Drechsler & Natter, 2012).

The relationship between a firm's age and innovation is not trivial, similar to the relationship between a firm's size and innovation. Older companies are likely to accumulate experience and knowledge over time, building absorptive capacity (W. M. Cohen & Levinthal, 1990), which may lead to more innovation than younger firms that have not yet built that foundation (Sørensen & Stuart, 2000). However, other research has shown that established procedures and routines in older firms may create resistance to the integration of significant external advances, hindering innovation (Freel, 2003). To circumvent this innovation dilemma, mature firms are recommended to allow disruptive innovations to be developed outside of the main core business operations (Christensen & Raynor, 2013). Additionally, within the OI paradigm, more mature firms can invest in startups as a way to stay relevant in the innovation game (Chesbrough, 2012).

2.1.8 Measuring innovation

Innovation being a very complex phenomenon with many processes and stakeholders involved under many different contexts has always been very difficult to conceptually grasp and even more challenging to objectively measure (Archibugi & Sirilli, 2001; Makkonen & van der Have, 2013). As a matter of fact, Dziallas & Blind (2019) analysed 226 publications between 1980 and 2015 about product and process innovation indicators and reported 82 unique indicators to assess different elements from the innovation process.

To provide standardised guidelines to the research community, the OECD introduced the Oslo Manual in 1992. This enabled country innovation benchmark as it enabled the CIS (OECD & Eurostat, 2018). To assess the impact of the Manual on the different studies involving innovation indicator, Becheikh et al. (2006) published a systematic review of the literature reporting on the different measures of technological innovation in the manufacturing sector from 1993 to 2003. They reported indirect measures such as R&D activities (6%) and patents (18%) from direct measures such as innovation count (25%) and survey innovation assessment based on the Oslo Manual definition (24%)[1]. All these metrics have important epistemic flaws (e.g., R&D activities are an input that may not be representative of outputs (Flor & Oltra, 2004; Kleinknecht et al., 2002), not all innovations stem from R&D activities (Michie, 1998), biased towards large firms as SMEs may not engage "officially" in R&D (Kleinknecht et al., 2002; Michie, 1998); patents measure invention not innovation (Coombs et al., 1996; Flor & Oltra, 2004), not all patenting activities lead to innovation and not all innovation can be patented, it varies greatly across sectors and firms' sizes (Archibugi & Sirilli, 2001; Armellini et al., 2014, 2017); the innovation count requires industry expert judgment which is idiosyncratic and subjective (Archibugi & Sirilli, 2001), that measure tends to favour radical & product innovation and dismisses incremental and process innovations (Flor & Oltra, 2004; Kleinknecht et al., 2002; OECD, 1997); and finally, survey innovation assessment validity is exposed to the answer rate, and the quality of the question has to properly discriminate innovation firms from one to another, which has been argued not to always been the case as 80% of firms reported to new or significantly improved products (Amara et al., 2004). Because of these shortcomings, other measurements were used such as index (15%), other measurements (9%) e.g., the sales due to innovation and no innovation measurement (4%) through assessing the main innovation determinants.

R&D intensity is a representation of the effort that a firm allocates to its own R&D. It is therefore a measure of intention, and therefore excludes direct measurement of the performance of the R&D process in generating knowledge and inventions. Patent information is a good performance indicator of the firm's inventiveness, or its ability to generate inventions. However, patent use is subject to variance across sectors and systematically disadvantages SMEs that have fewer financial resources to protect their IP. Although R&D intensity and information obtained from patent data

[1] "During the last three years, did your business unit introduce onto the market any new or significantly improved products?"

can be rightly used as a proxy for innovation performance in the case of firms with revenues above $100 million (Hagedoorn and Cloodt, 2003), they remain proxy measures that may overestimate innovation performance by not taking into account the execution i.e., the implementation process or commercialisation. On the other hand, SMEs may not be as effective in certain phases of the innovation process such as commercialisation. As innovation requires the introduction into the market or the firm of significantly improved new products and processes, which means that indicators of implementation success can be used as output indicators, i.e., the number of new products launched or the revenues generated by them. To account for effort, inventiveness and output, some authors have created indices that provide an overall score for these activities.

The degree of novelty is often used in combination with some quantitative aspects to better discriminate the impact of innovation results (Garcia & Calantone, 2002). For example, the introduction of products or services may be either new to the firm, new to the market or new-to-the-world (Augusto & Coelho, 2009; Brentani, 2001; OECD & Eurostat, 2018; OECD, 1997). The innovation number may also include the "new-to-the-world" label (Augusto & Coelho, 2009; Lukas & Ferrell, 2000) to better discriminate firms based on the radical nature of their product and services (see Amara et al., 2004, 2008). Another classic measure comes from the sales from new products and services (Ebersberger & Herstad, 2012; Matthias Inauen & Andrea Schenker-Wicki, 2011; Szambelan et al., 2020).

As the sequence of innovation activities constitute the process that results in innovation, it makes the time required from ideation until market launch as a measure for the productivity of the innovation process (Ivanov & Avasilcăi, 2014). Nevertheless, a better innovation process quality is more important than a speedy time to market for the success of new product development (Cooper & Kleinschmidt, 1995). Thus, the relationship between time to market and innovation performance is not trivial as it depends on several factors such as the attributes of the product or service, the level of competition in the market, the maturity of the market, the resources available to the company and the company's overall strategy and so on (Datar et al., 1997). In general, a shorter time to market allows a company to quickly introduce new products or services, potentially yielding a competitive advantage over rivals, as also described as the first-mover advantage (Lieberman & Montgomery, 1988). However, rushing to market may also lead to lower-quality products or services, which can negatively impact a company's reputation and ultimately its innovation performance (Bayus, 1997; M. A. Cohen et al., 1996). Finally, firms from the same

sector with similar product with a faster time to market generally have better processes and thus greater process innovation performance (Pisano, 1997).

2.2 Web-based studies

This section introduces the field of webometrics, providing a general background on quantitative studies that used data derived from the internet. Subsequently, web-based innovation studies are showcased as mini case studies to elucidate the methods employed, their objectives, and the rationale behind them. Key insights from these investigations that are pertinent to our research are also highlighted.

2.2.1 Webometrics

In the late 90s, social scientists in information science introduced the idea of using the internet as a data source for research, adapting "informetric" methods to the content of the World Wide Web and thereby paving the way for a new area of research known as "webometrics"(Almind & Ingwersen, 1997). With the mass digitalisation of various facets of society being pushed online at the pace of its worldwide adoption, the web quickly evolved into an embodiment of the real social structures of our societies. The rising new phenomenon of the web has brought complexity and nuances that have led researchers to distinguish the differences and overlaps between informetrics, bibliometrics, scientometrics, cybermetrics, and webometrics, the differences between the last two terms being that cybermetric was more generally allowing the use of digital data but non-web internet search such as email (Björneborn & Ingwersen, 2004; Chellappandi & Vijayakumar, 2018). The maturity in the cybermetrics/webometrics research allow for the dissemination of the practices outside of the field of library and information science opening the door for other social sciences (Thelwall, 2009). Numerous webometric research endeavours have encompassed the evaluation of diverse facets of the web, including websites, web pages, sections of web pages, terms within web pages, hyperlinks, search engine outcomes, and other related elements.

The primary areas of webometrics research in the late 2000s included link analysis, blog searching, and web impact evaluation. Link analysis entails examining the network of hyperlinks found on web pages in the hopes that they may reveal the existing relationships between different organisations (Hyun Kim, 2012; Katz & Cothey, 2006; Vaughan, 2004). This approach has worked well to shed light to existing collaborative network structures with the importance of actors within

them (Minguillo & Thelwall, 2012; Stuart & Thelwall, 2006). Blog searching became an area of research as a result of the vast diversity and amount of publicly generated information pushed by individuals at any given time. This data allowed researchers to understand prevailing trends in public opinion regarding particular issues. Blog searching is more common these days by evaluating public reaction towards different topics on Twitter or other social media platform (Thelwall et al., 2011; Veltri, 2013), while Google Trends is a useful tool for keeping track of topics' relative importance at any given moment (Choi & Varian, 2012). As for web impact assessment, it is a frequency method based on the online occurrences of topics or documents allowing researchers to assess documents or ideas with a greater level of influence will tend to be referenced or discussed more frequently online, provided no other significant factors are at play (Thelwall, 2009). It could also be used to better understand the diffusion of different authors, terms, scientific theories, candidates for office, books, journals, etc.

For innovation management studies, Thelwall (2009) suggested the use of product launches online media coverage, the number of attendees at an event, the number of requests for an institution's literature or the occurrences of concepts or publications of an institution are mentioned as potential could all be potentially proxies for innovative activities.

Web analysis typically requires three main areas of web mining: web structure mining for link analysis, web usage mining and web content mining for assessing web impact, and blog searching (Miner et al., 2012). This paper is particularly interested in web content mining, a technique that converts unstructured text data from web pages into structured data which enables quantitative research. Web content analysis is derived from content analysis, a method that finds connections between textual information and the context from which the information is sourced (Krippendorff, 1980). Weare & Lin (2000) discussed the advantages and difficulties of using the web as a data source for content analysis whereas the main benefit is the proliferation of sources for multimedia data such as text content, video, audio, and images that can be used for content analysis, which would potentially lower the cost of data acquisition for this type of analysis. The quality and validity of web content analysis, however, are especially difficult to evaluate and somewhat challenging to process due to the information's lack of structure, complete decentralisation, and lack of any kind of standardisation (Herrouz et al., 2013).

2.2.2 Web-based innovation studies

Some studies used a web-based methodology, mostly with the use of keywords occurrence on a firm's websites to derive contextual insights. What follows is a description of different papers with research designs and methodologies used, which will be considered for inspiration for our own research design. The papers are structured around the main methodology used to answer their research objectives. A summary is available in Table 2-1.

Hyperlink structures analyses

Using the web for studying innovation-related topics was introduced with the assessment of complex innovation systems by building scale-independent indicators with in-links and out-links of web pages as they follow a power law distribution (Katz, 2006; Katz & Cothey, 2006). The idea was to build a directed acyclic graph from these connections and perform social network analysis to assess the underlying characteristics of the system. The indicators were showing evidence of capturing the emergent properties and scaling characteristics of innovation systems.

Hyun Kim (2012) used hyperlink structures along with semantic network analysis in which co-occurrences of terms shed light on the triple helix communication ties from a convergence theory perspective. From the Wikipedia's nanotechnology page, the author started by retrieving all nanotech-related webpages resulting in 346 websites linked to that Wikipedia article. Then, the top 20 websites featuring the largest amount of inbound and outbound links, which include the keyword "nano" in their domain name and that were identified as one of the triple helix categories were selected. Starting with 871 webpages, each page was manually classified within a Triple Helix category (Etzkowitz & Leydesdorff, 2000) i.e., government, industry and university by investigating the "About" section. Then a co-occurrence words matrix was performed on each webpage. Study findings suggest that academic and business websites focus on different aspects of nanotechnology development, academic being more related to scientific discourse and businesses on development and applications. Then, a hyperlink network analysis was performed with a webometric analysis tool called LexiURL Searcher by parsing hyperlinks from Yahoo.com. It allowed to examine the structure of web-based discourse on nanotechnology by examining network indices such as centralisation, degree centrality, and density and so on. The hyperlink networks reveal significant connections and central players, highlighting the social construction of the field.

Methodological study

Arora et al. (2016) provides guidelines for social scientists to expand their toolkit using unstructured web-based text from the Wayback Machine to access historical website data, drawing on the example of a project that analyses indicators of innovation activities and strategies in 300 US SMEs in green goods industries. It presents six steps to plan a research design based on historical data. This will be further discussed in section 3.3.3. The authors discussed three main techniques to build indicators: Keyword frequency analysis, Named Entity Recognition and Topic modelling with LDA. Furthermore, they include these methods to analyse the results: Case studies, descriptive results, linear modelling, cluster analysis and dimension reduction techniques (e.g., PCA).

Cluster analysis

Libaers et al. (2016) generated a taxonomy of business models using exploratory factor analysis from the frequencies of keywords with a sample of small, highly innovative firms focused on technology commercialisation in the United States of America (US) economy. The authors describe three sources for the inspiration of the keywords selected which are the academic literature of technology commercialisation strategies and business models, industrial literature that stems from trade and business magazines and business descriptions, and the analysis of a sample of firms' websites. A custom-made algorithm was used to query Google using the preliminary set of 87 keywords with the websites' list. Multiple tests were performed resulting in dropping keywords that were either not enough or too much used resulting in a final list of 40 keywords which were used for business model classification purpose. The keywords used are the following: accessories, capability, clinical trial, collaboration, component, consulting, contract, cost effective, custom, discovery, efficiency, enabling, expertise, government, improve, integrated, integration, machining, maintenance, modular, novel, OEM, packaging, pipeline, product performance, professor, project, proprietary, provider, R&D, research, science, solution, supplier, system integration, system solution, technology development, turnkey, university, unmatched. They ended up with a taxonomy that describes six different commercialisation paths: research organization; development stage biosciences; highly specialized component supplier; specialized subcontractor; product solutions providers; and service solutions providers.

Based on a similar keyword-based methodology, Youtie et al. (2012) mined corporate SMEs websites from both the Wayback Machine and the current website to assess changes related with nanotechnology-related product and financing information pages to flag product or funding transitions. The methods employed to derive the dimensions and keywords were not explicitly delineated. Nevertheless, 17 dimensions were used: Conference/symposium (Conferences, symposiums, trade shows), Patent/Trademark (Patent application or grants or pending, copyrights, trademarks), Research (Review of various research advances), Demonstration (Demonstration along with terms such as conferences, symposiums, trade shows), Pre-Clinical Trials (Pre-clinical trials, animal studies and related), Clinical (Clinical trials, human trials and related), Manufacturing (Manufacturing, production facility), Venture Capital (Venture capital, series A funding and related), Partnership (Research and development, R&D and partnerships), Government Contract/Grant (Contracts or grants along with the names of federal or state agencies), Investors/General Funding (Investment, investors, finance, and related), General Revenue (Revenue), Private Contract (Review of contractual relationships with another company), Customer Sales (Sales), IPO/Shareholder (Initial public offering, IPO, investors, stock) and Purchase Agreement (Purchase agreement). An idealised linear model was used to evaluate research to commercialisation paths while contextual similarities in terms of sectors, age funding source and research intensity were assessed with a hierarchical cluster analysis.

Al-Hassan et al. (2013) used the text from the legal statements such as the privacy policies and terms of use retrieved from the corporate websites of 475 Fortune 500 firms in 2011 to attempt clustering firms into the same self-reported North American Industry Classification System code (NAICS). The authors first considered the whole text after removing stop words and the highest frequency terms across documents before applying a hierarchical cluster analysis. The authors then iterated with the number of clusters while removing terms that were not considered useful for their analysis. At the end, the firms were not easily grouped into clusters that would easily be interpreted as a NAICS code. The authors point to the different role of words in different contexts a problem called polysemy. Some manual filtering and cleaning process upstream can help enhancing the results. Furthermore, it raised questions about the methodologies used, the research design but also about the self-reported nature of NAICS code which is also subject to the idiosyncratic interpretations from different firms' representatives.

Arora et al. (2013) employed a web scraping method to collect information from the websites of 20 high-tech graphene SMEs from the US, United Kingdom (UK), and China in order to create clusters of firms based on common commercialisation terms. It resulted in a corpus of 790 pages in which a ratio corresponding to the occurrences of "graphene" compared to all nanomaterials including graphene was obtained. The following dimensions were analysed: applications (battery, composite, electrode, semiconductor, display, screen, transistor, diode, sensor, detector, detection, radio, RFID, capacitor, circuit and ink), sectors (electronic, energy, medicine, aerospace, automotive, construction and environment), globalizations (world, international, or Asia) and partnerships (partner, university and distributor). Hierarchical clustering was performed to generate three commercialisation profiles of these firms: High graphene, specialized product focused; High graphene, specialized materials focused and; Low graphene, existing nano product portfolios.

Correlation analysis

Gök et al. (2015) assessed R&D activities from the corporate websites of 296 SMEs specialised in green goods based in the UK between 2004 and 2011 using the Wayback Machine for the historical data. While the authors acknowledge the inherent self-reported bias of the data source, they also assert the quality of unobtrusiveness offered by corporate website data, as researchers can observe firm behaviour without firms being aware of being evaluated. The researchers developed a research design around many sources of data including time series enterprise database with financial information such as R&D expenditures, a database of public funding supporting R&D activities, publications and patent data and the firms' websites. The research focuses on comparing the level of R&D activities[2] from the websites by considering the following keywords R&D, research, development, product development, technology development, development phase, technical development, development program, development process, development project, development cent, development facility, technological development, development efforts, development cycle, development research, development activity with indicators such as R&D expenditures, R&D financial awards, and publications and patents from the firms. The paper discusses the limitations of the Wayback Machine exposures along with operations that were

[2] Other variables and keywords were also developed such as manufacturing strategy (products, manufacturing intensity, customization, greenness), linkages (universities, partnerships, membership organizations, regional/extra-regional links), investment strategy (venture capital, investment) and policy influence (regulation).

needed such as cleaning the relevant Uniform Resource Locator (URLs). The occurrences of keywords per webpages, counts of multiple instances on the same page, the total number of pages were all computed. Normalisation with the number of webpages and the number of noun phrases were performed to take into account the differences in website size, and it was decided to use the number of noun phrases as the pages were too inconsistent. To make sure the web-based indicators were really capturing R&D activity, the authors tested a set of different rules to associate the different terms together. The authors found that the word "development" alone was leading to many false positives such as the example of "property development" which led to the recommendation to perform the operationalisation of a more precise lexical field of "development" in a context that is known to be related to R&D activities which is argued by the authors to yield the best compromise between capturing essential concepts and minimising false positives. There were no significant correlations between the metrics used to assess R&D activities which were the number of publications, the number of patents, R&D expenditures and grants with the web-based R&D indicator. It is also noteworthy that 68.9 % of the firms reported R&D activities while a minority of firms scored in the other metrics, which hinted that firms may over-represent their activities. The authors suggest that this indicator may provide complementary additional information to current methods. They finally used a data quality framework (Batini et al., 2015; Batini & Scannapieco, 2006) to compare how these variables compared with traditional ones and argued they have advantages in coverage, currency, accessibility, quantity and flexibility while having disadvantage in terms of consistency, interpretability and handling.

Descriptive results and semantic analysis

Kinne & Axenbeck (2020) did a large-scale study based on a sample of 2.4 million German firms to map the artificial intelligence (AI) innovation ecosystem based on AI-related terms. The paper provides interesting descriptive statistics and probit regression about differences in URLs and firms regions, sectors, sizes and ages. In a nutshell, very young and small firms are less likely to have websites, and sector plays a significant role. Firms in areas with low broadband availability are less likely to have a website, with an estimated 30,000 German firms not having a website due to low high-speed internet availability. From the dataset, 0.6% of all firms are patent holders, and 71.47% of them have a URL, leading to their over-representation in web mining studies. The sector-level URL coverage of patent holder firms does not show significant differences, but

highlights the varying importance of patents as innovation indicators across different sectors. Furthermore, their analysis shows a highly skewed distribution of webpages per website, with a mean of 218.8 and a median of 15. Sector differences are considerable (Public services and Mining are especially high), and there is a positive correlation between firm size and the number of webpages. The average webpage has 3295.86 characters, and there is no significant relationship between text length per webpage and firm characteristics. Most websites are primarily in German, with 88.2% classified as such, and 60.8% of non-German language webpages in English. The share of German language on a firm's website is related to the firm's sector, but not to other characteristics. The number of hyperlinks per website is also highly skewed, with a mean of 252.17 and a median of 116. The number of hyperlinks is correlated with website size, and there is no significant relationship between the number of hyperlinks per webpage and firm characteristics. Then the paper investigated Berlin-based companies and scientific institutions involved in AI, and identified firms engaged in AI using a simple keyword search with AI-related terms in German and English.

Hamilton & Cavello (2021) used the Wayback Machine to assess the evolution of the adoption of ethical products in countries. More specifically, the study examines the ethical diamond market's impact on global inequities using the Wayback Machine, focusing on the market's evolution over 30 years and the contestation of Canada's early ethical monopoly, revealing that despite diversification, the market still does not significantly benefit the most marginalised diamond-producing countries and communities. The study assesses the number of retailers offering diamonds from each country of origin in the ethical diamond market since the late 1990s, focusing on the US, the UK, and Canada, and using a "discursive market share" proxy, which represents the number of retailers selling different narratives of ethical diamond production rather than the total volume or value sold from each country. The researchers identified 168 self-described ethical diamond retailers by searching for various ethical diamond-related terms (ethical diamonds, fair trade diamonds, eco diamonds, development diamonds, sustainable diamonds, green diamonds, clean diamonds, and conflict free diamonds) on Google, Twitter, Factiva news database, and membership/participant lists of ethical jewellery initiatives. They removed retailers that only claimed Kimberley Process compliance or lacked country of origin information. Using the Wayback Machine, they captured the diamond origin information for 2019 and each year the website was archived. However, the results show the evolution of retailers' ethical diamond

offerings over time, rather than a precise record of specific changes. Although it is more remotely related to innovation, the methodology could be used for responsible innovation studies.

Antonelli et al. (2022) combined the data from websites, press releases and patent document to perform a semantic analysis to understand the relationship between appropriation and OI unfold within the context of the Open COVID Pledge (OCP) platform. The OCP, launched in April 2020, called for organisations to make their patents and copyrights freely available to fight COVID-19. The initiative has garnered support from various global companies and led to complementary initiatives worldwide. The OCP is built on four pillars: open to all organisations, addressing the unrestricted public for COVID-19 purposes, allowing for pledgor or supporter commitments, and offering a non-exclusive, royalty-free license. The participating companies, which vary in size and commitment type, have pledged patents that are mostly recent and pertain to computing or electric communication techniques. This study uses semantic methods to analyse unstructured data related to patents, by applying text mining approaches and gathering data from multiple sources. It uses a proprietary semantic tool called "TAG" to process and analyse the information. The study uses flow charts to visualise semantic connections between words and to analyse the OCP initiative's impact on the web, pledgors' communications, and patent documents. The web analysis reveals the main word as "COVID" and its top linked words (open, pandemic, open covid, free, help, access and fight and property and intellectual), are in line with the OCP Manifesto. The semantic analysis of featured patents indicates that available technologies primarily relate to real-time search and analysis of contextual information. By expanding the patent pool, a wider range of domains and applications are identified, with many focusing on devices for detecting, displaying, and monitoring virus diffusion. The methodology seems to effectively unveil the correspondence between the messaging of the initiatives with the outcomes.

Regression analyses

Shapira et al. (2016) complemented traditional bibliometric methods with web-based data to examine the characteristics and factors associated with product entry related to graphene of 65 SMEs across 16 different countries, revealing novel insights into the development and commercialisation of this rapidly emerging technology. Among many steps, the stemmed version of the words is used along with performing a named entity recognition process. Normalisation was performed based on the total number of words multiplied by 1000 to facilitate interpretations.

Three key information were provided from the websites using rules. First, the lines of business were found with manufacturing (manufactur* or produc*) or service (consult* or servic*). Then, graphene targets and value-stream position in four categories: Material Producers (GNP OR graphene* dispersion OR graphene* powder OR nano* platelets) NEARBY (develop* OR introduce* OR *manuf* OR produc* OR provide*), Intermediate products (functional* OR ink* OR master*batch*) NEARBY (develop* OR introduce* OR manuf* OR produc* OR provide*), Equipment ((equipment* OR tool* OR CVD) NEARBY (develop* OR introduce* OR manuf* OR produc* OR provide*)), Graphene-enabled final products ((consumer*) NEARBY (develop* OR introduce* OR manuf* OR produc* OR provide*)). Finally, the graphene functionality was found by extracting the word nearby the word "graphe". Cluster analysis was performed with the variables to create typologies of firms based on their market orientation. A logistic regression was then performed to determine the factors that influence product introduction. The study also argued these web-based methods could potentially deliver real-time intelligence.

Nathan & Rosso (2017) combined UK administrative data, media content and website data to develop novel measures of SME firm innovation based on new products and services launch events. They retrieved product/service launches covered by 3,740 online news along with firms' website. They used supervised learning approaches such as structural topic modelling to cluster reporting by events and for other complementary classification. They ran econometric models with web-based and conventional metrics such as patenting and trademark to predict product launches. They found that web-based indicators are a useful complementary measure that reveal additional information, with tech SMEs that are actively patenting are substantially more active in launches than non-tech SMEs.

Li et al. (2018) used a sample of 271 green goods manufacturing US SMEs to analyse the micro-level characteristics and impacts of their Triple Helix interactions on their growth. The data stems from both in-house punctual scraping of the current websites from late 2012 to early 2014 and the historical corporate website from 2008 to 2011 using the Wayback Machine. The idea was to assess the strengths of these relationships by measuring how often companies mentioned keywords related to University (university, college, institute, academ*), Industry_financial (venture capital, VC, valuation, public offering, IPO, acquisition, merger, liquidity event, seed funding, round of funding, stage of funding, investor, investment firm, private equity, bank, adventure capitalist, bridge loan, buyout, corporate venture capital, corporate venturing, deal flow, debt financing,

direct financing, drive-by deal, due diligence, equity financing, financier, full ratchet, fund of funds, institutional investors, lead investor, leveraged buy-out, liquidity event, lock-up period, management buy-in, management buy-out, master limited partnership, mezzanine debt, mezzanine financing, owner-employee, portfolio company, private equity, private placement, raising capital, recapitalization, resyndication, risk capital, investment company, secondary purchase, syndication, term sheet, underwriter, vulture capitalist), Industry_relational (partner, stakeholder, distributor, collaborat*, alliance, joint venture, agreement, supplier, parts manufacturer, member, affiliation, association, club, participation, network) and geography (named entity recognition of places). Normalisation was performed based on the number of pages. The authors mention that small firms are motivated to showcase their collaborative relationships with the entities from the Triple Helix because it sends a signal of competitiveness. Furthermore, the paper highlights that firms invest in comparable signalling practices, like sharing R&D findings in academic publications to demonstrate their expertise (Allen, 1983; Hicks, 1995; Li et al., 2018; Muller & Pénin, 2006), while being progressively employed by not only large but also SMEs (Li et al., 2015). Furthermore, the authors used four case studies from companies' history to validate the web-based indicators. Ties between government and the industry were retrieved from sam.gov where firms that were qualified for public procurement received the value of 1. An Hausman-Taylor method was used to show that Triple Helix ties with two or more categories contribute to firms' growth.

Arora et al. (2020) conducted a topic modelling analysis of US-based green goods manufacturing SMEs' websites from 2008 to 2011 to detect successful entrepreneurial agility in sensing and reacting to market shifts, a process often referred to as "pivoting" in the startup domain. The paper contends that evaluating a dimension as fluid and dynamic as pivoting is particularly challenging using traditional data sources, as they lack granularity and feature excessively large time intervals between data points. These weaknesses are a strength of corporate websites as a data source. Website variables derived from keywords include: R&D difference as a proxy for strategic sensing, Topical difference for seizing and local difference. R&D difference measures research and development intensity (lab OR laboratory OR research OR development OR R&D OR researcher OR scientist) on company websites in 2008-2009, while local difference evaluates the change in local geographic focus between 2008-2009 and 2010-2011, in relation to the firm's headquarters. Variables were normalised based on the number of collected pages per year, excluding duplicates. Due to sparse Wayback coverage, variables were grouped into two time

periods, 2008-2009 and 2010-2011, by averaging each normalised-year variable within the period. Topical differences are obtained with topic modelling with latent Dirichlet allocation (LDA) assessing the difference in similarity measures between 2008-2009 and 2010-2011. As there was endogeneity in the model, an instrument variable using web-based information was used, notably the first year a website appeared on the Wayback Machine, the difference between the number of words in the two periods and the number of redesigns since 2004. The rationale behind the chosen instruments is that, website redesigns and new word additions, which are relatively low-cost activities, can introduce new topics as firms rebrand themselves, but these changes are unlikely to correlate with the structural error term in a way that explains sales growth; robust statistical checks and the correlation table in the results section confirm the instruments' face validity. A two-stage least square (2SLS) regression model performed show that firms' growth is positively associated with topical changes, with diminishing returns.

Supervised machine learning methods

Mirończuk & Protasiewicz (2016) presents a classification method for identifying innovative domains on the Internet by analysing company web pages using a Naïve Bayes classification, featuring Bernoulli and Multinomial distribution models. The results demonstrate improved classification quality for innovative companies when combined with information retrieval techniques. Similarly, Mirończuk & Protasiewicz (2020) introduces a diversified stacked generalisation method, combining a multi-view approach and genetic algorithm, to classify companies as innovative based on their public websites. This use case describes a system that processes a company website, which can be generalised to multiple websites. The system takes a website as input and undergoes a pre-processing phase to create three datasets: 1) a textual description from the index page, 2) link labels from the index page, and 3) a "big document" constructed from selected documents on the website. Each dataset is classified using a Naïve Bayes learning method, and the results are merged for use by a meta-classifier. Various learning methods, such as k-nearest neighbours, support vector machines, and decision trees, were tested to create the final decision classification function. The authors created an original test set of 2,747 real websites (509 innovative and 2,238 non-innovative) using annual rankings and awards like Forbes rankings and Business Gazelle. Three datasets were created, each described by different features. The classification quality was evaluated using 10-fold cross-validation, measuring error, accuracy,

precision, recall, and f1-score. Statistical tests were applied to compare the performance of different models for meta-learning classification and voting classification, with procedures to vary depending on the type of feature space.

Kinne & Lenz (2021) proposes a web-based innovation indicator generation method to overcome the limitations of traditional patents and survey-based indicators by using an artificial neural network classification model trained on labelled web texts from a German Community Innovation Survey. The model is then applied to predict product innovators among hundreds of thousands of German firms, with results showing its potential as a reliable, cost-efficient, and granular addition to existing innovation indicators. More specifically, the authors experimented with various neural network architectures, including convolutional neural networks, recurrent neural networks, and traditional models, for text classification tasks. An undercomplete autoencoder-like neural network with four hidden layers and intermediary dropout layers showed the best classification performance. The network had a "bottleneck" of two hidden layers with five neurons each, forcing the learning of a highly compressed representation. Scaled exponential linear units were used as activation functions, and a single neuron with a sigmoid activation function was used in the output layer for binary classification. The Adam optimising algorithm was employed for stochastic optimisation of the network weights. On the training set, the model has an average f1-score of 0.8 in detecting non-innovator and product innovator firms. The trained model was used to predict product innovator probabilities for 685,057 firms, resulting in a mean probability of 0.253, a median of 0.203, and ranging from the lowest predicted probability of 0.029 to the highest of 0.944.

Table 2-1. Summary web-based innovation literature

Paper	Objectives	Web data	Indicators	Methods	Results
(Katz & Cothey, 2006)	Innovation systems assessment	Websites of academic or scientific organisations	In-links and out-links frequency	Social network analysis	It demonstrates the properties and scaling characteristics associated with complex systems.
Libaers et al. (2016)	Taxonomy of business models	Corporate website text content	Keyword frequency	Exploratory factor analysis	Build a taxonomy based on six commercialisation paths
Youtie et al. (2012)	Detect product or funding transitions	Current and archived corporate website text content	Keyword frequency	Idealised linear model Hierarchical cluster analysis	Main commercial pathways were retrieved through ILM, but with notable divergences and contextual similarities revealed with cluster analysis.
Hyun Kim (2012)	Explore the structure of web-based discourse on nanotechnology	Websites related to Wikipedia's nanotechnology page	-Triple Helix manual classification -Co-occurrences of terms	Semantic network analysis	Patterns of language used by different triple helix were identified and the social network structure revealed.

Table 2-2. Summary web-based innovation literature (continued)

Paper	Objectives	Web data	Indicators	Methods	Results
Al-Hassan et al. (2013)	Explore whether firms from the same NAICS code can be associated together	Text content from legal statements from corporate websites	Frequency of terms	Hierarchical cluster analysis	Clusters are not easily interpretable because of polysemy. Disambiguation may be applied upstream through manual operations. Self-reported information does not easily transpose into standard categories.
Arora et al. (2013)	Explore commercialisation strategies from SMEs	Corporate website text content	-Occurrences of graphene vs all nanomaterials -Keyword frequency analysis by webpage	Hierarchical cluster analysis	Three commercialisation profiles were retrieved.
Gök et al. (2015)	Assessed R&D activities from the corporate websites	Current and archived corporate website text content	Rule-based keyword frequency analysis by webpage and noun phrases	Correlation analysis	No significant correlations between the metrics used to assess R&D activities suggesting novel and complementary information. Data quality framework analysed.

Table 2-3. Summary web-based innovation literature (continued)

Paper	Objectives	Web data	Indicators	Methods	Results
Arora et al. (2016)	Provides a methodological resource for using the Wayback Machine for social sciences studies	Archived corporate website text content	-Keyword frequency analysis -Named Entity Recognition -Topic modelling with LDA	Case studies, descriptive results, linear modelling, cluster analysis and dimension reduction techniques (e.g. PCA).	Six steps are presented: sampling, organising, and defining the boundaries of the web crawl, crawling, website variable operationalisation, integration with other data sources, and analysis.
Shapira et al. (2016)	Investigates firm development and commercialisation of graphene by examining firm characteristics, relationships, value chain positioning, and factors influencing product entry	Corporate website text content	-Stemmed Keyword frequency analysis by the number of words -Named entity recognition	-Cluster analysis -Logistic regression	Provide a broad picture of the introduction and scaling of downstream graphene products, along with financial and market support.
Nathan & Rosso (2017)	Evaluate innovation performance with novel data sources and methods, notably through product launches events	-Online news -Corporate websites	- Product launches -Events	-Cluster analysis -Structural topic modelling -Regression models	Web-based indicators are a useful complementary measure that reveal additional information, tech SMEs that are actively patenting are substantially more active in launches than non-tech SMEs.

Table 2-4. Summary web-based innovation literature (continued)

Paper	Objectives	Web data	Indicators	Methods	Results
Li et al. (2018)	Assess the Triple Helix impacts on SMEs growth	Current and archived corporate website text content	Keyword frequency analysis by pages	- Panel data regression with Hausman-Taylor method	-Authors used four case studies from companies' history to validate the web-based indicators - Triple Helix ties with two or more categories contribute to firms' growth
Arora et al. (2020)	Measure the ability to recognise and react to market shifts by startups	Archived corporate website text content	- Differences in keyword frequency -Topic model with LDA	2SLS regression model	Firms' growth is positively associated with topical changes, with diminishing returns.
Kinne & Axenbeck (2020)	Mapping the Berlin AI innovation ecosystem	Corporate websites	- Keyword frequency analysis -Hyperlink frequency	-Descriptive results -Correlation analysis	Many descriptive results related to corporate websites characteristics. -Identified firms engaged in AI
Hamilton & Cavello (2021)	Assess the evolution of the adoption of ethical products in countries	Archived corporate website text content	- Keyword frequency analysis	- Search method -Descriptive results	Results show the evolution of retailers' ethical diamond offerings over time, rather than a precise record of specific changes.

Table 2-5. Summary web-based innovation literature (continued)

Paper	Objectives	Web data	Indicators	Methods	Results
Antonelli et al. (2022)	Verify is the Open COVID Pledge manifesto matches the outcomes	Text from manifesto and patent related to the OCP	-	Semantic analysis	The methodology seems to effectively unveil the correspondence between the messaging of the initiatives with the outcomes.
Mirończuk & Protasiewicz (2016)	Use a supervised method to predict innovativeness	Corporate websites	Supervised machine learning model	Naïve Bayes classification Information retrieval techniques	Limited results, but shows potential in classifying firms based on unstructured data.
Mirończuk & Protasiewicz (2020)	Use a supervised method to predict innovativeness	Corporate websites	Supervised machine learning model	-Diversified stacked generalisation method -k-nearest neighbours -Support vector machines -Decision trees	High F1-scores show promising avenue
Kinne & Lenz (2021)	Use a supervised method to predict innovativeness	Corporate websites	Product innovation predictor	Neural network tools	Predict product innovator probabilities for 685,057 firms, resulting in a mean probability of 0.253

All these cases inspired to some level the research put forward in the thesis, which will be presented in Section 3. Most papers discuss briefly the benefits and limitations of websites as a source of data, and we aim to contribute to this discussion with an in-depth analysis of corporate websites as an object of study. This is the focal point of the next section.

2.3 Assessing corporate website as an object of study

As briefly discussed in the introduction, nowadays a firm's corporate website is the official communication and public relationship vehicle at its disposal, and as such is crucial for lead generation, sales and marketing, talent acquisition, investors' relationship, and so on. Therefore, the messaging strategy used represents its core intention towards each key stakeholder. Furthermore, firms' representatives select information and only reveal what they see as useful to meet their own idiosyncratic corporate objectives, which is nothing short of a very strong self-reporting bias. In the last few years, examples of firms engaged in the diffusion of misleading information were numerous as it will be presented in the next few pages, which may even question how corporations as a whole are honest in their information revealing. All these considerations may even question the epistemic value of corporate websites as a valid data choice.

This section aims to provide a theoretical background to: 1) enrich our analysis of the interpretation of the use of innovation language in corporate websites through established signalling theories literature; 2) assess the reliability of corporate website content as a source of trustworthy information; and 3) provide a theoretical ground for possible explanations when discussing our results. The section starts with a little section about the field of semiotics and the definition of signal. Then, an overview of signalling theories is presented followed by a review of signalling dynamics in language. Finally, an analysis of the self-reported nature of the website is offered by investigating relevant literature related to corporate websites.

2.3.1 Semiotics

The study of signs and how they are used to convey meaning stems from the field of semiotics mostly based on the work of Ferdinand de Saussure, Charles Sanders Peirce, Roland Barthes, Roman Jakobson, Charles Morris and Umberto Eco (Yakin & Totu, 2014). According to Charles Sanders Peirce's theory of semiotics, anything can be considered a sign as long as it can represent something else. This relationship, known as the sign action or semiosis, consists of three

interrelated elements: the sign (signifier), the object that the sign represents (signified), and the interpretant or the meaning or effect that the sign has on the mind of the person interpreting it (Atkin, 2022). For example, the word "tree" is a sign (signifier) that represents an object, a tree (signified) which has a specific meaning to the reader's mind. According to Sebeok (2001), there are six major types of signs: a symptom, which is a biological warning sign; a signal, which communicates and elicits a response from the receiver; an icon, which represents an object or concept directly, i.e., that looks somehow like its referent; an index, which represents an object or concept indirectly, e.g., smoke pointing to fire; symbols, which refer to their referent by a subjective, human-made representation; and a name, which assigns an identity to an entity. It is important to note that signs, more often than not, share properties that may enable the interpretation from more than one type of sign at the same time, which makes this typology not mutually exclusive, and somewhat fluid (Sebeok, 2001).

As it will be explained next, the index, symbol, signal and name are the types of signs that are present in corporate website textual content. Words symbolise the object to which they refer, for instance, the word "tree" is a symbol for a "tree". Symbols can be analysed and interpreted to derive scientific knowledge. For instance, historians, archaeologists, and anthropologists used an epistemic approach called evidential paradigm to derive knowledge about Palaeolithic human societies from cultural artefacts (Bertozzi, 2007; Ginzburg, 1989).

The index is another type of sign with an interesting epistemic property, as it implies a causal relationship to the object it signifies. However, the reliability and validity of the causal relationship between the signifier and the signified is obtained through experimentation. In that sense, we can see science as an enterprise to systematically interpret all the indices from nature to deduce testable hypotheses into theories and laws. Corporations, cities, and people are all idiosyncratic entities, each having their own name, which we should be able to find on a corporate website. Finally, the signal, being the more complex sign, requires more precision. It is defined as:

> […] a sign which mechanically (naturally) or conventionally (artificially) triggers some reaction on the part of a receiver. This is in accordance with the view that 'signals may ... be provided by nature, but they may also be produced artificially' (Kecskemeti 1952: 36). (Sebeok, 2001, p. 44)

Signals can be either unwitting, meaning the sender is unaware of the information being transmitted, or witting, meaning the sender is aware and intentionally transmitting the information. The corporate website is a complex semiotic product with design elements, several types of different media and text content, which can all be interpreted by analysing the specific names, icons, symbols, indices and signals. However, we view firms as agents with their own agenda and objectives within an ecosystem of stakeholders, and in this context, a corporate website serves as a means to elicit at least some level of desired reaction from its stakeholders, commonly called in communication theory as ostensive behaviour (Dégh & Vázsonyi, 1983). Therefore, we see sending signals, or signalling, at the core of all the firm's sign emissions. Thus, for the purpose of this paper, we will consider all the signs emitted by firms as signals.

2.3.2 Signalling theories

First derived from evolutionary biology, signalling theory aims to understand the use of signals within and between species in order to understand how they achieve successful communication (Carpenter, 1969; Grafen, 1990). Individuals with conflicting interests have different set of characteristics that allow them to play the evolutionary game, i.e., mating with the fittest and surviving from the fittest. Thus, the individuals engage in signalling games to communicate their better fitness quality, which should represent their superior characteristics. Although it may seem advantageous for a weaker individual to mimic the signals corresponding to the characteristics of the fittest individuals due to the nature of this dynamic, misleading behaviours are more the exceptions than the rules in the natural world (Godfray, 1991; Grafen, 1990; Zahavi, 1975, 1977).

Therefore, signalling theory, supported by game theory models, suggests that signalling games tend to converge to a strategic equilibrium, a state of signal stability where only the reliable signals remain effective over time because the receivers were able to detect that they were honest (Andersson, 2019; Summers et al., 2015). As a result, honest signalling assumes a correlation between the intensity of a signal and the quality it advertises (Számadó, 2011). Three types of honest signals are discussed in the literature: 1) the index signals, which provide genuine information about the sender because they can only be sent as a result of that inherent quality (although it was shown to behave in game theory models exactly as costly signals, see Biernaskie et al., 2014), 2) the costly signals, which showcase traits that cannot be copied or simulated i.e., honest signals that are too costly to fake (Zahavi, 1975, 1977), and 3) conventional signals, which

are low-cost signals designed in nature for increased detectability, discriminability and memorability of the consequence (or cost) of an interaction, such as defence mechanisms from individuals with very bright colours to dissuade potential predators (Guilford & Dawkins, 1995).

The theory of costly signalling has also interesting interpretations for social scientists as it proposes "*that animals (including humans) may send honest signals about desirable personal characteristics and access to resources through costly biological displays, altruism, or other behaviours that would be hard to fake"* (McAndrew, 2018). Economists and psychologists applied this theory to individuals and organisations, in which one can demonstrate its fitness or quality to others by incurring a cost that proves its ability to provide value to the receiver's eye, which can lead to superior treatment and status (Nelissen & Meijers, 2011; Tchernichovski et al., 2019). All different theories of signalling predict that over time, effective communication between senders and receivers is achieved through honest signals which carries a real and significant cost and the dynamic is reinforced through social rewards and punishments (Bliege Bird & Smith, 2005; Cronk, 2005; Nelissen & Meijers, 2011; Tibbetts & Izzo, 2010). Indeed, Számadó (2011) argued that honesty in signalling is maintained not by the actual cost incurred by honest signallers at equilibrium, but rather by the potential cost of lying and cheating. However, in human signalling dynamics, the level of social capital and social prominence may induce a halo or horn effect in the perception of other people which affect their signalling cost function, resulting in low-quality people with higher status potentially benefiting from a "reputation shield" and high-quality people potentially stuck in a "reputational poverty trap" (Dumas et al., 2021).

Similarly, signalling theory in labour economics aims to properly assess a situation by interpreting market signals in the presence of information asymmetry (Spence, 1973). In the market, information asymmetries are purposefully enabled by the actors with competing interests, such as corporate officials and the firm's stakeholders (Bergh et al., 2014). When assessing a firm, stakeholders may turn observable actions to signals as a way to gather more information about the firm's unknown characteristics to predict future actions (Spence, 1974). Interpreting the signals may facilitate the process of relating the existing knowledge about the firm with the information stakeholders want to acquire (T. Miller & Del Carmen Triana, 2009). The core idea of signalling theory is to establish a "separating equilibrium" in which market participants use signals to address an information problem (e.g., such as choosing the best firms to invest in), with the costs of credible signals being inversely related to the sender's quality so that receivers can use a Pareto

optimising logic to select the best senders based on the expected quality of the signal, which is confirmed through subsequent experiences (Bergh et al., 2014). Ideally, a signalling theory application has pareto optimal solutions when equilibrium is reached, meaning that changing the expectations of one party would result in a negative impact on its interests.

For example, consider a scenario where only truly innovative firms are able to secure investment from a very selective VC firm. In this case, receiving funding from this VC serves as a signal that allows other potential investors to differentiate between high- and low-quality firms in terms of innovation. In this case, the reputation of that VC firm maximises the quality of the signal, and thus, bring the signal acquisition cost down to other investment firms (i.e., it took on the cost of signalling). The value of this signal is confirmed when the firms that receive investment from VC subsequently outperform their peers in terms of innovation success. An alternative interpretation of that particular example is that developing a reputation over time is a costly process, and taking action from that position sends costly signals to other actors. If the VC makes an evaluation mistake, its reputation would suffer; therefore the signalling would be slightly less costly. If the VC makes more and more evaluation errors, its reputation would be damaged over time, causing its signalling to become less and less costly and, therefore, reducing its reliability as a valid source of information (Herbig & Milewicz, 1995). The VC has an incentive to keep its signal costly, not because it cares about its ability to communicate valid information about the innovation market to others, but because it would lose significant financial resources along with its long-built reputation, which would decrease its favourable treatment by others and thus limit future market success (Számadó et al., 2021).

2.3.3 Signalling from language

Language played a crucial role in human evolution because it enabled individuals to communicate and cooperate more effectively, hence facilitating the exchange of precise and accurate information on a larger scale (Giardini et al., 2022). Through the means of sound symbols called phonemes, language enabled high-volume communication which drove the marginal cost of the production of signalling towards zero, making it a “conventional” signal (Lachmann et al., 2001). Furthermore, as human language uses words, arbitrary symbols that are put together logically and subjectively into sentences with the intention to convey information, how one derives meaning from them is also arbitrary (Pinker, 2003). Because using words to build sentences is basically a signalling

commodity that is prone to many errors in interpretation, researchers typically did not consider costly signalling to be a proper framework in explaining how honest signalling in language reached stability through social dynamics (Bergstrom & Lachmann, 1998; Grassly et al., 2000). Lachmann et al. (2001) proposed that conventional signals are used in communication whenever there are coincident interests or aspects of conflicting interests that can be proven or verified easily while costly signals are used in other situations such as one-time interactions, communication about unverifiable properties and when the potential benefits of deception outweigh the expected costs of the community or individual punishment. As a result, most costly signals from humans are in the form of non-verbal communication, such as conspicuous consumption demonstration (Nelissen & Meijers, 2011; Veblen, 2017), emotional expressivity which is seen as an index for trustworthiness (Boone & Buck, 2003) and reputation risking which is built from an established reliability to meet expectation (Herbig & Milewicz, 1995; Számadó et al., 2021). Chaudhry & Wald (2022) proposes three types of costly signalling to deliberately use in conversations that would improve perceived honesty: 1) difficult-to-fake signals through genuine emotional expressivity, or the cost of emotional vulnerability; 2) verifiable signals can be validated from unbiased sources, or the cost from having done the research and remembering it; and 3) self-sacrificing signals that shows deeper personal involvement , or the cost in terms of time, effort, emotional energy or money or in reputation.

As mentioned earlier, honest signalling is especially enforced in situations when the social cost of dishonesty is high (Számadó, 2011). As such, language may help to propagate the sharing of information about an individual's reputation, which can serve as an incentive as a whole for honest communication and a deterrent against deception (Lachmann et al., 2001). In case of conflicting interests, one may choose language with reduced precision and/or accuracy in order to maintain the "honesty" of the signal from the receiver's perspective while avoiding sharing potentially damaging information and incurring social punishment (Bergstrom & Lachmann, 1998). If there is a significant discrepancy between the interests of a sender and receiver, the situation will lead to the famous Arrow's information paradox we described previously (Arrow, 1962). However, following evolutionary principles, it is believed that over time the efficiency in communication should improve (Nowak et al., 1999; Nowak & Krakauer, 1999). Lachmann et al. (2001) adds two important precision: 1) when the cost of a signal is imposed by the receiver or there are fewer strategic concerns about the information being shared, transparency can be increased and

unnecessary costs can be avoided by avoiding unhelpful communication patterns; and 2) in situations where social costs are the main regulatory force in honest signalling, misunderstandings become costlier and may lead to a preference for clearer and more reliable signals, which will converge towards the use of careful and redundant signals. Finally, their research suggests the use of complete sentences instead of words as an interpretable signal unit for language signalling system analysis.

2.3.4 Assessing the self-reported nature of corporate website

As just discussed, words are cheap conventional signals that cannot be used *per se* to determine whether a signal is honest. However, corporations have very strong incentives in maintaining an excellent reputation and image, through the careful coherence between a firm's own corporate identity and corporate communications (Chun et al., 2005; Gammoh et al., 2006; E. R. Gray & Balmer, 1998; Markham, 2015). This means that a firm's official communications may be seen as costly signals as it put its reputation on the line. Moreover, disclosure of sensitive information can be especially costly to reveal, and therefore it has been demonstrated it is best for a firm to only disclose information that would only improve the firm's value, especially among the firms with the best probability of success (Song Shin, 2003; Verrecchia, 1983). This suggests that firms are more likely to disclose information on their website that portrays them in a positive light, rather than revealing true aspects of their operations. This is what one can reasonably suspect when first thinking about the self-reporting bias induced in corporate written content. However, given our understanding of the dynamic that leads toward honest stable communication over time, and the fact we consider official communications as costly signals for reputation building, the most important question remains: on the whole, can we trust the information selectively disclosed by the firm?

One way to answer this question is to look directly at glaring examples of misleading signalling by corporations. As awareness of climate change, biodiversity loss and social injustice grows among the public, firms perceive increasing market pressure to reduce their environmental impacts and to engage in diversity, inclusivity and equity initiatives (Deng & Cheng, 2019; Dilevko, 2019). From a signalling perspective, a sender that communicates activities or initiatives based on moral excellence (or condemning actors of doing damaging activities) in order to improve or maintain one's reputation is called 'virtue signalling' (P. R. Berthon et al., 2023; Tosi & Warmke, 2016;

Westra, 2021). These efforts seem particularly evident when a public company with detrimental societal and/or environmental impacts attempts to send misleading signals to preserve their reputation (Dunlap & McCright, 2011; Fitzpatrick et al., 2022; Guix et al., 2022; Siano et al., 2017). This has been especially observed in the long history of 'greenwashing', notably when oil and gas giants engaged in minor activities that appear supportive of sustainability, such as investing in renewable energy or biodiversity protection, while simultaneously funding and lobbying for efforts to mislead the public about climate change, in an attempt to create a positive image about themselves and their operations (Dunlap & McCright, 2011). As we saw previously, a predictable strategy in case of conflict of interests is to reduce the precision and accuracy in the language, which is indeed one of the strategies reported in greenwashing practices (Guix et al., 2022; Siano et al., 2017). However, when the implication of a claim that attempt virtue signalling is not representing reality, and therefore the contradiction is seen as dishonest and manipulative, we can expect a cost on the sender's reputation. As we saw earlier, detected misleading signals come with social punishments, and when the cost penalty is high enough (by decreasing product demand, lowering share valuation and reducing market opportunities from damaged reputation), the economic dynamic should maintain the stability of honest corporate signalling. Unfortunately, the reality does not seem to perfectly match this dynamic. As an example, in the case of the Volkswagen scandal there was a 40% drop of the stock valuation in the following weeks, but the long-term outlook of the firm did not really change as the consumers and media attention span runs short (Jung & Sharon, 2019). As we saw earlier, it is possible also that significant brand benefits from Halo effects which may shield them for hefty reputational cost. However, Lyon & Maxwell (2011) predicts as formal environmental auditing increases in the future, the cost for greenwashing should increase and thus, reduce greenwashing occurrences. Because some firms clearly see the benefits of sending misleading signals outweighs the potential cost of being caught, the question about the current overall level of trustworthiness of corporate websites as a whole remains legitimate.

Moreover, there is evidence that virtue signalling is indeed effective, as it positively influences stakeholders and consumers' attitude as they encourage any form of social, environmental or governance initiatives, whether it does directly address areas of deficiency within the core business activities or not (S. G. Gray et al., 2020). This suggests we should see an increase in these activities, again within the core business activities or not, but also firms may determine that it is more

valuable for them to increase the signal intensity for their more 'virtuous' actions versus everything else they do, especially when operating in socially and environmentally sensitive industries (Cho & Patten, 2007). This suggests that even when firms might be sending honest signals, they might not be presenting an honest overall picture of the firm's most important characteristics, even after selecting only positive attributes.

Meanwhile, authenticity and transparency seem to be gaining literature evidence as the best approach to attain strategic goals (Chen et al., 2022; Fritz et al., 2017). This is not a new phenomenon as individuals have been seeking authenticity for hundreds of years (Grayson & Martinec, 2004). Moreover, the internet's ability to facilitate the communication and spread of information have increased market transparency, leading to informed consumers who expect consistency and authenticity from their brands and are less tolerant of insincere brand behaviour (Eggers et al., 2013; Holt, 2002). Furthermore, as the corporate website is the official storytelling vehicle of a firm, it is also signalling its identity and value to future and current employees (Nyagadza, Kadembo, et al., 2021). For recruitment purposes, the firms need to showcase their identity and important beliefs in their website. As employees are sensitive to their firm's signalling, they are more likely to align with a consistent and reliable signal (such as a corporate branding story), but an inconsistent or unreliable signal may lead to internal conflicts within the organisation due to cognitive dissonance during communication and sense-making processes (Nyagadza, Kabonga, et al., 2021). Thus, this self-reporting bias in situ may also represent an opportunity to assess the subjectivity of a firm toward itself and what they want to be seen as the key to their success. As mentioned previously, there is conceptual support that a firm's website may potentially be a vitrine to its culture, i.e., an outsider will notice cultural artefacts (Schein, 2004). However, even when a firm discloses information that may be viewed internally as authentic, it remains a subjective signal. For instance, a company may mention their innovation practices, or how good they are at developing new technologies, but these "insights" comes from their own biased idiosyncratic perspective.

In light of the analysis provided from this brief review of the literature, whether researchers may use corporate websites as a valid data source to completely replace such an objective epistemic tool as a questionnaire-based survey seems doubtful at the very least. However, as biased as this self-reporting is, we consider a corporate website as a centralised signalling platform where one can decode the firm's intentions and priorities in terms of communication with its key stakeholders.

As a result, important information may be gathered and be useful to researchers. For all the reasons mentioned above, it is especially challenging to accurately assess the quality of the signals obtained from corporate websites. As both elements of market strategic signalling (which may include misleading signals) and corporate culture signalling (which include honest, but subjective information) are likely to coexist in every firm's website, distinguishing how we can accurately use these signals for research purposes will be discussing next.

CHAPTER 3 RESEARCH DESIGN

3.1 Research question and objectives

Several successful innovation management quantitative studies have utilised web content analysis, using corporate web-based data to build new business model taxonomies (Libaers et al., 2016), assess product evolution (Shapira et al., 2016; Youtie et al., 2012), commercialisation patterns (S. K. Arora et al., 2013), product launches (Nathan & Rosso, 2017), explore Triple Helix impacts from firms' perspectives (Li et al., 2018), assess startups' pivots (S. K. Arora et al., 2020), map entire innovation ecosystem (Kinne & Axenbeck, 2020), and predict innovative firms based on supervised methods (Kinne & Lenz, 2021; Mirończuk & Protasiewicz, 2016, 2020). Other studies have attempted to evaluate specific innovation-related characteristics using firm websites, comparing their results with those obtained from traditional sources. For instance, Al-Hassan et al. (2013) attempted to match firms based on their NAICS code, but with limited success. Gök et al. (2015) compared a web-based R&D indicator created through keywords frequency analysis with traditional questionnaire-based R&D indicators, finding no significant correlation, suggesting that the indicators were not measuring the same concepts. This lack of direct correlation between information on corporate websites and traditional metrics raises questions about the validity of web-based indicators as evidence of innovative activities, highlighting the need for further investigation into the concepts they capture.

While the analysis in Section 2.2 raises concerns about the epistemic value of corporate websites as an objective data source, the signalling perspective may still offer valuable insights. The primary challenge lies in accurately evaluating the quality and meaning of signals from corporate websites. This gives rise to the central research question of this thesis:

To what extent does the innovation-related language found on corporate websites correspond with objective reality?

More specifically, can the innovation-related signals observed from firms' website translate as innovation measures for quantitative studies? If so, how? This exploratory research is structured around three primary objectives: firstly, to create and validate web-based indicators using questionnaire-based data; secondly, to present empirical evidence supporting credible interpretations of what is being measured; and thirdly, to evaluate the potential benefits and limitations of these web-based indicators and suggest possible applications.

3.2 Propositions

Due to the limited literature exploring the relationship between innovation signalling and actual innovation dimensions, we put forth multiple propositions to examine our research question from diverse perspectives. These propositions are organised around three validation approaches with both web-based and questionnaire-based data, which also correspond to the papers presented: first, statistical construct validity from comparing innovation factor measurements; second, using evidential paradigm and signalling theory to assess the web-based innovation evidence meaning; and third, using predictive validity by assessing how web-based measures match with the expectations from the literature.

Certain well-established innovation factors have been identified as advantageous for firms to publicly display and are measurable to some extent through keyword analysis. The factors considered include R&D (Allen, 1983; Gök et al., 2015; Hicks, 1995; Li et al., 2015; Muller & Pénin, 2006; OECD, 2015), IP (Kinne & Axenbeck, 2020; Teece, 1986), collaboration (Cropper et al., 2008; Li et al., 2018), external financing (Kor & Mahoney, 2005; Li et al., 2018; Wellalage & Fernandez, 2019), innovation culture (Herzog, 2011) and OI (Bogers et al., 2017).

The first category explores the differences in measurements of classic innovation factors, such as R&D, IP, collaboration, and external financing, between web-based and questionnaire-based measures. This approach allows for a more accurate understanding of what can be inferred from web-based indicators and refines our expectations regarding the relationship between these web-based indicators and objective measures of innovation.

Companies may use language on their websites to provide insights into their actual activities. The more frequently a company uses terms related to a particular factor, the stronger the signal is considered for that factor, suggesting a higher likelihood that the firm engages in activities associated with that factor. Hence, this proposition:

Proposition 1.1*: The more words related to a factor, in our case R&D intensity, IP, collaboration and external financing, that are used on a firm's website****, the more a firm is likely to perform activities related to that factor.***

Establishing a connection between keywords found on corporate websites and a company's concrete actions can be challenging, as some empirical evidence suggests that this may not be the case (Gök et al., 2015). Corporate websites often aim to showcase a company's best qualities to

both internal and external stakeholders, raising concerns about the objectivity of the information provided, which could be subject to self-reporting bias. However, it is reasonable to assume that the information conveyed on a company's website accurately reflects the firm's intentions regarding various concepts it wants to present to the public. From webometrics, web impact assessment suggests that online documents or ideas or concepts with greater importance offline will generally be cited or discussed more often online, assuming no other major factors are involved (Thelwall, 2009). Thus, the content on a website can provide a general sense of the importance of a particular factor for the firm. Thus, this proposition is suggested:

Proposition 1.2*: The more words related to a factor, in our case R&D intensity, IP, collaboration and external financing, that are used on a firm's website,* ***the more important a firm considers that factor.***

The second category explores whether innovation evidence in firms assessed with web-based indicators is associated with questionnaire-based evidence of innovation results and/or investment received, an indirect measure of innovation activities.

The rationale for using innovation evidence instead of precise innovation measurements is to attempt to achieve a closer epistemic correspondence between subjective web-based signalling and objective questionnaire-based information. For instance, a small biotechnology firm working on a single product and a large company with a diverse product portfolio may have very different R&D expenditures relative to their income. However, they might justifiably use similar innovation-related language on their websites. Assuming the firm has no misleading intentions, such innovation signalling could represent a self-reported idiosyncratic level of evidence of innovation.

To better illustrate this relationship, consider innovation-related signals as the smoke that indicates the fire, representing real innovation. Seeing smoke in the sky does not help to specifically determine the various attributes of the fire, such as its temperature, brightness, or combustion rate. However, we know that there is no smoke without fire. Likewise, while web-based innovation-related language may not directly correspond to specific innovation metrics, we can interpret the terms used as evidence of innovation. Archaeologists and anthropologists used a similar epistemic approach when they analyse cave painting to derive knowledge about Palaeolithic human societies. This meaning-making process is called evidential paradigm (Bertozzi, 2007; Ginzburg, 1989).

To determine if this innovation evidence obtained from a firm's website can be generalised as a valid, unbiased metric for real evidence of innovation, it would need to significantly be associated with an objective source of innovation evidence, such as classic innovation performance metrics derived from questionnaire-based survey data. As previously mentioned, established innovative firms are defined as those that reported innovation results (OECD & Eurostat, 2018). If firms with a stronger signal also have stronger objective evidence of innovation results, it would suggest that the stronger the signal, the stronger the evidence of innovation. Thus, the following proposition is tested:

Proposition 2.1*: A stronger innovation signal on a firm's website is positively associated with greater* ***evidence of innovation.***

On the other hand, the smoke represented by innovation-related language may point to a different fire than the one we expect. Instead of being related to evidence of innovation results, the smoke may be related to evidence of innovation activities. Indeed, innovation-active firms are those that engage in innovation activities without yet reporting innovation results (OECD & Eurostat, 2018).

By analysing the intentions behind the use of innovation-related signals, we can derive meaningful information from a signalling theory perspective. As previously established, the lack of external financing has long been recognised as a significant obstacle to innovation (Harhoff & Körting, 1998; Kalil, 2005; McNeil et al., 2007). Therefore, given the pressing need for such financing among innovation-active firms, it is plausible that web-based innovation-related signalling plays a crucial role in a firm's ability to attract external financing.

Nowadays, a corporate website is often one of the first sources of information that an investor uses to screen potential investment candidates. If a corporate website effectively signals a firm's potential growth, the investor may initiate a process that involves meeting with business leaders and requesting privileged information to analyse all parameters involved in making investment decisions. Consequently, if a firm's website facilitates this process and investment is received, the information on the website is validated by a third party. Moreover, if a firm's website uses innovation-related language and receives investment, the language should be representative of the innovation dimensions revealed by the firm. By investing, an investor assumes the cost of the signal, taking on the risk of potentially losing financial resources and their reputation. This implies

that investments received by innovation-active firms may be regarded as a costly signal for evidence of innovation activities. Furthermore, obtaining other forms of external financing, such as public grants or debt, requires an external decision-maker who allocates funding based on strict guidelines.

Investment can also enhance a firm's innovation capabilities, which is likely to impact its innovation outcomes. In one way or another, external financing received from a third party, such as a private equity firm, a bank, or public agencies, may serve as indirect evidence of the firm's innovation activities. Based on this analysis, the following proposition is tested:

Proposition 2.2*: A stronger innovation signal on a firm's website is positively associated with the* ***firm's receiving more investment****.*

The third category builds on the signalling perspective by examining how different typologies of firms are associated with stronger innovation signals. This analysis may enable the identification of the key similarities and differences between firms' innovation profiles and web-based signalling. These associations are based on the assumption that the innovation signal refers to real innovation activities, thus assessing the predictive validity of these indicators.

The association between evidence of innovation and the signal is assessed another time but this time in the context of exploring different levels of evidence of innovation as firms with different innovation profiles may have different signalling patterns. Nevertheless, the assumption is the same, the more a firm is signalling, the more likely it has a higher level of innovation evidence. Furthermore, the literature has established a positive correlation between innovation and a supportive culture (Dobni, 2008). As various market orientations differentiate firms based on their innovation profile showing different outcomes in terms of their performance (Aldas-Manzano et al., 2005; Hurley & Hult, 1998; O'Cass & Viet Ngo, 2007). These orientations emerge from processes and beliefs, all core elements from an organisation's culture (Schein, 2004).

Collaboration has a positive impact on various stages of the innovation process, including idea creation (Koen et al., 2014), R&D (Johnson & Filippini, 2009; Jordan & Lowe, 2004; Parker, 2000), and the performance such as the firm's competitiveness (Hagedoorn et al., 2000; Roja & Nastase, 2013), and other key innovation performance indicators such as patent numbers, sales growth, and the return on innovation sales (Arvanitis, 2012; Belderbos et al., 2004; Carboni, 2013).

OI describes practices that aim to optimise the innovation process by actively acquiring (outside-in) and distributing (inside-out) intellectual assets such as technologies, know-how and knowledge between firms (Chesbrough, 2003a; Chesbrough & Bogers, 2014). There is substantial evidence that acquiring external knowledge or technologies to augment internal R&D processes has a favourable effect on a firm's innovation performance (W. M. Cohen & Levinthal, 1990; Henkel, 2006; Laursen & Salter, 2006b). There is some indication that sharing knowledge and technologies can have beneficial impacts on a firm's innovation performance (Héroux-Vaillancourt & Beaudry, 2018). The concept of OI has been a topic of discussion in the literature, with some scholars contending that it is a rebranding of longstanding practices used by successful firms (Oakey, 2013; Trott & Hartmann, 2009), while others emphasise the difficulties and expenses of implementing OI practices, including the management of IP and the establishment of relationships with external partners (Birkinshaw et al., 2011). It is noteworthy that the connection between OI practices and firm performance differs across studies (Bogers et al., 2017), and that the advantages of OI appear to reach their maximum potential beyond a certain threshold before leading to diminishing returns (Fey & Birkinshaw, 2005; Laursen & Salter, 2004, 2006b). Overall, we anticipate that firms utilising OI practices will have a favourable impact on innovation. Additionally, the inclusion of OI and related concepts in the most recent version of the Oslo manual indicates their contemporary relevance (OECD & Eurostat, 2018, p.134).

R&D intensity is a classic indirect metric for innovation (Becheikh et al., 2006), which measures the extent to which a firm invests in developing its own inventive capabilities (Griliches, 1990, 1994, 1998; Hausman et al., 1984; Hitt et al., 1997) and improving its absorptive capacity (W. M. Cohen & Levinthal, 1990). Furthermore, a number of studies have demonstrated a positive association between R&D intensity and innovation performance (Baysinger & Hoskisson, 1989; W. M. Cohen & Levinthal, 1990; Deeds, 2001; Greve, 2003; Hall, 1990; Parthasarthy & Hammond, 2002). Although R&D intensity was previously regarded as a measure of innovation, it has been criticised for its limitations, as innovation can arise from sources beyond the R&D department (Michie, 1998; OECD & Eurostat, 2018), especially as companies increasingly adopt OI practices. Despite the limitations, R&D intensity still holds significant value as a component in comprehending innovation performance, particularly for technology-oriented firms (OECD & Eurostat, 2018).

Companies can achieve a competitive advantage by obtaining exclusive rights to commercially leverage their technologies, branding, copyrights, industrial designs, and so forth (Teece, 1986). Patents are typically a consequence of R&D efforts and are also frequently used as a measure of a company's inventiveness (Coombs et al., 1996; Flor & Oltra, 2004; OECD & Eurostat, 2018). Patents allow firms to enhance the return on their investment in technology development as they can leverage them to license their technology or sell their patents (A. Arora, 1995; Chesbrough, 2003a; Feldman & Florida, 1994; Fosfuri, 2006; Mazzoleni & Nelson, 1998; Merges, 1999; Rivette & Kline, 2000). Research has demonstrated that there exists a favourable correlation between appropriability and innovation performance, but this connection eventually plateaus and reaches a point of diminishing returns (Laursen & Salter, 2006a). Based on the aforementioned synthesis, we can propose the following proposition:

Proposition 3.1*: There is a positive association between a firm's* ***evidence of innovation level*** *and the strength of its innovation-related signal, including a) innovation culture, b) collaboration, c) OI, d) R&D and e) IP-related signals.*

Given the multifaceted nature of the relationship between time to market and innovation, which can depend on various factors such as market competition, product quality, available resources, and company strategy (Datar et al., 1997), a shorter or longer time to market may have an impact innovation performance which may be influenced by factors such as innovation culture, collaboration, OI, or R&D. In addition, it is anticipated that firms that rely on patents may require more time in the R&D phase (Roin, 2013). In light of these observations, we propose the following two propositions:

Proposition 3.2*: There is an association between a firm's* ***time to market*** *and the strength of its innovation-related signal, including a) innovation culture, b) collaboration, c) OI, d) R&D-related signals.*

Proposition 3.2e*: There is a positive association between a firm's* ***time to market*** *and the strength of its IP-related signal.*

The ability to turn knowledge into implemented ideas is critical to innovation, as actors with sufficient knowledge can add, alter, and combine previously assimilated knowledge from a new perspective to make it exploitable (Zahra & George, 2002). This process involves bissociation,

which refers to the creative combination or recombination of two seemingly unrelated ideas or pieces of knowledge to create a third, novel element necessary for innovation (Koestler, 1964). Furthermore, firms with more knowledge are better equipped to absorb external knowledge internally, which is essential for firms to innovate (W. M. Cohen & Levinthal, 1990), particularly in the context of OI opportunities (Bogers et al., 2017).

Proposition 3.3*: There is a positive association between* ***a firm's level of knowledge intensity*** *and the strength of its innovation-related signal, including a) innovation culture, b) collaboration, c) OI, d) R&D and e) IP-related signals.*

From a technological regime perspective (Breschi et al., 2000; Breschi & Malerba, 1997), the life science and health sectors are expected to be strongly associated with innovation activities due to their advanced knowledge base, diverse set of cumulative technologies, and strong appropriation mechanisms through patent activities (Castellacci, 2008; Coriat et al., 2003). Similarly, the digital media and ICT sector is expected to correspond with higher innovation activities due to its high opportunities, advanced knowledge base, and limited appropriation mechanisms (C. Antonelli, 1998). Innovation characteristics of the manufacturing sector depend on the scale of operations and product characteristics, with higher knowledge-intensive SMEs typically specialising their production, yielding high opportunities and strong appropriability through patents (Castellacci, 2008; Pavitt, 1984). Sectors such as natural resources, energy, or environmental services are typically process-oriented with lower opportunities for innovation (Baldwin & Rafiquzzaman, 1995). Therefore, the following propositions are considered:

Proposition 3.4*: A firm operating in either* ***Life science and health or digital media and other ICT*** *have a higher association with the strength of its innovation-related signal, including a) innovation culture, b) collaboration, c) OI, and d) R&D signals than firms from* ***Natural resources, or energy or environmental services or other sectors.***

Proposition 3.4e*: A firm operating in either* ***Life science and health or manufacturing sector*** *have a higher association with the strength of its IP-related signal than firms* ***from digital media and other ICT Natural resources, or energy or environmental services or other sectors.***

Proposition 3.5*: There is a positive interaction between* ***a firm's sector and a high knowledge intensity level*** *on the strength of its innovation-related signal, including a) innovation culture, b) collaboration, c) OI, d) R&D and e) IP-related signals.*

Firms may be classified with Pavitt (1984)' sectorial taxonomy with the adaptation from Baldwin (1995) and Castellacci (2008). Firms may be identified as either mass production (scale-intensive & science-based industries), or value-added knowledge-based (KIBS and specialised-suppliers), or supplier-dominated (resource-based and labour-intensive industries). It was shown in Section 2.1.5 that firms that were classified as either mass production or value-added knowledge-based were expected to be more innovative than firms listed as supplier-dominated. Thus:

Proposition 3.6*: A firm operating in either* ***mass production or advanced knowledge-based industries*** *have a higher association with the strength of its innovation-related signal, including a) innovation culture, b) collaboration, c) OI, d) R&D and e) IP-related signals than firms from* ***supplier-dominated industries****.*

Furthermore, from the 2002-2004 European Community Innovation Survey (CIS) data, Castellacci (2008) presented statistics that showed higher innovativeness for the firms classified as KIBS compared to specialised-suppliers, but lower patenting activities.

Proposition 3.7*: A firm operating within* ***knowledge-intensive business services*** *have a higher association with the strength of its innovation-related signal, including a) innovation culture, b) collaboration, c) OI, and d) R&D-related signals than firms operating as* ***specialised-suppliers.***

Proposition 3.7e*: A firm operating* ***as specialised-suppliers*** *have a higher association with the strength of its IP-related signal than firms operating as* ***knowledge-intensive business services****.*

Other contextual dimensions such as the impact of size, age and country may have an effect on a firm signal, as they do on innovation. According to Van Dijk et al. (1997) larger firms typically have more resources available for R&D and are less vulnerable to market fluctuations. Their comparative advantage is largely attributed to their greater access to resources, while the strengths of smaller firms are often associated with their behavioural traits (Vossen, 1998). Studies have shown that SMEs, especially those in the technology industry, play a crucial role in introducing new ideas and innovations to the market (Acs & Audretsch, 1990; Baumol, 2005; Hicks & Hegde, 2005).

Furthermore, larger firms tend to operate within a mechanistic structure, which is hierarchical and highly organized, whereas smaller firms typically operate within more organic structures that allow for greater flexibility and an innovative culture (Burns & Stalker, 1961; Herzog, 2011). Due to

their limited resources, SMEs tend to collaborate primarily with clients or providers, which may improve as they mature (Kaminski et al., 2008). However, research has demonstrated how much collaboration for SMEs can be differentiating factor (Ebersberger & Herstad, 2012; Hanna & Walsh, 2002; Li et al., 2018; Torkkeli et al., 2007). As for OI, it has been studied extensively in the context of large companies such as P&G, Xerox, IBM, and Intel (Chesbrough, 2003a). Furthermore, the adoption rate of OI was estimated as 78% for 125 large companies that responded to an OI survey in 2012 (Chesbrough & Brunswicker, 2014). Some argue that OI may be even more important for SMEs given their limited resources (Brunswicker & Van de Vrande, 2014), despite the difficulty of leveraging inter-organisational linkages for innovation, particularly in marketing (S. Lee et al., 2010).Given that SMEs typically have less formal and consistent R&D efforts than larger firms, it is expected that R&D signal may be less prevalent in SMEs (Kleinknecht et al., 2002; Michie, 1998). IP protection can also be costly, with the patent registration process requiring significant legal fees, and litigation to defend patents resulting in additional costs (Drechsler & Natter, 2012). Thus, it is expected that larger firms will use more IP-related signalling.

The relationship between a firm's age and innovation is similar as the relationship with size. On one hand, older firms are likely to have accumulated experience and knowledge over time, which can enhance their absorptive capacity (W. M. Cohen & Levinthal, 1990), which may imply more innovation than younger firms (Sørensen & Stuart, 2000). On the other hand, established procedures and routines in older firms may create resistance to integrating new cutting-edge knowledge, thereby hindering innovation (Freel, 2003). This is further demonstrated with mature firms that are encouraged to allow disruptive innovations to be developed outside their main core business operations (Christensen & Raynor, 2013), and invest in startups within the OI paradigm to remain competitive (Chesbrough, 2012). The relationship between age and innovation becomes even more complex when considering the need for signalling innovation-related concepts. However, because patenting is cumulative in nature, we expect a stronger relationship between older firms and IP-signalling.

The Canadian innovation landscape has raised concerns in recent years, as the country has exhibited a lower growth rate compared to other developed nations, despite having a highly educated population with strengths in critical areas of research and science. One notable aspect is the decline in financial support for R&D in all sectors, including firms' investments. In contrast,

Finland is often viewed as an innovation champion, with a successful economy that is more competitive than the US and other European Union (EU) countries on key technological fronts (Oinas, 2005). This success can be attributed to the Finnish government's early focus on national innovation systems and key public investments that have led to strong support for innovation (Fagerberg, 2017; Oinas, 2005; Torregrosa-Hetland et al., 2019). This difference is further supported with latest global innovation index ranking where Canada currently ranks 15th compared to Finland's 9th position. It is also important to note that differences in cultures and languages may affect the signal strength assessed on corporate websites. These dimensions may also be evaluated without necessarily forming formal propositions.

3.3 Data and variable-related methodologies

3.3.1 Building indices

Indices are very useful as they allow the combination of several measures to build a latent construct. There are two types of indices: reflective indices and formative indices (Bagozzi, 2011).

The reflective construct represents an existing latent variable at the intersection of two or more items (Hagedoorn & Cloodt, 2003). In other words, a portion of the items' variance is shared among a group of items, and it is that shared variance *to discover* an underlying dimension. The most important elements to respect are the internal consistency (or reliability) of the items to reflect a meaningful dimension which ensures the convergent validity (Bagozzi et al., 1991). When all the dimensions are mutually exclusive, than there is discriminant validity (Bagozzi et al., 1991)

The Principal Component Analysis (PCA) and Factor analysis (FA) are both popular methods, the first being used to reduce the number of items that explains most of the combined variance from a dataset, and the second to generate latent factors maximising the common variance per factor (Hair et al., 1998). Both these methods share the main underlying assumption that the relationships between the items are linear by nature. Moreover, the items in a PCA need to be on the same scale and multivariate normal distributed, an assumption required for interpreting inferences (T. W. Anderson, 1963; Jackson, 2005). Exploratory Factor analysis (EFA) does not have a strong requirement for normality within the input data with the exception of the ML extraction method (Hair et al., 1998).

The methods regroup the variables into components or factors which can be fine-tuned by removing items that are not loading at all, then the items loading significantly on many dimensions and finally the items that are on their own dimension. Then, the rule of thumb is to consider only the items that load higher than 0.3 for the EFA and 0.5 for PCA after a varimax rotation which ensures the discriminant validity. With the final construct, a Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy is obtained to assess the construct validity at establishing meaningful dimensions for which a minimum value of 0.6 is usually recommended to keep the construct (Kaiser, 1974). Then, the reliability of each dimension is obtained with Cronbach's alpha values which should be higher than 0.7 (Hair et al., 1998).

Confirmatory Factor Analysis (CFA) can also be considered for building reflective variables, but because of the numerous underlying assumptions it requires (normality, adequate sample size, linearity, no multicollinearity, independence of the errors, homoscedasticity, and correct model specification) it is mainly used with well-established construct, not for exploration (Hair et al., 2011; Kock, 2022). Finally, the reflective construct is typically computed with the average of all the items.

A formative index can be created by obtaining the sum of several independent items *to produce* a latent variable. A useful example is provided by Bagozzi (1994, p.332). A researcher who wants to study alcoholism would not obtain a useful metric by obtaining the shared variance of individuals' consumption of different alcohol beverages as people have different preferences. Therefore, it is the total consumption of different alcohol beverages that constitutes a better metric to assess alcoholism.

When it comes to measuring a complex conceptual object such as innovation, another metaphor can be useful. The famous parable of the blind men and the elephant depicts a group of visually impaired individuals who are introduced to an elephant for the first time. In an effort to comprehend what the elephant is like, each blind man feels a distinct part of its body, yet only one facet, such as the tusk or the side or the tusks, or the legs and so on. As a result of their restricted exposure, their depictions of the elephant differ from one another. It is their shared experience that constitutes the elephant. Similarly, independent items related to a concept can be added to produce an aggregated measure that represents the concept.

So how to put together items that are statistically independent of each other to constitute a meaningful dimension? A formative construct relies heavily on conceptual reasoning and field expert judgment when it comes to assembling items together (Bagozzi, 1994; Hair et al., 2011). In other words, first and foremost, it is assumed that a researcher through its knowledge of the literature should know how an "elephant" can be accurately assessed considering the items available for the research. By definition it is considered that there are no errors of measurements, so there is no internal consistency and convergent validity to assess with the formative construct (Edwards & Bagozzi, 2000).

Some scholars also proposed to use a Partial Least Square (PLS) regression model to provide some empirical support for a formative construct (Hair et al., 2011). That can be especially useful if there is some level of uncertainty in cases some items can be related to two or more dimensions. PLS was initially called nonlinear iterative partial least squares and it focuses on increasing the amount of variance that is explained by the independent variables instead of just reproducing the empirical covariance matrix as would be done in a CFA (Wold, 1975). Like a covariant-based Structural Equation Modelling (SEM), a PLS model consists of three parts: a structural component, which shows the connection between latent variables (an inner model); a measurement component (an outer model), which reveals how the latent variables and their indicators are associated; and a weight relation component, which is used to calculate case values for the latent variable (Hair et al., 2011). PLS has shown to be a robust method when performing complex SEM models when the conditions for CFA was not met, such as a small sample size, non-normality or having nonlinear relationships (Hair et al., 2011; Kock, 2022). The software WarpPLS offers Factor-based PLS algorithms, the classic PLS algorithm and PLS Mode A and Mode B (Kock, 2022). Factor-based algorithms are only useful to take into consideration any possible error measurements from a reflective construct. PLS Mode A and Mode B account for the "Good neighbour assumption" in the case of performing a SEM where the inner model interacts with outer models (Kock, 2015). In our case, we are only interested in building a formative measurement model, we only have an outer model. Therefore, the use of either the classic PLS, or PLS Mode A or Mode B would yield the same results estimating our formative outer model (Wold, 1980). Finally, to estimate the p-values we used Stable3 which has shown better accuracy than the bootstrap method in the calculation of p-value with smaller and larger sample, especially considering our sample is not big enough to perform a proper bootstrapping sampling (>5,000 as suggested in Hair et al. (2011)) (Kock, 2018).

PLS regression measurement models can be performed in two steps. First, a formative latent variable is built similarly than with a CFA by allocating each item to its corresponding dimension. Then, the items not loading positively and significantly are removed. In the second step, all the latent dimensions can be generated as variables by the software and then be used to build other latent variables from a second order perspective.

Following the formative construct guidelines from Hair et al. (2011), we need to ensure that all the resulting loadings and weights are positive and significant showing empirical support for our construct. We can see that all Variance Inflation Factor (VIF) are all well below 2.5 which shows the absence of multicollinearity, which is necessary for formative construct. All the indicator effect sizes should be equal or superior to 0.02, which implies that no indicator is too weak to be considered irrelevant (Kock, 2014). Finally, the resulting distributions for each formative variable should be unimodal. This can be done with the Rohatgi-Székely test (Rohatgi & Székely, 1989) and the Klaassen-Mokveld-van Es test (Klaassen et al., 2000). Finally, the average variance extracted should be high enough to demonstrate that the measure is meaningful.

3.3.2 Keywords analysis

All the research papers presented in this thesis use web-based indicators with a keyword frequency method. The lexical field of innovation-related factors was explored through both peer-reviewed publications and seminal books such as the Oslo Manual for innovation (OECD & Eurostat, 2018) and the Frascati Manual for R&D (OECD, 2015), as well as by investigating corporate website content. Keywords considered related to R&D (Gök et al., 2015; Meyer, 2000; OECD, 2015; Rietsch, 2015; Schmoch, 2007), collaboration (Arvanitis, 2012; Belderbos et al., 2004; Carboni, 2013; Ramdani, 2014), external financing (Gök et al., 2015; Li et al., 2018; Rietsch, 2015), OI (Bogers et al., 2017; Chesbrough, 2003a, 2006b; Liyanage et al., 1999), innovation culture (Herzog, 2011; OECD & Eurostat, 2018) and IP protection (Rietsch, 2015; Teece, 1986) were either directly derived or inspired by these publications. Additionally, for the external funding factor, the Government of Canada was offering numerous public programs and funding opportunities for nanotechnology projects at the time. Consequently, these funds and programs' names were converted into keywords related to the external funding factor. Keywords for Paper 1 are available in Table 4-3 and keywords for Paper 2 and 3 are available in Appendix C.

The first paper in this thesis follows a similar approach to Gök et al. (2015) in assessing the correspondence between web-based and questionnaire-based variables on the same factor. To minimise false positives, the precision of the terms used was prioritised, utilising lemmatization versions of the terms within specific contexts. Gök et al. (2015) employed a rule-based variable analysis to determine that the construct with the largest number of different instances, with keywords used in various specific innovation-related contexts, yielded the best compromise between relevance and minimising false positives. These results align with what is expected from a formative index, i.e., the more independent but complementary parts are added, the clearer the picture becomes. Thus, we directly start from this premise and therefore, we did not consider using any rule-based systems. A classic keyword frequency analysis is performed by the firm's website as done previously (Gök et al., 2015; Li et al., 2018).

In Article 2 and 3, we opted for a different approach, as the prior method minimising false positive rates resulted in numerous false negatives. For instance, a sentence like "We develop products that our customers […]" was not identified as an instance of development.

Identifying specific instances of various versions of words in innovation-related contexts proved to be increasingly time-consuming and complex, particularly for five distinct constructs. Additionally, terms associated with practices such as OI had a high number of zero occurrences. As a result, our main focus when evaluating innovation-related signals were to maximise recall, ensuring we did not miss any potentially relevant documents. Stemming may yield higher recall due to its simpler, more aggressive reduction approach, capturing a wider range of word variations. Although lemmatization can produce more accurate results, it may overlook relevant documents because of its stricter morphological analysis. Similarly to Shapira et al (2016), we processed textual content using stemmed versions of keywords to capture most variations of a given concept (e.g., using "innov" instead of "innovation", "innovative", "innovation culture", "innovator", "innovate", and so on). We applied a more conservative stemming approach when necessary to avoid false positives, such as using "create", "creativ", and "creation" instead of "creat" due to potential confusion with "creatine". We performed multiple tests with our original list and removed keywords that were either consistently unused by firms (e.g., "risk-taking", "market introduction", etc.) or more likely to lead to false positives (e.g., experience, test) or that would give excessive exposure to terms frequently repeated in legal pages such as terms of service, privacy policy, or cookie policy (e.g., new product, new service, intellectual property). We acknowledge that this

approach to minimise false negative will likely generate false positives, which is a limitation of these indicators. Finally, we concentrate on sentences as they offer a structure that better identifies firms' signals, as suggested by Lachmann et al. (2001). Consequently, we employed a document keyword frequency analysis that considers instances of stemmed keywords per sentence per website. Here is an example of a strong signal with three instances of keywords related to innovation in one sentence:

*Our core expertise in biotechnology **develop**ment equips us with the intellectual acumen to generate moments of brilliance, reevaluate the status quo, challenge traditional beliefs, and **create innov**ative solutions for enduring problems.*

To account for variations in website structure and size, several standardisation methods were considered, such dividing by the total number of pages, dividing by the total number of words, and finally dividing by the total number of sentences. The number of pages (Li et al., 2018) and word sentences (Gök et al., 2015) were considered, but ultimately not used in our analysis. The number of pages was quickly dismissed due to significant variations in the use of text versus images on webpages. Identifying word sentences proved to be complicated to implement and arguably does not provide substantial improvements compared to simpler methods, such as the total number of words or the number of sentences. The first paper considers all text retrieved from websites; hence, standardisation by the number of words is appropriate as it was done by (Shapira et al., 2016). For Paper 2 and 3, dividing by the number of sentences is considered the appropriate standardisation method, yielding a number of signals per sentence.

Term Frequency-Inverse Document Frequency (TF-IDF) was also considered as it evaluates the significance of a word within a document relative to its prevalence across all documents. Words that appear in a majority of documents are considered less valuable and thus receive a penalty (Ramos, 2003). This method is typically employed for information retrieval, classification and clustering among other natural language processing tasks. However, our goal is to build the aggregated signalling effects from websites for our analyses, rather than comparing terms among each other across different websites. Thus, we did not use TF-IDF.

Finally, keywords were analysed in English, French, and Finnish, with consistency ensured by employing automatic translation software, as done in previous studies involving multiple languages (Kinne & Axenbeck, 2020; Shapira et al., 2016). We utilised Google Translate for

Article 1 and LibreTranslate for Articles 2 and 3. Although our tests did not reveal significant issues, it is more than likely that the translation software may not perfectly capture the intended context of the original text, potentially leading to false positive or false negative results. This is another limitation of our methodology.

3.3.3 Data and treatment

The first article, which addresses the first two propositions, is based on a database derived from a questionnaire-based survey on nanotechnologies and advanced materials in Canada, enabling the use of a sample of high-tech companies from various industrial fields. The second and third papers utilise a database provided by The Evidence Network (TEN), a company that offers impact assessments for innovation intermediaries. Between 2010 and 2016, TEN collected a database of over 4,700 observations that engaged with innovation intermediaries during various mandates. This database presents an opportunity to apply web-based methodology indicators to a larger sample size. This section outlines the data acquisition process and variable creation. Note that for all continuous variables used in parametric methodologies, Kurtosis and Skewness measures are calculated to determine if the resulting variables follow a normal distribution. If not, necessary transformations (e.g., using the natural logarithm or the inverse function) are performed for further statistical analysis.

Nanotechnologies and advanced materials questionnaire data

The initial drafting of the questionnaire was carried out by Marie-Pier Lambert-Desjarlais during her master's project and by Leila Tahmooresnejad for her Doctor of Philosophy (Ph.D) degree at Polytechnique Montréal in 2014. The body of the questionnaire was largely based on the Oslo Manual (OECD & Eurostat, 2005) and was elaborated in collaboration with a Canadian company, AGY Consulting, which specialises in emerging technologies such as nanotechnology, clean technologies and biotechnology. The questionnaire was theoretically validated by Catherine Beaudry and Carl St-Pierre, statistician of the Chair, and operationally validated by Charles-Anica Endo from AGY Consulting. The questionnaire was tested and validated by conducting individual interviews lasting from 1h30 to 2h00 with five Quebec companies developing nanotechnologies and/or advanced materials. Thirteen questionnaires were fully completed by companies that produced nanotechnologies in August 2014.

In the winter of 2015, Mikaël Héroux-Vaillancourt began working on the topic of OI and nanotechnology. The August 2014 data collection was considered a pre-test. Because the response rate was especially low, Catherine Beaudry concluded that the questionnaire needed to be reduced in size since it was particularly time consuming to complete (at least 30 minutes). The initial questionnaire was therefore slightly modified to take into account the new research realities as well as the respondents' reactions to the difficulty of answering certain questions in the questionnaire. In all, 8 questions were eliminated from the questionnaire in order to reduce the completion time. The following themes were explored: innovation, commercialization, collaboration and IP.

The final survey had a total of 13 pages, including an introduction and definitions page. The questionnaire was answered by managers of firms working in fields that develop and use nanotechnologies and advanced materials. Respondents were asked for the total number of product and process innovations and those specifically related to nanotechnologies and advanced materials. In the questionnaire, we defined product innovations as new or significantly improved products (goods or services) and process innovations as new or significantly improved processes. Changes of a purely aesthetic nature and the resale of new products purchased from other companies were excluded from the survey. Also, we asked about the number of nanotechnology-related R&D projects and the time required to develop and commercialise the most recent and most important product innovation. The innovation section of the questionnaire includes the importance of knowledge sources and the importance of OI activities which are measured by a 1-to-7 Likert scale, from not important to essential. Other questions include the importance of commercialisation actions, the proportion of funding allocated to R&D and commercialisation, the importance of market impacts, the number of exports, the importance of barriers to commercialisation, whether the company has collaborated to develop or commercialise the company's most recent and most important nanotechnology product innovation, the importance of the collaborators, the importance of the reasons for the collaboration, IP protection mechanisms and patenting, and finally, control items such as the number of employees, revenue, and industry of the company.

To ensure that the response rate was maximised (the goal was 100 respondents), the Polytechnique team contracted the firm Léger360 to administer the questionnaire and find new companies eligible for our survey. Due to the ubiquitous nature of nanotechnology and the fact that it is being adopted

in a very large number of industrial sectors, identifying companies that use or develop nanotechnologies is not an easy task. Indeed, companies that use or develop nanotechnologies are not labelled and cannot be searched for in an obvious way. First, a list of 583 companies acquired from AGY Consulting was used.

The Leger360 team began data collection in September 2015 using the AGY Consulting list using a convenience sampling method with the Snowball method. Companies that agreed to participate in the survey, but had not yet completed the questionnaire were given reminder calls at one-week intervals on two occasions. In an initial call with R&D managers from the companies on this list, their eligibility to be included in the survey was verified by asking if they were developing nanotechnology or using nanotechnology products or processes. After 332 solicitations, the incidence rate was 28% which means that 93 companies were eligible. Of these, 33 respondents completed the questionnaire in full and 42 accepted the invitation but did not complete the questionnaire at that point in the collection. Since more than half of the sample from the AGY Consulting list was used and only one third of the target (100 respondents) was reached, a second list was composed using the common characteristics of eligible firms.

To do this, firms that were eligible for the study (the 93 firms mentioned above) were listed and classified according to their NAICS codes. A total of 23 NAICS codes representing 67% of all cumulative frequencies were identified. Leger360 acquired from InfoGroup a list of 3345 companies based on the NAICS code frequencies. Of this list, 306 companies were duplicates and 703 records were not used. By the end of December 2015, a total of 2971 Canadian high-tech firms were solicited. Of this total, 973 companies did not respond, 1459 were ineligible, 380 refused to participate, and 168 eligible companies agreed to participate, of which 80 respondents completed the entire questionnaire. Because it is not known whether respondents refused to participate before or after answering the eligibility questions, and also because of the unknown nature of the non-respondents, an approximation of the response rate was made. Assuming a uniform distribution, a response rate of 11% was obtained by assigning the same characteristics to respondents who answered eligibility questions as to nonrespondents and respondents who refused to participate (see Table 4-1). Of course, it is assumed that respondents were honest and completed the survey in good faith. After eliminating duplicates and checking for history bias, 9 respondents from the August 2014 survey were added to the sample bringing the total to 89 companies. Several

additional tests were conducted to ensure that the potential for inherent bias was minimal. The questions relevant to this research is available in Appendix A.

Nanotechnologies and advanced materials web data collection

We selected the 89 companies obtained and employed a web scraper, Nutch 2.2.1[3], an open-source web scraping software, to extract and record all text content from their websites as of February 2016. The scraper was initialised with a list of URLs corresponding to these companies, which served as the first level of indexing traversal. The scraper then navigated the entire website structure to extract all URLs, Hypertext Markup Language (HTML) code, and links to other pages.

The scraper prioritised text collection by ranking pages based on a score calculated using the number of links pointing to them. We restricted the scraper to the vertical links of the domain names present in the initial list, allowing us to modify the Nutch configuration file to exclude other links. We also utilised the Tika plug-in to index text from Portable Document Format (PDF) and Microsoft Word documents and identified and translated pages in languages other than English using the Compact Language Detector software and Google's programming interface, Goslate.

Nutch's technical limitations prevented us from extracting content from 10 websites that seemed to be associated with the use of a JavaScript framework not recognised by the software. From the initial 89 firms in our study, we successfully extracted text content from 79 firms (88%).

Nanotechnologies and advanced materials variables

For the questionnaire data, first a PCA is performed to assess whether some items can be reduced (especially those assigned to R&D) into latent variables. However, both the KMO and Cronbach's Alpha were inconclusive. Thus, the items were treated individually to test proposition 1.1.

We retrieved all the R&D-related activities with the following variables: the number of R&D projects in nanotechnology/advanced material (*nb_R&D_proj*), a dichotomised version of the importance of internal R&D as a source of knowledge (*dInt_R&D_source*), the level of importance of Commercial laboratories/R&D firms/Technical Consultants as a source of knowledge (*Ext_R&D_source*), the level of importance of contracting of external R&D service providers (*Contract_In_R&D*), the level of importance of providing R&D services to third parties

[3] Apache Nutch. (n.d.). Apache Nutch 2.2.1 API [Documentation]. Retrieved June 4, 2020, from https://nutch.apache.org/apidocs /apidocs-2.2.1/index.html.

(*Contract_Out_R&D*), the time allocated to R&D (*time_R&D*), dichotomised importance of Private research laboratories/Research and Development firms as collaborators for the development and the commercialisation (*dCollab_R&D*), dichotomised importance of accessing research and development from collaborators for the development and the commercialisation (*dCollab_R&D_Reason*) and the proportion of Canadian employees assigned primarily in R&D (*propEmp_R&D*).

The items related to IP are the number of IP mechanisms used (*nbIP*) and the number of patents (*nbPatent*).

Only one item is directly and solely related to collaboration which is whether a firm used collaboration for the latest innovation (*dCollab*).

For external three relevant questions were retrieved: the proportion of external financing for R&D (*PropExtFin_R&D*), the proportion of external financing for commercialisation (*PropExtFin_Comm*), and the total proportion of external financing versus all financing (*propTotExtFin*).

To test proposition 1.2, we considered enlarging the scope of the measurements by considering all the questions related to two factors: R&D and IP. This approach provides a relative importance attributed to these factors. Consequently, we combined single items into a single measure (i.e., a formative index with the sum of the weighted variables). The items were transformed into z-scores. PLS regressions were performed, and the sum of all items with their corresponding weights was calculated. This resulted in *R&D_INDEX* and *IP_INDEX*.

Each web-based variable was derived from a data processing pipeline created using RapidMiner – Community Edition software. The software was chosen because it provides an intuitive user-friendly interface along with powerful text mining tools for beginners. The software connected to a MySQL database, and the text content underwent cleaning steps, translation, and direct computation of the total occurrences of each keyword for each factor per website. The results were then standardised by dividing the total number of words then multiplied by 1,000 as done by Shapira et al. (2016) . This process produced four continuous variables: *Web_R&D*, *WEB_IP*, *WEB_COLLAB* and *WEB_ExtFin*.

TEN questionnaire-based data

TEN provided our research with a dataset collected between 2010 and 2016 from 29 investigations (also named programs) containing 4,740 entries from a range of Canadian and Finnish organisations, including businesses, nonprofits, and public institutions.

The data in this study was collected from 13 tailored questionnaires, designed and contextualised by TEN representatives to suit target respondents across 29 investigations. The online questionnaires contained an average of 30 questions, divided into four sections: preliminary questions, questions on the impact of the intermediary on the company's resources and capabilities, questions on the impact of the intermediary on market performance and concluding questions. The questionnaires were designed to be completed within an average of 15 minutes, and the investigations were conducted within a maximum of one month. Due to the close relationship between companies and intermediaries, along with follow-up calls and email reminders, our sample achieved an average response rate of 39% across all investigations (Appendix D).

From all the investigations, we combined the results from the questions that were relevant to our study that include: the control variables such as the foundation year, employment, revenue, sector, the variables of interest such as the number of innovations introduced in the past 3 years, the percentage of the company’s revenues that is attributed to innovations, the degree of innovation, the time to market, and the investment received. Examples of the questions can be seen in Appendix B.

While the questionnaires cover the same topics, there are notable differences in their design, necessitating a harmonisation process when combining data from different investigations. Some variables on the same topic were not measured with the same type of questions, requiring conversion to common scales. For example, the number of employees was sometimes measured by directly asking respondents for the number of full-time and part-time employees, while in other cases, multiple-choice options with intervals were offered. When we had direct entries of employee numbers, we summed all full-time and part-time employees to obtain a single number, representing total employees. In the case of interval scale measurement, we added the midpoint of the selected scale to the new combined variable of total employees. All other items, except for the sector and the degree of innovation, were also combined either by taking the midpoint in the case of interval

measurement, or by using the direct value provided by the user, either as a directly entered number or the number corresponding to a multiple-choice value.

The harmonisation of various sectors necessitated special attention due to the differing, non-mutually exclusive categories used in the questionnaires. The study featured questionnaires with different levels of specificity, from targeting subsectors like Green Energy Technologies, Environmental - Clean Technologies, and Environmental - Water Technologies, to broader categories like Energy and Environment, and even more general classifications like Energy or Environment. To ensure consistency, we grouped these categories into their most general form: Energy or Environment. Ultimately, this resulted in six primary sectors: Materials & Manufacturing, Life Science & Health, Energy & Environment, Digital Media & ICT, Natural Resources, and Other. The "Other" group encompasses a variety of sectors, including financial services, food and beverage, construction, retail, entertainment, and more.

The degree of innovation was measured using either a 1-to-7 or 1-to-5 ordinal scale based on the respondent's perceived level of innovation[4]. The resulting item, *InnovationDegree*, resulted into a 1-to-3 ordinal scale: 1 for firms that had not yet developed an innovative product, process, or service (11.57%); 2 for national and regional leaders combined (57.31%); and 3 for all international leaders combined (31.12%).

Since the surveys were not specifically designed for this study, we needed to gather some additional web-based data. First, we had to obtain the websites of the organisations represented by the respondents. Second, to better segment our sample for our research, we categorised the organisations into six mutually exclusive types: tech company, low-tech company, advanced knowledge service, other services, nonprofit/ Non-Governmental Organisation, public institutions, and undefined[5]. Our priority was to first identify the organisation's website, which would allow us to determine the organisation type. Consequently, we processed each entry individually to collect

[4] It is either measure with a 1-to-7 ordinal scale where 1 is “International leader, we rank first in the world”, 2 is “International advanced, there are few companies in the world with our capabilities”, 3 is “National leader, we rank first in Canada”, 4 is “National advanced, there are few companies in Canada with our capabilities”, 5 is “Regional leader, we rank first in our region”, 6 is “Regional advanced, there are few companies in our region with our capabilities” and 7 is “We have not yet developed an innovative new or improved product, process or service” or a 1-to-5 ordinal scale where 1 is “International leader, we rank first in the world”, 2 is “International advanced, there are few companies in the world with our capabilities”, 3 is “National leader, we rank first in Canada”, 4 is “National advanced, there are few companies in Canada with our capabilities” and 5 is “We have not yet developed an innovative new or improved product, process or service”.

[5] undefined: The information from the website may not be enough to properly identify an organization type.

this additional data for our sample.

To retrieve the organisation's website, we primarily used the organisation's name when available and found a valid URL through Google's search engine. If the name was not available for a given investigation, we used the domain name from the respondent's email[6]. If the domain name corresponded to a popular email service, such as Gmail or Hotmail and so on, we searched for the respondent's LinkedIn profile or used Google's search engine to match respondent's name affiliated with an organisation that matched the timing of the questionnaire completion. If there was any reasonable doubt of a false positive, we would not include a website.

Out of 4740 observations, we retrieved 2949 websites, leaving 1791 observations not included in this study. Assessing the organisation type was mostly trivial, as we found this information directly on most homepages or "About" and "Career" sections. We made distinctions between tech companies, low-tech companies, and advanced knowledge service firms based on evidence of STEM discipline involvement.

We removed responses from nonprofits (64), public institutions (75), and undefined (30) respondents, resulting in a sample of 2780 respondents (58.6%) corresponding to 2483 unique companies with a working website domain. Two hundred and sixteen companies (8.7%) were represented by more than one respondent, with 171 companies by two respondents, 34 companies by three respondents, five companies by four respondents, one company by five respondents, one company by six respondents, one company by seven respondents, one company by eight respondents, one company by 10 respondents, one company by 11 respondents and one company by 12 respondents.

To increase the size of the sample from the questionnaire data, imputation was performed on the following variables: Revenue (2,296/2,190), the Number of innovations (990/919), the Percentage of the company's revenues that can be attributed to innovations (453/398), the Time to Market (2,243/1,830), and the Investment received (2,010/1,777). We generated 2,784 imputation classes based on the unique combination of the following variables that correspond to the nearest neighbour that shares some common intrinsic characteristic: Sector (6 categories), Country (2 categories), Size (4 categories), whether it is a tech firm (2 categories) and the Program ID (29

[6] It corresponds to the characters at the right of the "@"

categories). For each imputed variable, we calculated the means within each class and attributed that value to all the missing values from the same class and we ensured that all the imputed values represent less than 30% of the original sample which is the threshold recommended by the literature (Särdnal et al., 1992).

Our final database from the questionnaire-based surveys with imputed values include the following items: Respondent identification (ID) (2,780), Program ID (2,780), Organisation type (2,780), Website ID (2,780), Website (2,780), Investigation Year (2,780), Country (2,780), Sectors (2,457), Foundation Year (2,320), Age at the time of the questionnaire (2,320), Revenue (2,296), Number of Employees (2,282), Number of Innovations introduced in the past 3 years (990), Percentage of the company's revenues that can be attributed to innovations (453), Degree of Innovation (752), Time to Market (2,243), and the Investment received (2,010).

TEN questionnaire-based and complementary web-based variables

To test proposition 2.1 and propositions 3.1, we created two variables based on the concept of evidence of innovation results based on questionnaire data. These are not a direct measure for a specific output or outcome, but rather it provides a general assessment of a level of innovativeness of the firm (OECD & Eurostat, 2018). We generate two sets of variables for each group of variables: a categorical variable and the dummy variables corresponding to each category.

First, a continuous variable called Evidence of innovation (*InnovEvidence*) is built by converting all our innovation-related items into ratios and combining them. These items are the innovation degree (*InnovationDegree*), the number of innovations (*InnovationNumber*) and the percentage of a company's revenues that can be attributed to innovations (*InnovationProportion*). First, we subtracted 1 from *InnovationDegree,* so that the ordinal scale was going from 0 to 2. Then, we divided all the three variables by their maximum value, so they all share the same 0 to 1 scale. Since the variables are all mutually exclusive, the sum of the variables was calculated to obtain the firm's corresponding value for *InnovEvidence*. The resulting variable follows a normal distribution which can be understood as a measure of the intensity of evidence of innovation results, where 0 is the absence of innovation evidence and 1 is the strongest innovation evidence we obtained.

The second variable of evidence of innovation results is an ordinal index called *oInnovEvidence*[7]. The resulting index ranges from 1 to 3, with 1 indicating no evidence of innovation, 2 indicating some evidence of innovation, and 3 indicating strong evidence of innovation. Firms in the first group have not developed any new or improved products, processes, or services or did not declare any innovations that are new-to-the-world or did not have any revenue that can be attributed to product, process, or service innovation. Firms in the second category exhibit moderate regional or national innovation novelty or have either introduced one or two innovations considered to be new-to-the-world or generate less than 50% of their revenue from product, process, and service innovation. The third group of firms represents firms with a strong international leadership in terms of innovation novelty or that have introduced three or more innovations considered to be new-to-the-world or generate more than 50% of their revenue from product, process, and service innovation. Dummy variables were generated for each category of firm, resulting in *dNoInnovEvidence*, *dSomeInnovEvidence* and *dStrongInnovEvidence*.

The time between conceptualisation and market launch is referred to as the time to market. Firstly, time to market is used as a productivity measure of the innovation process that also explains the evidence of more innovation results (Ivanov & Avasilcăi, 2014). Similar to the innovation evidence, two variables are generated. A continuous variable called *TimetoMarket* is obtained directly in years, with a range going from 0.5 to 9 years from multiple choices with intervals. It was converted to months to allow for better granularity.

Second, firms with a different time to market profiles may have an impact on the signalling. Given the variation in time to market across different industries, such as pharmaceutical products taking significantly longer to develop than software products, the 20th and 80th percentile values were obtained for each sector. To classify firms based on their time to market, we assigned a value of 1 to firms that fell closest to the 80th percentile value of their corresponding sector, 3 to firms that fell closest to the 20th percentile value, and 2 to the remaining firms. This resulted in an ordinal variable *oTimeMarket*, where a value of 1 denotes firm with a longer time to market, a value of 2 represents firms with an average time to market, and a value of 3 indicates firms with a shorter time to market. However, because the measurement was obtained with intervals, the classification

[7] In paper 3, we refer to this variable as a nominal index. To enhance the clarity of the methodology, we did not name the variable and just presented the dummy variables corresponding to each category.

of firms was attributed to the closest interval of the corresponding percentiles and we ended up with three relatively close groups (30%, 36% and 34%). Again, dummy variables for each category of firm were also generated, which resulted in *dLongTMkt*, *dAverageTMkt* and *dShortTMkt*.

To test proposition 2.2, we obtained the Investment received (*Investment*) directly with an interval measure. Approximately half of the investment measures considered in our sample are exclusively in the form of equity financing, while the remainder is the total amount of external financing, including both equity, grants, and loans. This variable is only needed to test propositions 2.1 and 2.2.

Technological companies typically require more financial resources to engage in their innovation activities (Kor & Mahoney, 2005) and often need external financing (Kalil, 2005; McNeil et al., 2007). They also tend to attract more private equity investments (Brown et al., 2009). As such, it is expected that these factors contribute to explaining the investment received. To test proposition 2.2, tech firms are identified with a dummy variable (*TechFirm*).

Furthermore, as propositions 3.3 and 3.5 consider the level of knowledge intensity, consider the level of knowledge intensity, tech firms and advanced knowledge service firms were combined into a categorical variable called Knowledge Intensity (*nKnowledgeIntensity*: 1 to 2), where 1 is considered low and 2 is considered high. Dummy variables, *dLowKI and dHighKI* are also computed.

Six sectors were identified, including Digital Media & ICT (*dDigMediaICT*), Life Science Health (*dLifeSciHealth*), Materials & Manufacturing (*dMaterialMnfg*), Natural Resources (*dNaturalRes*), Energy and Environment (*dEnergyEnv*), and Other (*dOther*). A combined group consisting of Natural Resources and Energy & Environment (*dEnerEnvironNat*) was also assembled for some analyses due to their relative similarity in innovation-related characteristics and to increase the number of firms. Nominal variables for the sectors (*nSector*: 1 to 5) were generated, with 1 corresponding to Materials & Manufacturing, 2 to Life Sciences & Health, 3 to Digital Media & ICT, 4 to Natural Resources, Energy & Environment, and 5 to Other.

In order to test propositions 3.6 to 3.7, an adaptation of the sectorial taxonomy previously discussed (Baldwin & Rafiquzzaman, 1995; Castellacci, 2008; Pavitt, 1984) was derived based on the

available information about the sector, organisation type, and number of employees[8]. The process began by classifying all large and medium firms as either scale-based (*dScaleManuf*), science-based (*dScience*), KIBS (*dKIBS*), specialised-suppliers (*dSpecializedSupplier*), or supplier-dominated (*dSupplierDominated*). For the remaining small firms, those from the Digital Media and ICT sector [9] and those operating as advanced knowledge services were combined as KIBS. Small tech firms not included in the Digital Media and ICT sector were labelled as specialised-suppliers. All remaining firms were categorised as supplier-dominated.

Since there were a limited number of scale-based and science-based firms, they were combined into another category called *dMassProd*. In the end, a nominal variable *nSectorialCategory* was generated with 1 for supplier-dominated, 2 for specialised-suppliers, 3 for KIBS, and 4 for mass production.

Statistics Canada classifies firms with fewer than 100 employees as small firms, those having between 100 and 499 employees as medium firms, and those with 500 or more employees as large (Government of Canada, 2020). Because 92% of the sample consists of small firms, of which approximately 66% have fewer than 10 employees, it was decided to distinguish very small firms from small firms. In the end, a nominal variable *nSize* was generated with 1 corresponding to firms with 1 to 10 employees (*dVerySmall*), 2 for firms with 11 to 99 employees (*dSmall*), 3 for firms with 100 to 499 employees (*dMedium*), and 4 for firms with more than 500 employees (*dLarge*). Additionally, the annual revenue (*Revenue*), obtained directly with interval measures, provides another way to assess a firm's size.

We generated firm's age profiles based on their distribution. Three intervals representing approximately 33% of the age distribution of the company at the time of the questionnaire were generated. We generated a categorical variable (*nAge*: 1 to 3), with firms that were 3 years old or younger at the questionnaire year in group 1, between 4 and 8 years old at the questionnaire year in group 2, and at least 9 years old at the questionnaire year in group 3. Dummy corresponding to these categories are also generated with *dEarlyStage*, *dYoung* and *dMature.*

[8] The adaptation of the taxonomy presented is meant to be a useful framework for exploration and to provide a general indication of the relationships among the different signaling dimensions and sectorial attributes as described by this adaptation of the taxonomy. We acknowledge this operationalization of the taxonomy may not precisely reflect the intended dimensions. Nevertheless, we think it is enough representative to test the propositions.

[9] The vast majority of these firms are doing software, internet business and consulting. It is possible that the Digital media and ICT group includes some firms in telecommunications.

Two categories referring to Finnish (1) and Canadian (2) firms (*nCountry*: 1 to 2) were added. Both *dFinland* and *dCanada* were also generated.

Finally, since the investigations are distributed overtime, we generated variables associated to years (*d2010*, *d2011*, *d2012*, *d2013*, *d2014*, *d2015* and *d2016*) and time periods (*d2010_2011*, *d2012_2013* and *d2015_d2016*). The categorical variable generated for the time periods include 2014 as a category on its own because about half of the respondents answered that year (51.8%). Therefore, *nPeriod* with 1 corresponding to 2010 and 2011, 2 corresponding to 2012-2013, 3 corresponding to 2014, and 4 corresponding to 2015-2016.

TEN web data collection

The web-based data acquisition process for this part of the research was made possible through the participation and close collaboration with Davide Pulizzotto. A process was built from scratch to maximise our output. The following is an extensive description of the data acquisition process.

Due to the significant time lag between when the study was conducted and when the questionnaire data was originally collected, it was determined that obtaining current website information would not be methodologically sound. The context in which these companies operated may have changed dramatically, particularly considering the impact of the COVID-19 pandemic on the economy. Furthermore, some firms may have been acquired, closed, or changed their corporate names, which would result in a significant drop in the number of corresponding companies. To support this reasoning, a scraping test was performed in early 2022. From the main sample, a sub-sample of 1900 websites was selected, and only 788 companies (41%) could be scraped. The information retrieved online was heavily biased towards COVID-19 messaging. To maximise the relevance and amount of data from the websites, it was necessary to close the gap as much as possible between the time the questionnaires were administered and the web content data needed.

To collect the data, we utilised the Wayback Machine, an online repository containing snapshots of publicly accessible web pages from the past 26 years. However, there are many limitations with the service, the more important is the lack of rigorous consistency in the crawling process which causes significant selection biases. In fact, the number of snapshots for the same website can exhibit high variance on a year-to-year basis (Gök et al., 2015). Furthermore, certain pages within a website, such as homepages, appear to be crawled more frequently than others (Arora et al., 2016). This leads to another limitation: snapshots of different pages within the same website may

not be taken simultaneously, potentially causing gaps in the chronological record of events and communications. Additionally, the Wayback Machine crawler seems to treat websites differently, with larger websites being crawled more often and some not being crawled at all (Gök et al., 2015). Finally, there is often a significant delay, ranging from six months to two years, between when snapshots are captured and when they become publicly accessible, making the Wayback Machine unsuitable for studying recent history (Arora et al., 2016). Despite these challenges, researchers have successfully employed and validated the Wayback Machine as a reliable data source for social sciences (Murphy et al., 2007). Arora et al. (2016) provides methodological guidelines for using the Wayback Machine in social sciences, outlining a six-step process consisting of sampling, organising and defining the boundaries of the web crawl, crawling, website variable operationalisation, integration with other data sources, and analysis. We drew upon this structure to inform our web-based data collection methodology. The first three steps are adapted from Arora et al. (2016)'s structure with 1) sampling, 2) organising and defining the boundaries of the web crawl and 3) crawling. Then we added further data processing steps with 4) translation and clean-up & pretreatment. The website variable operationalisation will be discussed in the subsequent section, and additional steps concerning the integration with other data sources and analyses are extensively covered in the presented research papers.

Step 1. Sampling

The initial step pertains to the validity of employing the Wayback Machine in the context of our specific sample. To ensure the validity of this step, the following questions are presented:

1. Do the organisations in the sample have websites?
2. Do their websites provide information to answer the analyst's research questions?
3. Does the Wayback Machine capture the relevant historical websites? Arora et al. (2016, p.1907)

The initial two questions were swiftly addressed, as our sample consists solely of companies with functional websites, and historical data is crucial to answering our research question. To fulfil the third prerequisite, we chose a random subsample of 50 homepages and examined them on the corresponding dates using the graphical user interface provided by the Wayback Machine website.

We discovered that in nearly every instance, there was at least one usable snapshot of the homepage available within a three-year timeframe.

Step 2. Organising and defining the boundaries of the web crawl

The second step involves determining the scope of the query, which can be summarised by the following questions:

> 1. What kind of information presented on webpages needs to be crawled?
>
> 2. Which webpages on an organisation's website should be included in the collection? Arora et al. (2016, p.1908)

For this study, we are solely focusing on text content that is directly visible to visitors, contained within the HTML file under the title tags h{1-9} and the paragraph tag <p>. We are not considering text within images, videos, or other formats such as Adobe Flash or JavaScript, as processing these formats at scale is highly challenging. Regarding the second question, we determined that the homepage and all pages directly accessible from it represent the most significant content presented to the public by the firm. In fact, Chartbeat, a data analytics company, conducted a study in 2014 analysing two billion website visits, finding that 55% of visitors spent less than 15 seconds actively engaging with a page[10]. This shows that most visitors do not go further than the homepage, and firms are most likely aware of this as well. This suggests that most visitors do not venture beyond the homepage, a fact likely recognised by firms as well. Some researchers have deemed the homepage alone to be sufficient data for their analyses, as demonstrated in a paper predicting firms' survival (Blazquez et al., 2018). However, based on our experience with our sample of companies, some websites contain minimal textual information on their homepage, particularly in the case of very small enterprises. Including a second layer would significantly increase the amount of text content, we can collect while remaining just one click away from the homepage, ensuring that the content remains strategically important from the firm's perspective. Lastly, to minimise the scraping process's duration and limit potential outliers, we capped each query at a maximum of 50 pages, which should encompass the firms' most critical communications.

[10] Haile, T. (2014, March 9). What you think you know about the web is wrong. Time. Retrieved March 15, 2023, from https://time.com/12933/what-you-think-you-know-about-the-web-is-wrong/

According to a 2015 survey conducted by RBC Capital Markets involving 379 SMEs in the US, it was found that most companies do not update their websites frequently; in fact, it is common for these websites to be updated only once a year or even less often[11]. To ensure the web data aligns as closely as possible with the temporality of the questionnaire data while maximising the number of distinct snapshots obtained for a given website, we chose to cover a three-year period with July 1 of a questionnaire year as the midpoint. Another consideration relates to multiple snapshots of the same pages (particularly homepages) of a website within a specific time range, which could artificially inflate the occurrences of keywords and introduce a methodological bias. To mitigate this issue, the nearest snapshot of the homepage from July 1 of a questionnaire year is selected. Then, a breadth-first search algorithm is employed to crawl the website based on the criteria established in step 1. All unique URLs are gathered, and only those within the three-year range are retained. These URLs correspond to pages that make up the second layer depth for our analysis.

Although we are confident that this method minimises duplicates, it may not capture all potential URLs from the existing second layer for the given three-year period, leaving us vulnerable to homepages in different languages. To maximise second-layer scraping, we explored an alternative method: applying a breadth-first search algorithm to all homepage snapshots within the time range and processing the resulting URLs. However, this would significantly increase the time required for data acquisition and complicate our methodology. Indeed, addressing duplicate pages, minor updates, or significant updates would necessitate sophisticated rules, which is not a trivial task. Due to these reasons and the fact that we have no guarantee that this process would substantially improve data acquisition outcomes, we decided against pursuing this step.

Step 3. Crawling

This step operationalises the crawling strategy as defined in Step 2. Since no existing crawler software provided the flexibility needed to execute our crawling strategy, we opted to create our own Wayback Machine Crawler in Python tailored to our use case. The Wayback Machine offers a free public Application Programming Interface (API), but it does impose limits on the number of requests per minute. Our Wayback Machine Crawler performs gentle multi-retry crawling in several steps to obtain the maximum number of pages in the shortest amount of time possible

[11] Statista. (n.d.). Frequency of updating their company website according to small business owners in the United States as of March 2017. Retrieved March 15, 2023, from https://www.statista.com/statistics/472340/smb-frequency-update-website-usa/

(approximately 10,000 pages per hour). The crawler has three main processes: 1) managing interactions with the Wayback Machine server, 2) storing results in a MongoDB database, and 3) writing log files from console activity. The open-source code is available on GitHub[12]. From our dataset, we combined all the websites and their corresponding questionnaire years into a list of 2,684 unique pairs structured as follows: website.com, 2014.

The process begins by dividing the list into *n* sub-lists, where *n* is the number of parallel processes set by the user. Each process sequentially takes the first item from its list to make the query, using the server's API to send a request. While the exchanges between the crawler and the API are more detailed and sophisticated, it is crucial to understand that for each request, we receive either a positive server response (indicating successful retrieval of textual data) or a server error response. In both cases, we populate a different collection in our MongoDB database: 1) Data and 2) Server error responses. After the crawler has processed the initial list, we create a new list from a selection of all Wayback Machine URLs with server error responses related to the server's temporary inability to process the request, such as "#429 – Too many requests". We performed several iterations until no new text content was obtained. Upon completing the scraping process, we amassed a corpus of 45,241 pages.

However, we observed that the Wayback Machine also crawls broken links on some websites, resulting in numerous pages with server errors being captured by its crawler. Fortunately, we identified many page titles containing terms related to broken links, such as "password", "wayback machine", "404", "error", "not found", etc. We compiled a list of these keywords to detect and remove residual server errors. After cleaning the corpus of these pages, we were left with 39,027 pages (86%).

Step 4. Translation

As our keywords are in English and the majority of the firms in our sample communicates in English as their primary language, we decided to focus this study on English content. We used Google's Compact Language Detector 2 (cld2) R package[13] to process all the text content in our corpus, identifying 28,381 pages as English (73%), 10,351 pages (26.5%) in languages other than

[12] 4point0-ChairInnovation-Polymtl. (n.d.). Wayback Machine Crawler [Repository]. GitHub. Retrieved March 15, 2023, from https://github.com/4point0-ChairInnovation-Polymtl/Wayback_Machine_Crawler

[13] rOpenSci. (n.d.). cld2: Google's Compact Language Detector 2 [Documentation]. Retrieved March 15, 2023, from https://docs.ropensci.org/cld2/index.html

English, and the remaining pages undetected by cld2 (<1%). To optimise the number of pages retained in our analysis, we employed automatic translation to process non-English pages. Automatic translation comes with methodological challenges and systemic biases, one of which is polysemy. Words can have multiple meanings depending on their context, and translation software may not accurately translate content as a domain expert would. Since the lexical set related to innovation is domain specific, ensuring the accurate translation of intended meanings in English is difficult and could lead to false positive or negative interpretations.

This issue is particularly relevant when using bottom-up Natural Language Processing (NLP), such as topic modelling, where context is crucial. However, using keyword frequency analysis mitigates the risk of false positives, as the focus is on finding relative occurrences of terms rather than deriving meaning from interrelationships between words and concepts. As mentioned earlier, the keywords were also tested in the most important languages other than English, and the results were found to be acceptable.

In conclusion, while there is a possibility that duplicates from homepages' translated versions may artificially amplify the signal for certain websites, it is important to note that the Wayback Machine crawler tends to give more exposure to snapshots of the main domain language, thereby limiting the number of snapshots taken for other languages' versions. Based on this observation and the other factors considered in the study, the use of automatic translation has been deemed appropriate for the purpose of this research.

To translate the 10,351 non-English pages into English, we used LibreTranslate from Argos translate, which is an automatic translation software based on OpenNMT an open source neural engine system for translation and sequence learning (Klein et al., 2017, 2018, 2020). OpenNMT has its origins in a collaboration between the Harvard NLP group and a firm called SYSTRAN in 2016. The GitHub is frequently updated and is currently maintained by SYSTRAN and another company called Ubiqus. The original article appeared to be cited 1,822 times on Google Scholar, demonstrating its relevance to its research area.

LibreTranslate offers an online service with a user interface, public API, and a premium API version, which we utilised to process up to 80 queries with a maximum of 1,500 characters per

minute for a monthly fee of $19. LibreTranslate[14], Argos translate[15] and OpenNMT[16] are all frequently maintained open source initiatives.

We built a function to interact with the LibreTranslate API using R and RStudio to translate the non-English pages in the dataset. We used the R package Parallel to optimise the number of processes performed simultaneously. Pages with more than 1,500 characters were segmented into groups with temporary unique IDs to respect the API's threshold. The query returned the temporary unique ID (allowing for recombination of the translated text into the original pages), the translated content, and the API response code. As we did with the Wayback Machine process, we performed several rounds of queries on the API responses that experienced one-time server errors.

From the initial 10,351 non-English pages, 10,221 were written in languages supported by LibreTranslate. We successfully translated 10,210 pages (26.4%), bringing the English corpus to a total of 38,591 pages. Out of the initial sample of 2,684 unique website-questionnaire year pairs, we were able to obtain the English textual content for 2,318 unique pairs (86%), which corresponds to a total of 2,379 observations from the questionnaire.

Step 5. Clean up and pretreatment

We performed a clean-up process on the resulting corpus with the goal of obtaining intelligible text segments separated by line breaks. Line breaks from the online text content were obtained in the HTML and coded as "\n", which served as the basis for structuring the segments. To address potential encoding errors, applied the uniformisation of the characters of the text content in Unicode Transformation Format-8 (UTF-8) before the cleaning process.

Using regular expression (regex) with functions of the stringr R package[17], we started by removing patterns of HTML and JavaScript code, standardising spaces, punctuation, quotation marks, etc. We removed all the dots that were not used for punctuation such as dots following a capital letter directly (e.g., M Smith) with the exception of acronyms. We also noticed that some translation process generated some systematic error in the translated text. Indeed, some sentences started with

[14] LibreTranslate. (n.d.). LibreTranslate [Repository]. GitHub. Retrieved March 15, 2023, from https://github.com/LibreTranslate/LibreTranslate
[15] Argos Open Tech. (n.d.). Argos Translate [Repository]. GitHub. Retrieved March 15, 2023, from https://github.com/argosopentech/argos-translate/
[16] OpenNMT. (n.d.). OpenNMT [Repository]. GitHub. Retrieved March 15, 2023, from https://github.com/OpenNMT
[17] https://stringr.tidyverse.org, accessed on November 23rd 2022

a lower-case letter and some line breaks (\n) were introduced in the middle of a sentence. Thus, we needed to perform specific clean-up operations on the translated text such as replacing the lower case with upper case at the beginning of a sentence and eliminate line breaks that were in the middle of intelligible segments. Finally, we removed recurring patterns of text that added noise, such as sentences like *The JavaScript in your browser is currently disabled.*" or "*All rights reserved*" or phone numbers and so on.

After the clean-up process, we ran again the language detection function from cld2 again to remove all segments in languages other than English (114,020), resulting in 822,299 segments. Then, we focused only on sentences, as they are considered the best signalling unit according to the signalling theory literature (Lachmann et al., 2001), resulting in a final corpus of 511,815 sentences. The descriptive statistics for the final corpus are as follows:

- Min: 1
- Max: 4,785
- Mean: 220.98
- Standard deviation: 331.64
- Median: 110

Twenty percent of all cumulated frequencies contain observations with 22 sentences or fewer, while 80% of all cumulated frequencies contain observations with 345 sentences or fewer. The last 20% of the cumulated frequencies have an especially long tail, ranging between 345 and 4,785 sentences. This high variance can be explained by differences in structure and size among the websites, as well as the varying exposure to the Wayback Machine crawler.

TEN web-based data variables

From NLP, document frequency is the number of documents containing a particular keyword. In our case, a sentence is considered as a document. To generate our document frequency table, we have created a function that performs two operations on an entered keyword:

1. Adds a new column in the data frame for the analysed keyword with the name of the keyword.

2. Uses a regex to detect the pattern of the keyword[18], ignoring the case, for each sentence in the corpus. If the keyword is detected in a sentence, a value of 1 is assigned in the column of that keyword; otherwise, it is assigned 0.

By applying this process to each keyword on the list, we obtained a document frequency of their keywords. It is important to note that a sentence may contain more than one instance of an innovation-related term, which would result in a value for that sentence that is greater than 1. We finally aggregated all the document frequencies of all the keywords by unique pair website-questionnaire year.

We obtained standardised frequencies by dividing the document frequencies of keywords for each firm by the total number of sentences obtained. We operationalised four variable constructs for the purpose of testing our propositions.

To test propositions 2.1 and 2.2, a first construct is obtained with the sum of all the standardised frequencies for each factor to generate the web-based indicator of that factor *(webInnov*, *webCollab*, *webOpen*, *webRD*, and *webIP)*. We then generated an overall innovation index with the sum of all the factors (*webIndex)*.

A second construct is obtained from a PLS regression. Six variables are generated from the first order. Each item is multiplied by the weight obtained from the regression, and the result is divided by the sum of all the weights for each factor resulting in *webInnovPLS*, *webCollabPLS*, *webOpenPLS*, *webRDPLS*, and *webIPPLS*. Then, the overall innovation index (*webIndexPLS)* is obtained from the second-order PLS regression with the same process applied with the five variables previously generated. This method may enhance the resolution of the signal by increasing the relative contribution of the more important items.

For testing most propositions in the third article, we used the original discrete values for the web-based variables, which allow us to control for the size of the website by adding a variable to control that effect in the right side of the equation. Therefore, the third construct use simply the sum of all the frequencies for each factor (*CountInnov*, *CountCollab*, *CountOpen*, *CountRD* and *CountIP*) to build the web-based indicator of that factor, and we obtain the overall index (*CountAll*) by summing the factors.

[18] We used str_detect from the R package stringer.

The fourth set of variables began with the standardised versions obtained from the first construct. Ratios were obtained by dividing each variable by its corresponding maximum value which facilitates the assessment of the relative impacts for each dimension resulting in: *AllRatio*, *InnovRatio*, *CollabRatio, OpenRatio, RDRatio* and *IPRatio*. These variables were not used for formal proposition assessment, but rather for exploring one-way Analysis of Variance (ANOVA) analysis. Table 3-1 presents a summary of all the variables used to test the propositions in the thesis.

Table 3-1. All variables used to test the propositions

Description	Source	Type	Paper 1	Paper 2	Paper 3
Correlation variables					
The number of R&D projects in nanotechnology/advanced material.	Questionnaire	Continuous	ln(nb_R&D_proj)		
The importance of internal R&D as a source of knowledge	Questionnaire	Binary	dInt_R&D_source		
The level of importance of Commercial laboratories/R&D firms/Technical Consultants as a source of knowledge	Questionnaire	Discrete	Ext_R&D_source		
The level of importance of contracting of external R&D service providers	Questionnaire	Discrete	Contract_In_R&D		
The level of importance of providing R&D services to third parties	Questionnaire	Discrete	Contract_Out_R&D		
The time allocated to R&D	Questionnaire	Continuous	ln(time_R&D)		
The importance of private research laboratories/Research and Development firms as collaborators for the development and the commercialisation	Questionnaire	Binary	dCollab_R&D		

Table 3-1. All variables used to test the propositions (continued)

Description	Source	Type	Paper 1	Paper 2	Paper 3
The importance of accessing research and development from collaborators for the development and the commercialisation	Questionnaire	Binary	dCollab_R&D_Reason		
The proportion of Canadian employees assigned primarily in R&D	Questionnaire	Ratio	propEmp_R&D		
The number of IP mechanisms used	Questionnaire	Discrete	nbIP		
The number of patents	Questionnaire	Continuous	ln(nbPatent)		
Whether a firm used collaboration for the latest innovation	Questionnaire	Binary	dCollab		
The proportion of external financing for R&D	Questionnaire	Ratio	PropExtFin_R&D		
The proportion of external financing for commercialisation	Questionnaire	Ratio	PropExtFin_Comm		
The total proportion of external financing versus all financing	Questionnaire	Ratio	propTotExtFin		
Relative importance that a firm assigns to R&D	Questionnaire	Continuous	R&D_INDEX		
Relative importance that a firm assigns to IP	Questionnaire	Continuous	IP_INDEX		
Total occurrences of keyword related to R&D and standardised by the total number of words	Corporate website	Continuous	ln(Web_R&D)		

Table 3-1. All variables used to test the propositions (continued)

Description	Source	Type	Paper 1	Paper 2	Paper 3
Total occurrences of keyword related to IP and standardised by the total number of words	Corporate website	Continuous	1/WEB_IP		
Total occurrences of keyword related to collaboration and standardised by the total number of words	Corporate website	Continuous	ln(WEB_COLLAB)		
Total occurrences of keyword related to external financing and standardised by the total number of words	Corporate website	Continuous	WEB_ExtFin		
Dependent variables					
Measure of the intensity of innovation evidence, ranging from 0 corresponding to no innovation evidence, to 1 with the strongest innovation evidence	Questionnaire	Continuous		InnovEvidence	
Amount of funding or investments received	Questionnaire	Continuous		ln(Investment/50,000+1)	
Number of all sentences with innovation-related terms on a firm's website	Historical corporate website	Discrete			countAll

Table 3-1. All variables used to test the propositions (continued)

Description	Source	Type	Paper 1	Paper 2	Paper 3
Number of all sentences with innovation culture-related terms on a firm's website	Historical corporate website	Discrete			countInnov
Number of all sentences with collaboration-related terms on a firm's website	Historical corporate website	Discrete			countCollab
Number of all sentences with OI-related terms on a firm's website	Historical corporate website	Discrete			countOpen
Number of all sentences with R&D-related terms on a firm's website	Historical corporate website	Discrete			countRD
Number of all sentences with IP-related terms on a firm's website	Historical corporate website	Discrete			countIP
Independent variables					
Number of months needed to introduce new or improved products, processes, or services to the market	Questionnaire	Continuous		ln(TimetoMarket /100+1)	TimetoMarket

Table 3-1. All variables used to test the propositions (continued)

Description	Source	Type	Paper 1	Paper 2	Paper 3
Categorical variable, where 1 is longer time to market, 2 is average time to market, and 3 is shorter time to market	Questionnaire	Ordinal			oTimeMarket
Dummy variable that takes the value 1 if the firm is in the group with a longer time to market (TimetoMarket), and 0 otherwise.	Questionnaire	Binary			dLongTMkt
Dummy variable that takes the value 1 if the firm is neither in the group with a longer or a shorter time to market (TimetoMarket), and 0 otherwise.	Questionnaire	Binary			dAverageTMkt
Dummy variable that takes the value 1 if the firm is in the group with a shorter time to market (TimetoMarket), and 0 otherwise.	Questionnaire	Binary			dShortTMkt
A measure of the degree of innovation evidence, where 1 is no innovation evidence, 2 is some innovation evidence and 3 is strong innovation evidence.	Questionnaire	Ordinal		oInnovEvidence	oInnovEvidence

Table 3-1. All variables used to test the propositions (continued)

Description	Source	Type	Paper 1	Paper 2	Paper 3
Dummy variable that takes the value 1 if the firm exhibits no innovation evidence, and 0 otherwise.	Questionnaire	Binary			dNoInnovEvidence
Dummy variable that takes the value 1 if the firm exhibits some innovation evidence, and 0 otherwise.	Questionnaire	Binary			dSomeInnovEvidence
Dummy variable that takes the value 1 if the firm exhibits strong innovation evidence, and 0 otherwise.	Questionnaire	Binary			dStrongInnovEvidence
Dummy variable that takes the value 1 when a firm uses or develops technologies that require the involvement of STEM disciplines, and 0 otherwise.	Questionnaire	Binary		TechFirm	
The sum of all standardised keyword frequencies on a firm's website	Historical corporate website	Continuous		ln(10*webIndex+1)	
The sum of all standardised keyword frequencies referring to innovation culture on a firm's website	Historical corporate website	Continuous		ln(100*webInnov+1)	
The sum of all standardised keyword frequencies referring to collaboration on a firm's website	Historical corporate website	Continuous		ln(100*webCollab+1)	

Table 3-1. All variables used to test the propositions (continued)

Description	Source	Type	Paper 1	Paper 2	Paper 3
The sum of all standardised keyword frequencies referring to OI on a firm's website	Historical corporate website	Continuous		ln(10,000*webOpen +1)	
The sum of all standardised keyword frequencies referring to R&D on a firm's website	Historical corporate website	Continuous		ln(10*webRD+1)	
The sum of all standardised keyword frequencies referring to IP on a firm's website	Historical corporate website	Continuous		ln(10,000,000*webIP+1)	
The PLS weighted sum of all standardised keyword frequencies on a firm's website	Historical corporate website	Continuous		ln(100*webIndexPLS+1)	
The PLS weighted sum of all standardised keyword frequencies referring to innovation culture on a firm's website	Historical corporate website	Continuous		ln(200*webInnovPLS+1)	
The PLS weighted sum of all standardised keyword frequencies referring to collaboration on a firm's website	Historical corporate website	Continuous		ln(500*webCollabPLS+1)	
The PLS weighted sum of all standardised keyword frequencies referring to OI on a firm's website	Historical corporate website	Continuous		ln(100,00*webOpenPLS+1)	
The PLS weighted sum of all standardised keyword frequencies referring to R&D on a firm's website	Historical corporate website	Continuous		ln(50*webRDPLS+1)	

Table 3-1. All variables used to test the propositions (continued)

Description	Source	Type	Paper 1	Paper 2	Paper 3
The PLS weighted sum of all standardised keyword frequencies referring to IP on a firm's website	Historical corporate website	Continuous		ln(1,000,000,000*webIPPLS+1)	
Categories of knowledge intensity where 1 is for low-tech industries and other services, and 2 is for advanced knowledge services and tech firms.	Questionnaire	Nominal			nKnowledgeIntensity
Dummy variable that takes the value 1 when a firm require the involvement of STEM disciplines, and 0 otherwise	Questionnaire	Binary			dHighKI
Categories of sectors where 1 corresponds to Materials & manufacturing, 2 to Life sciences & health, 3 to Digital media & ICT, 4 to Natural resources, energy & environment and 5 to Other.	Questionnaire	Nominal			nSector
Dummy variable that takes the value 1 if the firm is in the Materials & Manufacturing sector, and 0 otherwise.	Questionnaire	Binary		dMaterialMnfg	dMaterialMnfg
Dummy variable that takes the value 1 if the firm is in the Life sciences & health sector, and 0 otherwise.	Questionnaire	Binary		dLifeSciHealth	dLifeSciHealth
Dummy variable that takes the value 1 if the firm is in the Digital media & ICT sector, and 0 otherwise.	Questionnaire	Binary		dDigMediaICT	dDigMediaICT

Table 3-1. All variables used to test the propositions (continued)

Description	Source	Type	Paper 1	Paper 2	Paper 3
Dummy variable that takes the value 1 if the firm is in the Natural resources, energy & environment sector, and 0 otherwise.	Questionnaire	Binary		dEnerEnvironNat	dEnerEnvironNat
Dummy variable that takes the value 1 if the firm is in any other sector, and 0 otherwise.	Questionnaire	Binary		dOther	dOther
Sectorial categories where 1 is for Supplier-dominated, and 2 is for Specialised-suppliers, 3 for KIBS and 4 for Mass production.	Questionnaire	Nominal			nSectorialCategory
Dummy variable that takes the value 1 if the firm is in the Supplier-dominated sector, and 0 otherwise.	Questionnaire	Binary			dSupplierDominated
Dummy variable that takes the value 1 if the firm is in the Specialised supplier sector, and 0 otherwise.	Questionnaire	Binary			dSpecializedSupplier
Dummy variable that takes the value 1 if the firm is in the KIBS sector, and 0 otherwise.	Questionnaire	Binary			dKIBS
Dummy variable that takes the value 1 if the firm is in the Mass production sector, and 0 otherwise.	Questionnaire	Binary			dMassProd

Table 3-1. All variables used to test the propositions (continued)

Description	Source	Type	Paper 1	Paper 2	Paper 3
Dummy variable that takes the value 1 if the firm is in the scale manufacturing sector, and 0 otherwise.	Questionnaire	Binary			dScaleManuf
Dummy variable that takes the value 1 if the firm is in the science-based sector, and 0 otherwise.	Questionnaire	Binary			dScience
Control variables					
Company's annual revenues	Questionnaire	Continuous		ln(Revenue/1,000 + 1)	
Total number of sentences divided by maximum sentences by size and sector	Historical corporate website	Ratio			SentenceRatio
Categories of countries where 1 corresponds to firms located in Finland, 2 for firms located in Canada	Questionnaire	Nominal			nCountry
Dummy variable that takes the value 1 if the firm is located in Canada, and 0 otherwise.	Questionnaire	Binary			dCanada
Dummy variable that takes the value 1 if the firm is located in Finland, and 0 otherwise.	Questionnaire	Binary		dFinland	
Categories of sizes where 1 corresponds to firms between 1 and 10 employees, 2 for firms between 11 and 99 employees, 3 for firms that have between 100 and 499 employees and 4 for firms with more than 500 employees.	Questionnaire	Nominal			nSize

Table 3-1. All variables used to test the propositions (continued)

Description	Source	Type	Paper 1	Paper 2	Paper 3
Dummy variable that takes the value 1 if the firm has fewer than 10 employees, and 0 otherwise.	Questionnaire	Binary			dVerySmall
Dummy variable that takes the value 1 if the firm has 11-99 employees, and 0 otherwise.	Questionnaire	Binary			dSmall
Dummy variable that takes the value 1 if the firm has 100-499 employees, and 0 otherwise.	Questionnaire	Binary			dMedium
Dummy variable that takes the value 1 if the firm has 500 and more employees, and 0 otherwise.	Questionnaire	Binary			dLarge
Categories of ages where 1 corresponds to firms between 0 and 3 years old, 2 for firms between 4 years old and 8 years old, 3 for more than 9 years old.	Questionnaire	Nominal			nAge
Dummy variable that takes the value 1 if the firm is 3 years old or younger at the time of the questionnaire, and 0 otherwise.	Questionnaire	Binary		dEarlyStage	dEarlyStage
Dummy variable that takes the value 1 if the firm is between 4 and 8 years old at the time of the questionnaire, and 0 otherwise.	Questionnaire	Binary		dYoung	dYoung
Dummy variable that takes the value 1 if the firm is at least 9 years old at the time of the questionnaire, and 0 otherwise.	Questionnaire	Binary		dMature	dMature
Dummy variable that takes the value 1 for an investigation that happened in 2010, and 0 otherwise.	Questionnaire	Binary		d2010	
Dummy variable that takes the value 1 for an investigation that happened in 2011, and 0 otherwise	Questionnaire	Binary		d2011	
Dummy variable that takes the value 1 for an investigation that happened in 2012, and 0 otherwise	Questionnaire	Binary		d2012	

Table 3-1. All variables used to test the propositions (continued)

Description	Source	Type	Paper 1	Paper 2	Paper 3
Dummy variable that takes the value 1 for an investigation that happened in 2013, and 0 otherwise	Questionnaire	Binary		d2013	
Dummy variable that takes the value 1 for an investigation that happened in 2014, and 0 otherwise	Questionnaire	Binary		d2014	d2014
Dummy variable that takes the value 1 for an investigation that happened in 2015, and 0 otherwise	Questionnaire	Binary		d2015	
Dummy variable that takes the value 1 for an investigation that happened in 2015, and 0 otherwise	Questionnaire	Binary		d2016	
Categories of period of year where 1 corresponds to 2010-2011, 2 for 2012-2013, 3 for 2014 and 4 for 2015-2016	Questionnaire	Nominal			nPeriod
Dummy variable that takes the value 1 if the firm has responded to the survey in 2010 or 2011, and 0 otherwise.	Questionnaire	Binary			d2010_2011
Dummy variable that takes the value 1 if the firm has responded to the survey in 2012 or 2013, and 0 otherwise.	Questionnaire	Binary			d2012_2013
Dummy variable that takes the value 1 if the firm has responded to the survey in 2015 or 2016, and 0 otherwise.	Questionnaire	Binary			d2015_d2016

3.4 Methods

This section describes the main methodologies that are used to test our propositions with the Multi-Trait Multi-Method (MTMM) matrix and followed with the regression models. It follows up with a description of the one-way ANOVA analysis and the contribution of the paper published.

3.4.1 Correlation analysis and MTMM matrix

To assess whether the web-based method can accurately represent concrete actions measured by a questionnaire, we compared each pair of variables related to the same concept from both methods (web-based vs. questionnaire-based). We used Pearson's correlation when the variables followed a normal distribution, and Spearman correlation when they did not.

For all outcomes that proved significant and important, we adhered to the methodology outlined in Ortiz de Guinea et al. (2013). We conducted a multitrait-multimethod (MTMM) matrix analysis to evaluate the convergent and discriminant validity of the measures. This was followed by a confirmatory factor analysis (CFA) to determine if any method effects might result in congeneric bias.

Introduced by D. T. Campbell & Fiske (1959), the MTMM matrix is employed to test the convergent and discriminant validity, as well as the reliability, of a construct where a set of traits "T" (interchangeable with factors in the case of this research) is measured using a set of "M" different methods. The primary purpose of this method is to assess the validity of various measures obtained through different methods. Consequently, it ensures that the validity of measurements obtained by vastly different methods (for example, a measurement derived from a Likert scale and one from an electrocardiogram) captures the intended trait. This method is widely used in various social science fields, including psychology, organisational research, marketing research, and information systems. The matrix facilitates easy comparison of how methods and measures correspond with one another.

Reliability is assessed using Cronbach's Alpha, while all other measures are correlations, which enable the evaluation of convergent and discriminant validity. Section 4.5.3 provides an in-depth explanation of how to interpret the matrix results.

3.4.2 Regression methods

Paper 2 and 3 requires two different regression methods.

First, the interrelationship between innovation evidence, investment received, and the web-based innovation signal from propositions 2.1 and 2.2 is examined. *Proposition 2.1* is assessed using a model that takes into account the influence of the web-based innovation signal, time to market, and investment received as factors contributing to the evidence of innovation. Subsequently, proposition 2.2 is tested by evaluating the effect of the web-based innovation signal and whether the firm is a technology-based entity on the investment received. In general, innovation results are likely to contribute to receiving investments (Aghion et al., 2004). However, since the sample primarily consists of smaller firms using the services of innovation intermediaries to enhance their capabilities and access resources, it seems more plausible that they would undergo a process of obtaining investment, which would lead to more innovation results. This approach also simplifies the dynamics and makes it easier to model.

Solving this simultaneous equation system also allows to determine whether the web-based innovation signalling has a greater impact on explaining the variation in either one of the dependent variables.

Because *Investment* is both a dependent variable and is a function of our other dependent variable *InnovEvidence*, it violates the assumption of no correlation between the disturbances or error terms and the explanatory variables in the model assumed with Ordinary least squares (OLS) regression.

As a quick reminder, the disturbance or error term is the difference between the observed values of a dependent variable and the values predicted by the model. If the disturbances are correlated with the explanatory variables, the OLS estimates of the model parameters may be biased and inefficient. In simultaneous equations system, endogenous variables are determined by the values of other variables in the system, and the errors in those equations will be correlated with the errors in the equations that use the endogenous variables as explanatory variables. Thus, *Investment* is an endogenous variable as the error terms are expected to be correlated.

In order to produce consistent estimates of the model parameters, the reg3 function within Stata 16 uses an instrumental-variable approach and generalised least squares (GLS) to account for the correlation structure in the errors across the equations (StataCorp., 2019). This allows the model

to properly account for the correlated errors and estimate the parameters in a way that is unbiased and efficient. Two stage least squares (2SLS) and three-stage least squares (3SLS) regressions are both types of instrumental variable (IV) approaches used to estimate the parameters of a model with endogenous variables in a way that accounts for the endogeneity and produces consistent and efficient estimates. These techniques assume that all dependent variables in the model are endogenous and correlated with the errors in the equations, while all other variables are assumed to be exogenous and uncorrelated with the errors. The exogenous variables are used as instruments to estimate the values of the endogenous variables. The 2SLS and 3SLS method are executed in stages whereas the first two stages are the same for both techniques: 1) developing instrumented values for all endogenous variables; 2) estimating the covariance matrix of the equation disturbances; and 3) in the case of the 3SLS, executing a GLS-type estimation using the instrumented values and the estimated covariance matrix obtained from the second stage. The third stage is especially important when they are contemporaneous correlations suspected between the disturbances (Henningsen & Hamann, 2007).

Thus, 3SLS limits the impact of non-independence of the errors (independently and identically distributed disturbances), the potential bias of non-independence of the errors (heteroskedastic residuals) as well as the non-homoskedasticity of the error variances. The 2SLS estimates are generally consistent without the third stage, but using the covariance matrix improves the efficiency of the three-stage estimator assuming the covariance matrix estimates are consistent, which depends on the errors in the individual equations. If one of the equations in a structural equation model is misspecified, the resulting coefficients will be biased and inconsistent; alternatively, only the coefficients in the misspecified equation will be affected if each equation is estimated separately by 2SLS (StataCorp., 2019). Finally, in order to be able to perform the estimation, the model must not be under-identified, i.e., there must be at least as many noncollinear exogenous variables in the remaining system as there are endogenous right-hand-side variables in each equation, a condition satisfied with our just-identified model.

To ensure the reliability of our models, we will follow a rigorous testing procedure. First, we will examine the independence of our variables of interest and control by assessing the correlation between them. We will then test for multicollinearity by calculating the VIF for each variable after running an OLS regression that clusters observations by company to account for repeated

measures. If VIFs are below 2, we will conclude that there is no multicollinearity (Hair et al., 2010). Then, using Stata16 we tested our models with 2SLS and 3SLS with ireg to minimise the sum of squared errors (it iteratively refines the estimated disturbance covariance matrix and the parameter estimates until it converges to a stable solution).

Next, we will test our models using Stata16 with 2SLS and 3SLS methods. We will compare the results between the 2SLS and 3SLS methods to ensure consistency in the coefficients and their significance level. To test for heteroskedasticity, we will use the Breush-Pagan test with the sureg function.

We will then test for endogeneity using the endogenous post-estimation tests from the ivreg 2SLS, running a Durbin (score) and a Wu-Hausman metric (StataCorp., 2019). If the results of these tests are not significant, for robustness purposes, we will also perform OLS regressions of the model treating each equation separately and compare the results to those obtained with the other methods.

Lastly, to test the majority of the propositions from the third paper, we employ models that explain the variations in signal strength for overall innovation, innovation culture, collaboration, OI, R&D and IP. These models utilise the following explanatory variables: *oInnovationEvidence*, *oTimeMarket*, *nAge*, *nSector*, *nSectorialCategory*, *nKnowledgeIntensity*, *nSize*, *nCountry*. Additionally, *nPeriod* and *SentenceRatio* serve as control variables.

Poisson (Hausman et al., 1984) or Negative binomial (NB) (Greene, 2008; Long, 1997), and their zero-inflated variations (Foltz et al., 2000) are all regression models that can be considered for fitting count data. The NB model exhibits a better fit if the alpha is significantly different from zero (Hilbe, 2011; Long, 1997).

The zero-inflated model consists of two parts: a logit model that fits the excess of zeros distribution and a negative binomial (ZINB) or Poisson (ZIP) model that fits the remainder of the count data distribution. A zero can be interpreted as an absence of signalling. This absence may occur because a firm did not signal any innovation-related terms at all or because we failed to detect it. As the web-based data collection process relies on the Wayback Machine, which introduces selection and non-selection biases, some pages containing innovation-related signalling may not have been retrieved. Furthermore, different innovation-related factors are likely not uniformly distributed among firms. For instance, a significant number of firms may not actively engage in formal intellectual property mechanisms, resulting in a substantial number of zero observations. Similarly,

some terms related to open innovation appear to be used more infrequently in our tests, increasing the likelihood of having no observations for that factor. Even with a factor that has a well-established lexical field, such as R&D, it is possible that some firms do not see the value in signalling its importance on their website. Finally, not all firms in our sample are necessarily innovation-active firms, which means they may not have any innovation factors to signal. The function incorporating both models (logit and Poisson/NB) iterates over the likelihood (LL) of the function until it converges to a maximum value.

The NB/Poisson component of the zero-inflated model is specified in the same manner as it was with the standalone NB/Poisson models. For the logit part, also referred to as the "inflate" component, contextual variables are used to explain the excess of zeros. These variables include *nAge*, *nSector*, *nSectorialCategory*, *nSize*, *nKnowledgeIntensity*, *nCountry*, *nPeriod* and *SentenceRatio.*

The goodness of fit assessment was previously conducted using the Vuong test (Vuong, 1989), but recent research has shown it to be inadequate (Wilson, 2015). Therefore, to determine which model better fits the data, a deviance test[19] is performed:

$$D=-2(LL_{ZI}-LL_{NB/Poisson})$$

where D follows a χ^2 distribution at *p* degree of freedom where *p* is the number of variables in the logit model.

Additionally, since the regression results only display the variations in incidence rate compared to the omitted group, predicted margin effects are calculated using Wald tests. These tests allow for the assessment of differences among other categories. The signal is predicted in terms of occurrences for a target group of firms that remain constant across variables.

3.4.3 One-way ANOVA analysis

To explore the differences within each categorical variable in terms of different signalling dimensions, a one-way ANOVA analysis was conducted, as shown in the supplementary materials of the third paper (see Appendix E). In cases where a categorical variable has more than two categories, a Kruskal-Wallis test is performed to assess the overall differences in mean rank. Due

[19] Test performed with Fitstat (see Long & Freese, 2001)

to the non-normal distribution of the variables and the presence of categories with a small number of firms, the Mann-Whitney U test is used to evaluate one-on-one differences between categories on different signalling dimensions. Lastly, Wilcoxon rank tests are employed to assess the differences in ranking of different signalling dimensions by category. These non-parametric tests are suitable for evaluating the differences in signalling between items with a small number of observations, such as scale-based and science-based firms.

CHAPTER 4 ARTICLE 1: USING WEB CONTENT ANALYSIS TO CREATE INNOVATION INDICATORS – WHAT DO WE REALLY MEASURE?

Mikaël Héroux-Vaillancourt, Catherine Beaudry and Constant Rietsch

This chapter introduces a paper published in the Quantitative Science Studies open-access journal on December 1st, 2020. It describes how we built different web-based indicators based on keyword frequency analysis and assesses their meaning from cross-examination from a questionnaire-based survey. We used a classic construct validation method, the multi-trait multi-method matrix along with a confirmatory factor analysis for robustness check. The results show that web-based indicators seem to better represent higher concepts rather than specific measurements.

4.1 Abstract

This study explores the use of web content analysis to build innovation indicators from the complete texts of 79 corporate websites of Canadian nanotechnology and advanced materials firms. Indicators of four core concepts (R&D, IP protection, collaboration and external financing) of the innovation process were built using keywords frequency analysis. These web-based indicators were validated using several indicators built from a classic questionnaire-based survey with the following methods: correlation analysis, Multi-Traits Multi-Methods (MTMM) matrices and confirmatory factor analysis (CFA). Results suggest that formative indices built with the questionnaire and web-based indicators measure the same concept, which is not the case when considering the items from the questionnaire separately. Web-based indicators can act either as complements to direct measures or as substitutes for broader measures, notably the importance of R&D and the importance of IP protection, which are normally measured using conventional methods such as government administrative data or questionnaire-based surveys.

Keywords: Web content analysis, word frequency analysis, Multi-Traits Multi-Methods, construct validity, Web-mining, innovation measurement

4.2 Introduction

The majority of researchers in innovation and technology management today still rely on public databases and questionnaire-based surveys to obtain most of their data to perform quantitative

studies on industrial strategies and innovation activities. However, public databases are often incomplete or too general in nature. While questionnaires remain precise instruments, the process of designing, testing and administering questionnaire-based surveys can be especially time consuming and costly for researchers. Furthermore, the over solicitation of respondents and of their time plays against questionnaire-based surveys, which suffer from increasingly lower response rates (less than 10%) and thus threaten the validity of studies performed using such methods for instance because of the potential of significant non-response biases.

To complement questionnaire-based data, social scientists have often used secondary data. With the development of "Big Data" analytical tools, websites are increasingly recognised as additional information gold mines. Researchers in innovation and technology management are now investigating whether they can use the information that organisations provide on their websites to acquire valuable additional data for their research. Technology companies generally maintain websites that allow the media, as well as potential investors, customers, suppliers and collaborators, to learn about the nature of their activities. The information provided on company websites is as rich as it is diversified, including products, services, business models, R&D activities, and more. Since these corporate websites are freely available to anyone with Internet access, researchers need to evaluate their value as data sources. Before fully validating their usage, a few questions need to be answered: Is it possible to extract the information contained on these websites and convert it into useful data for research purposes? Is the available information reliable and sufficient to provide an accurate picture of the firm's specific characteristics? In other words, can the contents of a corporate website be used to identify different business innovation attributes?

This research aims to study high technologies that should have emerged by now − nanotechnologies and advanced materials − but that are still difficult to assess because of their pervasiveness throughout the industry. Nanotechnologies and advanced materials, as so-called enabling technologies, have innumerable applications and are present in all major industrial sectors such as food, agriculture, electronics, renewable energy, the environment, biomedical, health care, and so on. The ability to develop and use nanotechnologies and advanced materials is expected to give impetus to innovation performance and economic growth (Hwang, 2010). The versatility of these advanced technologies' applications and the sectors affected by them make them particularly interesting for studies in innovation management since they allow an intersectoral bird's eye view of the industry. However, owing to the pervasiveness of nanotechnologies and advanced materials

throughout the economy, national statistics and other data repositories often fail to accurately measure their impact. A potential solution to alleviate this perceived lack of accurate data lies in the information contained on company websites.

The clear majority of companies working in high-technology fields, such as those producing or using nanotechnology and advanced materials, maintain updated websites. Although the online information is made available by the companies themselves, which suggests the possibility of a strong self-reporting bias, this source of information is deemed suitable for the study of the emerging technologies (Gök et al., 2015). For instance, Youtie et al. (2012) noted that small businesses tend to have less content-heavy websites, which facilitates the handling of data. A successful web mining analysis has several advantages over questionnaires, scientific publications and patents. First, the population covered by searching the Web (Web crawling) is very large (Herrouz et al., 2013) compared to questionnaire-based studies, which generate few returns (low response rate). Contrary to government data, the frequency of updates is high, even daily, in most cases (Gök et al., 2015). One significant drawback, however, is that companies do not disclose all strategic and business data on their websites. The main disadvantage of such web-based data stems from problems inherent in organising and interpreting highly unstructured information, since each site organises various types of information differently. In addition, we do not yet know exactly what is the significance of what we are measuring, hence the necessity to concurrently use both traditional and webometrics methods and to compare them with one another.

In this study, we analysed and compared four sets of measures of innovation and commercialisation respecting nanotechnology and advanced materials firms in Canada stemming from two different data gathering techniques: word frequency analysis from web content and questionnaire-based surveys. Comparisons between the results from both methods were obtained via correlations. To ensure a convergent and discriminant validation of our results, a Multi-Traits Multi-Methods (MTMM) technique was then performed on the most significant measures. Formative indices were built to determine the best representation of the web-based indicators. One final MTMM matrix was used, along with an MTMM confirmatory factor analysis, as a post-hoc analysis to ensure the robustness of our results.

The remainder of the article is organised as follows: Section 4.3 presents the theoretical development of the innovation and commercialisation of high technologies, such as

nanotechnology and advanced materials, along with web content analysis; Section 4.4 describes the methodologies used to collect data and the MTMM method; Section 4.5 analyses the results surrounding propositions 1.1 and 1.2, and performs a post hoc analysis; Section 4.6 discusses the results and conclusions; and lastly, Section 4.7 addresses the research limitations and future research.

4.3 Theoretical development

4.3.1 Using the web as a data source for social sciences

A few years after the invention of the world wide web, its use as a means to get valuable data for social science studies was introduced (Almind & Ingwersen, 1997). With more and more online content from all aspects of society being produced, the web becomes a digital representation of the social construct that embodies our civilisation. With such a rich repertoire of information, social scientists started to explore ways to leverage the web as a data source for research (Björneborn & Ingwersen, 2004). As a result, measuring different elements of the web such as web sites, web pages, parts of web pages, words in web pages, hyperlinks, web search engine results and so on, has been undertaken by a number of scholars which paved the way to a new field of research: webometrics, as an evolution from the classic bibliometric studies (Thelwall, 2009).

Link analysis, blog searching and web impact assessment were the main focus of late 2000s webometrics research. The first, link analysis consists in analysing the network of hyperlinks contained in web pages in the hope that they may reflect an organisation's connections (Hyun Kim, 2012; Katz & Cothey, 2006; Vaughan, 2004). This method has proven useful to understand the structure of the collaborative network and the importance of a given actor within its network (Minguillo & Thelwall, 2012; Stuart & Thelwall, 2006). The interest in the second method, exploring the content of blogs, resides in their vast heterogeneous public-generated information at a given time. Their understanding provides information on the trends in public opinion on a given set of topics. Nowadays, blogs searching is more prevalent in social media platforms such as Twitter (Thelwall et al., 2011), and topics importance can be effectively monitored through Google Trends (Choi & Varian, 2012). The third, web impact assessment, offers a method for measuring the direct online impact of specific topics or documents by counting their presence on the web. These could be a good proxy for offline impact due to the dominant penetration of the internet as means of information and communication (Thelwall, 2009). Web impact assessment could help

identify patterns of diffusion and relative impact of authors, terms, scientific theories, political candidates, books, journals and so on. In the case of innovation studies, the web could be viewed as a way to obtain indirect indicators such as the number of attendees at a given event, the media coverage for a given product launch or the number of requests for an organisation's literature or the number of times an organisation's concepts or publications are mentioned (Thelwall, 2009).

In order to perform web analysis, three main areas of web mining are generally used: web structure mining for link analysis; both web usage mining and web content mining for web impact assessment; and blog searching (Miner et al., 2012). Of particular interest for this paper is web content mining, which is a technique that turns unstructured information in the form of text contained in webpages into structured data useful for research purposes.

Content analysis is an objective and quantitative method used to find relationships between textual information and the context from which the information is sourced (Krippendorff, 1980). The benefits and challenges of using the web as a data source for content analysis were already discussed at the beginning of the millennium (Weare & Lin, 2000). The main opportunity from using the web as a data source stems from the explosion of possible sources of multimedia information (text, video, audio and image) that can be used for content analysis, which was anticipated to reduce the cost of data acquisition significantly. However, the information contained on the web is highly unstructured, completely decentralised and not standardised in any way which make the quality and validity of web content analysis particularly challenging to verify (Weare & Lin, 2000) and somewhat costly to process.

Despite the limitations, some web content analyses were performed in innovation studies. For example, Arora et al. (2013) used a web scraping method to obtain data from small and medium-sized high technology graphene firm websites in the US, UK, and China. More recently, Gök et al., (2015) proposed a web content analysis based on a keywords frequency analysis to assess R&D activities of 296 UK-based green goods small and mid-size enterprises. They compared the results of their R&D web indicator with the results obtained through a questionnaire-based survey. The results of this study showed that the web-based indicator did not correlate significantly with the non-web-based indicator, which suggests that these two indicators did not reflect the same concept. It is therefore possible that web-mining indicators provide new information that was not captured through classical methodologies and thus be shown to be complementary.

4.3.2 Construct definition

Following in the footsteps of Gök et al., (2015), in this study, we focus on parameters influencing innovation and commercialisation in Canadian high-technology firms. We consider the four most important factors known to influence innovation and commercialisation in high-tech firms, in our case, nanotechnologies and advanced materials: Research & Development (R&D) intensity, intellectual property (IP) protection, collaboration and external financing (Lee, Lee, Jhon, & Shin, 2013).

R&D intensity refers to the effort a firm makes to generate inventions (Griliches, 1990, 1994, 1998; Hausman et al., 1984; Hitt et al., 1997) as well as investment in its own absorptive capacity (W. M Cohen & Levinthal, 1990). Greater R&D efforts are likely to yield better returns to innovation. Indeed, the positive relationship between R&D efforts and innovation performance has been demonstrated in several quantitative studies (Baysinger & Hoskisson, 1989; Deeds, 2001; Greve, 2003; Griliches, 1998; Hagedoorn & Cloodt, 2003; Hall, 1990; Parthasarthy & Hammond, 2002). Innovation and R&D efforts have been shown to positively influence firms' commercialisation and financial performance (Geroski et al., 1993; Klette et al., 2000). R&D efforts should therefore give nanotechnology and advanced materials firms a technological superiority as compared to the market. R&D intensity is a measure of intention, and, as such, not a direct measurement of the performance of the R&D process as concerns the production of knowledge, inventions and innovations (R. Adams et al., 2006; Cebon et al., 1999; Flor & Oltra, 2004; Kleinknecht et al., 2002). In addition, not all innovations are systematically derived from this internal R&D process (Michie, 1998), as highlighted by the rising popularity of OI practices (Chesbrough, 2003a). Moreover, measures based solely on R&D are not the best suited for SMEs since their efforts are often neither formal (Kleinknecht et al., 2002) nor constant (Michie, 1998). For these reasons, researchers use R&D intensity less as a proxy for innovation, as suggested by (Becheikh et al., 2006), and more as a means to achieve higher innovation performance.

IP protection confers a competitive advantage to companies by offering exclusivity for the commercialisation of technologies, branding, copyright, industrial design, and so on. (Teece, 1986). Patents enhance the return on investment of the amounts dedicated to technology development through temporary monopoly power, license granting or the sale of patents (Arora, 1995; Chesbrough, 2003a; Feldman & Florida, 1994; Fosfuri, 2006; Mazzoleni & Nelson, 1998;

Merges, 1999; Rivette & Kline, 2000). Laursen and Salter (2006a) observed a positive relationship between appropriability and innovation performance, but one which eventually leads to decreasing returns. A patent, as one of the outputs of the research process, is often considered as a measure of inventiveness (Coombs et al., 1996; Flor & Oltra, 2004; OCDE & Communities, 2005). While patent statistics are frequently used as proxies for innovative activities (Pavitt, 1985) or innovation (Becheikh et al., 2006), they have also been the subject of considerable criticism (Archibugi, 1992; W. M Cohen. & Levin, 1989; Dosi, 1988; Griliches, 1998). The use of patents varies considerably from one sector to another (Archibugi & Sirilli, 2001; Armellini, Beaudry, & Kaminski, 2017; Armellini et al., 2014; Michie, 1998) and according to firm size since SMEs, which have less financial resources to protect their IP, are systematically disadvantaged.

In contrast to the first two types of indicators, collaboration has a much broader impact on the innovation process: from idea generation and research to commercialisation. Collaboration thus has a positive impact on several innovation performance indicators, such as the number of patents, sales growth and the return on innovation sales (Arvanitis, 2012; Belderbos, Carree & Lokshin, 2004; Carboni, 2013), and has been shown to improve competitiveness (Hagedoorn et al., 2000; Roja & Nastase, 2013). A constant collaborative effort is essential for the development and deployment of emerging technologies, which are becoming increasingly complex, as well as for reducing R&D costs, sharing risk and increasing performance (Johnson & Filippini, 2009; Parker, 2000). For example, the aerospace industry requires all actors to adopt a policy of inter-organisational cooperation in order to take advantage of all the necessary knowledge and know-how available (Jordan & Lowe, 2004). In this field, collaboration is so important that it spans all types of partners: clients (Armellini et al., 2017), including suppliers (Bozdogan et al., 1998); competitors (Esposito, 2004; Frear & Metcalf, 1995); and universities and institutes (Armellini et al., 2014, 2017). McNeil, Lowe, Mastroianni, Cronin, and Ferk (2007) have shown that collaboration with universities or government institutes enables young high-tech firms to access particularly expensive tools. In addition, Kim, Lee, and Marschke, 2014 highlighted the impact of university research on scientists associated with the training and development of skilled labour, patents and innovation in industry.

Finally, since most nanotechnology and some advanced material projects are still in the early development stages, they rely heavily on private and/or public funding to reach the commercialisation and marketing phases (Kalil, 2005; McNeil et al., 2007). In surveys on barriers

to innovation, firms very often refer to the lack of external financing as a major obstacle to their innovation activities (Harhoff & Körting, 1998). VC investment in young US high-tech firms had a positive and significant effect on their R&D productivity in the 1990s (Brown et al., 2009). For more than a decade, substantial public investments have targeted nanotechnologies worldwide (Crawley, 2007); more recently, in 2018, USD 1.2 billion were invested in the US (National Nanotechnology Coordination Office, 2017). As the development of these emerging technologies requires public support, the Innovation, Science and Economic Development Canada (ISED, formerly known as Industry Canada) website proposes 13 programs dedicated to funding the development of nanotechnologies and advanced materials. These public funds often have a lever effect and firms that receive such funding in grant form are more likely to also successfully raise private funding and receive better external financing offers (Meuleman & De Maeseneire, 2012).

These four pillars of the innovation process (R&D, IP, collaboration and external financing) have been the focus of much literature and have been extensively measured by both researchers and national statistics offices. But do they transpire in corporate websites? A number of scholars have attempted to analyse web content from the information contained on companies' websites to build indicators and innovation metrics (Gök et al., 2015; Herrouz et al., 2013). Yet understanding the "real" meaning of measures created from data gathered with web mining techniques is not a trivial task. Gök et al. (2015) showed that their web-based indicator based on R&D-related keywords did not correlate significantly with non-web-based indicators, which suggested that these two indicators did not reflect the same concept. Therefore, the web mining indicator could represent new information that was not captured through classical methodologies and could possibly be complementary. In this paper, two propositions will be tested to assess whether these web-based indicators can be used as substitutes for indicators obtained by classical questionnaire-based methods.

Within corporate websites, innovation words/factors are referred to using several synonyms and other related terms. Companies may use words on their website in order to provide insight into what they actually do. The more a company uses terms related to a specific factor, the stronger the signal is deemed for that factor, and therefore, the more likely the firm is liable to perform activities related to that specific factor. Thus, for each factor mentioned above, we suggest the following proposition:

Proposition 1.1: The more words related to a factor, in our case R&D intensity, intellectual property, collaboration and external financing, that are used on a firm's website, the more a firm is likely to perform activities related to that factor.

Connecting keywords found on websites with tangible actions taken by those companies is a bold proposition that may not be valid empirically. The content of corporate websites is written with the aim of signalling the company's "best" characteristics to all its internal and external stakeholders. This represents a rather modest but reasonable claim against the validity of the information to be found, which may suffer from a possible self-reporting bias. However, it is reasonable to assume that the information provided on companies' websites will be true to the company's intentions as to the various concepts it wishes to share with the world. The information that stands out on a website can give a broad sense of the importance of a specific factor for the firm.

Proposition 1.2: The more words related to a factor, in our case R&D intensity, intellectual property, collaboration and external financing, that are used on a firm's website, the more important a firm considers that factor.

The two propositions will be tested and compared to the results obtained with a classic questionnaire-based survey using the methodology explained in section 3.

4.4 Data and Methodology

4.4.1 Questionnaire-Based Data Collection

For a more general study of innovation in nanotechnologies and advanced materials, we first conducted a classic questionnaire-based survey, the core of which is based on the Oslo Manual (OCDE & Communities, 2005), to explore the following themes: innovation, commercialisation, collaboration and IP protection, amongst other topics. The questionnaire includes the importance of the sources of knowledge and the importance of innovation activities measured by a Likert scale from Not important (1) to Essential (7). Other questions cover the importance of commercialisation actions, the proportion of financing for R&D and commercialisation, the importance of market impacts, the number of exports, the importance of obstacles to commercialisation, whether or not the firm collaborated in order to develop or commercialise the latest most significant product innovation, the importance of collaborators, the importance of the reasons for collaboration, IP

mechanisms, patent management, and general questions such as the number of employees, revenue, and the firm's business sector. A sample of the questionnaire is provided in Appendix A.

Due to the ubiquitous nature of nanotechnology and advanced materials, combined with the fact that they are adopted in a large number of industrial sectors, identifying companies that use or develop these technologies is not straightforward. Firms that either use or develop nanotechnology and advanced materials are not labelled as such, nor are they searchable in any obvious way. As a consequence, we used all possible means to build an exhaustive list of high tech companies with a higher probability to be involved in nanotechnology and/or advanced materials with the help of several sources including Industry Canada (now the Innovation, Science and Economic Development Canada), Nano-Québec (now Prima-Québec), Nano Ontario, Nanowerk, futuremarketsinc.com, and AGY Consulting, a Canadian firm specialised in emerging technologies such as nanotechnology, clean technology and biotechnology. Combining all these data sources resulted in more than a thousand unique entries. After manually removing all obvious non-eligible companies, a final list of 592 high tech companies was then put together.

To maximise our response rate, we contracted Léger360, a well-known survey company, to administer the questionnaire and find new companies eligible for our study. Their professional approach is appreciated by firms that have responded to surveys in the past. Data collection began in September 2016 based on a convenience sampling method. In order to even cover more ground, we also used the snowball method. More precisely, respondents first had to answer the eligibility question: "Does your company develop or/and commercialise nanotechnologies and/or advanced materials?" Then, if the firms were eligible, the respondents were asked whether or not they were interested in participating to our study. Finally, they were asked whether they had any recommendations for any potential respondents who might be eligible to our investigation as an attempt to increase our sample size. Firms that agreed to participate in the survey but did not complete the questionnaire received two telephone reminders at one-week intervals. After 332 solicitations, the incidence rate was 28%, which indicated that 93 companies were eligible.

To maximise our sample size, we added companies to the list by using the common characteristics of the eligible respondents. Companies that were eligible for the study (the 93 eligible companies listed above) were grouped and classified according to their North American Industry Classification System (NAICS) codes. A total of 23 six-digit NAICS codes corresponding to 67%

of the eligible companies were identified. Léger360 acquired a list of 3,345 companies representative of the frequency distribution of the NAICS codes obtained from InfoGroup.

A total of 2,971 Canadian high-tech companies were solicited through the entire data acquisition process. Of these, 973 companies did not respond, 1,459 were not eligible, 380 refused to participate, and 168 eligible companies agreed to participate, of which 89 respondents completed the survey. Because of Leger360's code of ethics, we have no means of identifying whether respondents refused to participate before or after answering the eligibility questions. The unknown nature of the non-respondents necessitates an approximation of the response rate. Assuming a uniform distribution, a 12% response rate was obtained by assigning the same characteristics to respondents who answered the eligibility questions to non-respondents and respondents who refused to participate (see Table 4-1). Since the population is unknown, we have a non-probabilistic convenience sample for which the methodology may have induced a selection bias. In addition, we assume that the respondents were honest and answered the survey with goodwill.

Table 4-1. Estimated response rate

Number of high technological Canadian companies:	
Reached (A)	2,971
Did not answer (B)	973
Refused to participate (C)	380
Unknown status (A - B - C = D)	1,618
Not eligible (E)	1,459
[Rate of Not Eligible (E/A = F)]	[49%]
Estimated not eligible (D x F = G)	795
Estimated eligible (A - E - G = H)	718
Total accepted	168
Total completed	89
[Estimated response rate]	[12%]

The resulting sample represents a diverse selection of Canadian high-tech firms that develop or use nanotechnologies or advanced materials. Moreover, 74% of the firms are considered nanotechnology or advanced materials intensive, which means that at least 80% of their revenues come from nanotechnology/advanced materials innovations. The different application domains are

as follows: 54% in advanced materials, 21% in biotechnology and medicine, 24.4% in electronics, 23.3% in equipment and devices, 13.3% in photonics and 33.3% in other domains. More than 50% of respondents are small businesses and 83.5% are SMEs. The $94M revenue average decreases to $31M when the three largest firms are excluded. Finally, 85% of the firms come from Quebec (54.5%) and Ontario (30.7%) and 12% are from British Columbia and Alberta.

In order to test several types of biases, such as self-reporting bias and non-respondent bias, we selected 79 eligible enterprises that did not participate in the study as a control sample. The Government of Canada's[20] Directory of Canadian Companies was used as an external source of data to rule out the possibility of non-response bias. The database of companies from different sectors comprises information provided by the companies themselves on a voluntary basis. While the Government of Canada does not guarantee the accuracy or the reliability of the content, we assumed that the companies that willingly update information in an official public database input accurate information. We therefore assume that this mitigates the self-reporting bias that might arise from this type of source. The database provides the number of employees for 37 firms and revenues for 30 firms from our main sample, as well as the number of employees for 29 firms and revenues for 26 firms from our control sample. Using a two-tailed Mann-Whitney U test, we compared the main sample with the control group for these two metrics. We did not find any significant difference (at the 5% level) between the two samples for either metric (p-value = 0.115; p-value = 0.166 for the number of employees and revenues respectively), which suggests the absence of a non-response bias: there are no significant differences between the characteristics of the respondents and the non-respondents from our list and thus that the sample of eligible companies is homogeneous.

We then compared the data obtained from our questionnaire-based survey with the data from the Directory of Canadian Companies using the same two metrics in order to verify whether an important self-reporting bias exists. For every firm for which we had data from both our questionnaire-based survey and from the Directory, we tested each data pair with a two-tailed Wilcoxon Signed Ranks Test. Once again, we did not find any significant difference (at the 5% level) between our questionnaire-based survey results and the data from the Directory (p-

[20] https://www.canada.ca/en/services/business/research/directoriescanadiancompanies.html, accessed June 4, 2020

value = 0.058; p-value = 0.714), which suggests that the self-reporting bias issuing from the questionnaire-based survey is no different than that found in an official public database.

In order to build the four factors described above, R&D, IP protection, collaboration and external financing, we identified all the relevant questions from the questionnaire-based survey and transformed the answers to these questions into different types of variables. The 12 questions used are listed in Appendix A. Moreover, we transformed every continuous variable from the survey that did not follow a normal distribution by applying a natural logarithm (ln) or an inverse function (inv). Some Likert-scale questions could not be normalised since they were skewed on one tail or the other. We thus transformed them into dummy variables by attributing the value of 1 to answers associated with important (5), very important (6) and essential (7) and the remaining were given the value of 0.

A Principal Component Analysis (PCA) was then performed to validate reflective constructs. Only constructs with sufficiently high Kaiser-Meyer-Olkin –KMO– measures (KMO > 0.6) were deemed valid, and only dimensions with Cronbach's alpha above 0.7 (Hair et al, 1998) were considered reliable. We used PCA with a Varimax rotation on 7-point Likert-scale questions that described the concept of R&D (R&D questions 2, 3, 5 and 6 in Appendix A) to explore whether any particular combination of variables could lead to relevant reflective dimensions corresponding to specific factors of the concept examined. Two factors were created, but both the KMO measure and Cronbach's alpha did not reach an acceptable level (KMO < 0.6, alpha < 0.7) to satisfy the validity and reliability of the constructs. In addition, these combined variables did not correlate with each other, which suggested using a formative construct. We thus proceeded to treat each item individually rather than as a composite indicator.

In the end, we generated a total of nine variables pertinent to R&D, one variable relating to collaboration, two variables corresponding to external financing and finally, two variables measuring IP. Table 4-2 describes the details of the questionnaire-based survey constructs. Descriptive statistics are presented in Appendix A in Table A-2.

Table 4-2. Questionnaire-based survey construct

Concepts	Indicators	Variables
R&D	Number of R&D projects in nanotechnology/advanced materials	ln(nb_R&D_proj) (Continuous)
	Dichotomised importance of internal R&D as a source of knowledge	dInt_R&D_source (Dummy)
	Level of importance of Commercial laboratories / R&D firms / Technical Consultants as a source of knowledge	Ext_R&D_source (Ordinal)
	Level of importance of contracting of external R&D service providers	Contract_In_R&D (Ordinal)
	Level of importance of providing R&D services to third parties	Contract_Out_R&D (Ordinal)
	Time of R&D	ln(time_R&D) (Continuous)
	Dichotomised importance of Private research laboratories / Research and Development firms as collaborators for the development and the commercialisation	dCollab_R&D (Dummy)
	Dichotomised importance of accessing research and development from collaborators for the development and the commercialisation	dCollab_R&D_Reason (Dummy)
	Proportion of Canadian employees assigned primarily in R&D (%)	propEmp_R&D (Continuous)
IP	Number of IP mechanisms used	nbIP (Count data)
	Number of patents	ln(nbPatent) (Continuous)
Collaboration	Use of collaboration for the latest innovation	dCollab (Dummy)
External financing	Proportion of external financing for R&D (%)	PropExtFin_R&D (Continuous)
	Proportion of external financing for commercialisation (%)	PropExtFin_Comm (Continuous)
	Total proportion of external financing (%)	propTotExtFin (Continuous)

Note: All continuous and ordinal variables are normal.

4.4.2 Web-Based Data Collection

While we were processing the questionnaire-based survey data, we initialised the indexing robot with Nutch[21], an open source Web scraper software. We provided the robot with an initial list of uniform resource locators (URL) corresponding to the 89 enterprises that compose our questionnaire-based dataset. This list is the first level of indexing traversal by the robot, which then browses the structure of the entire website to extract all the URLs. To begin the data collection process, the URLs linked to a company's website are written in a text file. The URLs are then injected into Nutch's database.

Once the database has been populated with the list of sites to be visited, the robot's route is generated by indicating the maximum number of links per page that we want to collect. The robot ranks the pages with a score that it calculates to prioritise the text collection. For instance, a page will have a high score if it is pointed to by many other pages. In addition, the higher the scores of pages pointing to another page, the higher the score of the page pointed to. Nutch then selects the links of the pages with the highest score to be collected first. When the links to browse are generated, Nutch will collect all the HTML code from the pages as well as the links to the other pages.

There are mainly three different types of links that structure the Internet network. Upstream links start from a page to one or more pages not necessarily on the same site. Downstream links are all links pointing to the same page. Finally, vertical links are all links within the same website (Van de Lei & Cunningham, 2006). The indexing robot scans all links without distinguishing between domain names and ranks them by score to first scan the links it considers most interesting. We have therefore limited the robot to the vertical links of the domain names present in the initial list. This is possible by modifying the Nutch configuration file limiting the robot to the initial list of domains.

By default, Nutch indexes only the text of pages in HyperText Markup Language (HTML) format. However, relevant information was also stored in other digital formats. We used the Tika plug-in

[21] We used Apache Nutch 2.2.1 version. Documentation can be found on https://nutch.apache.org/apidocs/apidocs-2.2.1/index.html, accessed June 4, 2020

to browse and index the text of documents in Portable Document Files (.pdf) and Word documents (.doc) format.

A further difficulty stems from the fact that our sample included companies that use multiple languages in their websites. Most Canadian companies will have English or French content and some will have other languages. Therefore, we did take into account the fact that the information could be given in languages other than English. There were several different cases to deal with. If a website was entirely in English, which corresponds to a majority of Canadian company websites, we then kept all the data from the website. If a website was available in several languages, including English, then only pages written in English were taken into account. If the website was exclusively available in French, the text was translated in English. Finally, the last case was a page whose language was neither French nor English. Such a page was not considered eligible for our study and, therefore, was not kept for the rest of the study.

The language detection was performed by the Compact Language Detector software while the translation was performed with Google's programming interface, Goslate. Translation with Goslate was also used by Arora et al (2013) to analyse the content of Chinese graphene-producing company websites.

We assigned to each recovered page a label of language, "French", "English" or "other" and translated the French pages into English for the relevant sites according to our assumptions.

Due to technical limitations, such as the structure of the websites, only 79 of these firms (88%) provided enough information to be included in our study. In the end, more than 9.7 million words from 27,000 pages were extracted from the complete texts of corporate websites. We then used a word frequency analysis on the text present in the websites. More specifically, for the 79 websites captured, we gathered information on the four innovation and commercialisation factors mentioned above. For each factor, we listed all the relevant keywords that appear in corporate websites. Figure 4-1 shows the complete process to build the web-based innovation indicators to be compared with the questionnaire-based indicators described in the previous section.

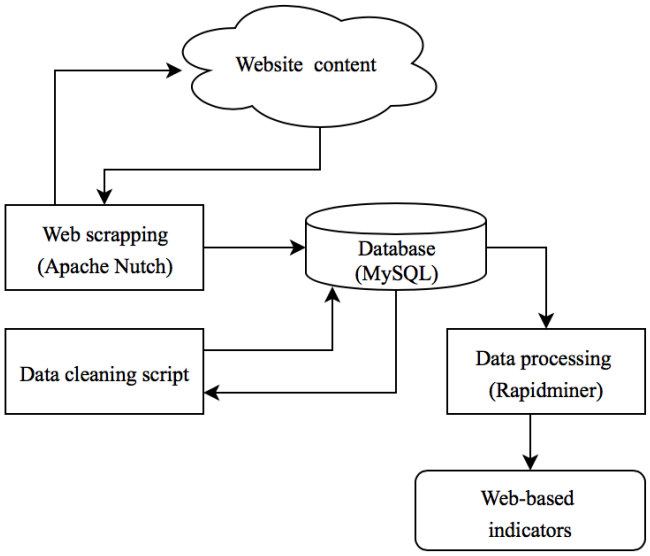


Figure 4-1. Data mining and treatment process.

Factors, keywords and the web mining constructs are described in Table 4-3. R&D (Gök et al., 2015) and collaboration (Ramdani, 2014) keywords were selected from the literature, while IP and external financing were identified from our own investigations of the literature. The most relevant keywords of any paper, which are generally listed on its first page, particularly in the abstract[22], serve as a basis for the list of keywords used for the construction of our factors. We used additional keywords related to specific public funds and programs. As mentioned earlier, the Government of Canada's public programs and funding opportunities for companies for the development of nanotechnology and advanced materials projects are listed on the Innovation, Science and Economic Development (ISED) Canada website. Specific information on these funds and programs was converted into keywords tied to the external funding factor for the individual firms for which we extracted website data. Keywords were manually tested to ensure they did not lead to false positive results.

[22] For this pilot/exploratory study, we did not mine entire articles to select the most common or important keywords related to specific contexts.

Table 4-3. The web mining construct

Factors	Keywords	Indicators	Variables
R&D	research & development, research and development, r&d, researcher, product development, technology development, technical development, development phase, development program, development process, development project, development cent, development facility, technological development, development effort, development cycle, development research, development activity, fundamental research, basic research (Gök et al, 2014)	Natural logarithm of the number of keywords frequencies divided by the total number of words multiplied by 1000	ln(WEB_R&D)
IP	patent, intellectual property, trade secret, industrial design	Inverse (1/x) of keywords frequencies divided by the total number of words multiplied by 1000	inv(WEB_IP)
Collaboration	affiliation, collaboration, cooperation, partners, partnership, consorti, international consorti, global consorti	Natural logarithm of the number of keywords frequencies divided by the total number of words multiplied by 1000	ln(WEB_Collab)
External financing	atlantic canada opportunities agency, business development bank of Canada, sustainable development technology, venture capital, atlantic innovation fund, nrc-irap, fednor, Industrial research assistance program, grants, private investment	The number of keywords frequencies divided by the total number of words multiplied by 1000	WEB_ExtFin*

Note: * All transformed variables are normal except WEB_ExtFin which could not be normalised by any transformation.

Clustering using keyword frequency analysis with RapidMiner, a text mining software, enabled us to count the number of occurrences of each keyword for each factor. We transformed these clusters of occurrences into four continuous variables. Because the 79 companies differ in structure and size, and therefore present different quantities of information on their websites, we standardised each variable by dividing all occurrences by the total number of words appearing on their website

and multiplied the resulting value by 1,000. For each continuous variable, we calculated the Kurtosis and Skewness measures to determine whether they followed a normal distribution. Given that none of the four variables did so, they were transformed by applying a natural logarithm (ln) or an inverse function (inv). In the case of external financing, since we did not reach normality with either transformation, this variable was therefore treated using non-parametric tests that do not require normality.

A possible selection bias arises from the fact that we selected only those companies that answered the survey. To verify whether this was the case, we ran our web mining program on the websites of the control sample and generated the same variables. We then used a two-tailed Student's t-test to test the difference in means for the following variables: ln(WEB_R&D) (p-value = 0.130), inv(WEB_IP) (p-value = 0.083) and ln(WEB_Collab) (p-value = 0.144), and concluded that there is a non-significant (at 5% level) difference between the two samples for these three variables. Finally, using a two-tailed Mann-Whitney U test, we compared the means of the variable WEB_ExtFin (p-value = 0.008) from both our sample and the control sample and found a significant difference. For this particular variable, we cannot conclude that the means of the two samples are the same. A selection bias is therefore present for the variable WEB_ExtFin and will be included in the limitations of our research.

4.4.3 The Multi-Trait Multi Method (MTMM) matrix technique

The literature describes two types of biases that can threaten construct validation: 1) mono-method biases can occur when the by-products from the unique combination of a trait (in our case an innovation factor such as R&D, IP, collaboration and external financing) and a measure introduce a systematic variance; and 2) by-products that are congeneric, i.e. inherent to the methods used, generate anomalies among the items forming the measures (Ortiz de Guinea et al., 2013; Reinig et al., 2007; Straub et al., 1995; Straub & Burton-Jones, 2007).

First introduced by D. T. Campbell & Fiske (1959), the Multi-Trait Multi-Method (MTMM) is designed for the convergent and discriminant validation of a construct where a set of t traits (interchangeable with factors in our case) are measured with m different methods. It has been suggested that this method is an effective way to verify the ability of new measurement methods to successfully measure what they are supposed to, while testing for the presence of mono-method biases in social science studies such as psychology (Bar-Anan & Vianello, 2018; Guo et al., 2008),

education (C. M. Campbell et al., 2019; Gulek, 1999), organisational research (Bagozzi et al., 1991), marketing research (Lugtig, 2017), and information systems (Ortiz de Guinea et al., 2013). Even though its use for construct validity has been employed selectively since its introduction in 1959, the MTMM matrix is still one of the very few possible ways to attempt construct validation and this is why we are witnessing a resurgence of its use in the past few years. To our knowledge this validation method has not been used in innovation studies, nor to assess new innovation metrics compared to the traditional means of measuring innovation, à la Oslo Manual (OCDE & Communities, 2005; OCDE & Eurostat, 2019).

An example of an MTMM matrix with three traits (A, B, C) measured using three different methods is shown in Table 4-4. The matrix easily allows the visualisation of four different means of assessing the quality of a measure: reliability, convergent validation, absence of either a combination of trait and method effects or a mono-method bias and, finally, discriminant validation (D. T. Campbell & Fiske, 1959).

Table 4-4. The MTMM matrix technique

	Traits	**Method 1**			**Method 2**		
		A1	B1	C2	A2	B2	C2
Method 1	A1	$\propto_{A1}$					
	B1	$r_{A1,B1}$	$\propto_{B1}$				
	C1	$r_{A1,C1}$	$r_{B1,C1}$	$\propto_{C1}$			
Method 2	A2	$r_{A1,A2}$	$r_{B1,A2}$	$r_{C2,A2}$	$\propto_{A2}$		
	B2	$r_{A1,B2}$	$r_{B1,B2}$	$r_{C2,B2}$	$r_{A2,B2}$	$\propto_{B2}$	
	C2	$r_{A1,C2}$	$r_{B1,C2}$	$r_{C2,C2}$	$r_{A2,C2}$	$r_{B2,C2}$	$\propto_{C2}$

The diagonal of the matrix (also called reliability diagonal) comprises six Cronbach alpha values (α_{A1}, α_{B1}, α_{C1}, α_{A2}, α_{B2} and α_{C2}), which measure the correlation of the traits with themselves and are thus an indication of the reliability of each measure for each trait (D. T. Campbell & Fiske, 1959). These monotrait-monomethod values indicate the reliability of a reflective measure (D. T. Campbell & Fiske, 1959), which is required to ensure that a measure will produce the same result under the same conditions. High Cronbach's alpha values on the matrix diagonal confirm the

reliability of the traits, that is, that the variance of the measures behaves systematically (Churchill, 1979). In other words, any possible occurrence of a random error is minimised.

The diagonal at the cross-section of Method 1 and Method 2 (also called monotrait-heteromethod, or validity diagonal) comprises three correlation values ($r_{A1,A2}$, $r_{B1,B2}$, and $r_{C1,C2}$) that compare one trait measured with two different methods. High and significant values confirm the convergent validity of the methods (D. T. Campbell & Fiske, 1959), that is, that two different methods measure the same traits (Churchill, 1979).

Two triangles, called heterotrait-monomethod triangles, represent the cross-sections of traits that belong to the same method within the matrix. They comprise three correlation values per triangle. For instance, the heterotrait-monomethod triangle of method 1 comprises $r_{A1,B1}$, $r_{A1,C1}$, and $r_{B1,C1}$, while that of method 2 comprises $r_{A2,B2}$, $r_{A2,C2}$, and $r_{B2,C2}$ (D. T. Campbell & Fiske, 1959). High correlation values (higher than the validity diagonal) within the heterotrait-monomethod triangle question the validity of the construct because of either a combination of trait and method effects or a mono-method bias.

Finally, two other triangles that represent the cross-sections of traits that belong to different methods within the matrix (also called heterotrait-heteromethod triangles) comprise three correlation values per triangle. The heterotrait-heteromethod triangle of method 1 comprises $r_{A1,B2}$, $r_{A1,C2}$, and $r_{B1,C2}$, while that of method 2 comprises $r_{B1,A2}$, $r_{C2,A2}$, and $r_{C2,B2}$ (D. T. Campbell & Fiske, 1959). Again, high values (higher than the validity diagonal) question the discriminant validity of the methods. The discriminant validity is used to assess whether a method fails to measure something it is not supposed to measure effectively (Peter & Churchill, 1986).

Bagozzi et al. (1991) have raised questions about the reasonability of the criteria needed, the lack of standard practice, the negligence of the amplitude of the differences between pairs of correlations in the analysis, and the lack of information about the nature of the variation in measures due to traits, method or random error. Fiske and D. T. Campbell (1992) listed the exceptional conditions required to successfully benefit from the method. Bagozzi et al. (1991) suggested the confirmatory factor analysis (CFA) as an alternative to the MTMM for construct validation, which, although it does not share the same limitations, is unable to differentiate the random error from the variance measure and unable to verify interactions between traits and methods. Therefore, performing both methods will mitigate the limitations inherent in either

method taken individually and ensure robustness of the results. In this paper, we adopted the methodology used by Ortiz de Guinea et al. (2013), who opted for an MTMM analysis followed by a post-hoc analysis with a CFA to validate their results.

4.5 Results

4.5.1 Keywords analysis

More than 9.7 million words from 27,000 pages were extracted from the complete texts of 79 corporate websites. The details of the frequency analysis for each indicator is provided in Table 4-5. For the R&D factor, a total of 1,974 keywords were counted, from which six keywords (r&d, scientist, laboratory, research and development, researcher, laboratorie) account for more than 85% of the total frequency distribution. For IP, 75% of the total frequency distribution (397) is embodied by the keyword "patent", hence suggesting that most of the signal provided by the IP indicator is actually related to patenting. This is hardly surprising considering our sample of nanotech firms. For collaboration, a total of 1,769 keywords were counted, from which, more than half is associated with partnering activities. On external financing, we only obtained a total of 251 keywords, from which, more than 80% of the frequency distribution was dominated by the word "grant".

Table 4-5. Keywords Frequencies

R&D		**IP**		**Collaboration**		**External financing**	
r&d	480	patent	300	partner	967	grant	202
scientist	368	intellectual propert	84	partnership	422	venture capital	31
laboratory	274	trade secret	8	collaboration	132	business development bank of canada	7
research and development	246	patent pending	4	alliance	88	industrial research assistance program	4
researcher	185	patent protect	1	collaborative	62	private investment	4
laboratorie	165			cooperation	35	atlantic canada opportunitie	1

Table 4-5. Keywords Frequencies (continued)

R&D		**IP**		**Collaboration**		**External financing**	
product development	99			partnering	34	public fund	1
technology development	64			collaborating	18	public funding	1
research & development	33			cooperative	9		
development phase	15			cooperating	2		
development project	12						
development effort	10						
basic research	5						
technological development	4						
development program	4						
fundamental research	2						
development cycle	2						
development center	2						
development research	1						
development process	1						
development facility	1						
development faciliti	1						
Total	1,974	397		1,769		251	
Mean	89.72	66.17		147.41		31.38	
Std	138.23	118.91		283.20		69.67	

4.5.2 Testing proposition 1.1

Construct validation

Each pair of variables related to the same concept from the two methods (web mining and questionnaire-based survey) was compared, using a Pearson correlation analysis when the variables followed a normal distribution, and a Spearman correlation when they did not, to assess whether the variables stemming from the web mining technique could be used as a proxy for similar concepts measured by a survey. Details on the construct comparison of the web-based indicators and the questionnaire-based indicators are provided in Table 4-6.

Table 4-6. Validation construct for testing proposition 1.1

Concepts	Web-based Variables	Questionnaire Variables
R&D	ln(WEB_R&D)	ln(nb_R&D_proj) dInt_R&D_source Ext_R&D_source Contract_In_R&D Contract_Out_R&D ln(time_R&D) propEmp_R&D dCollab_R&D dCollab_R&D_Reason
IP	Inv(WEB_IP)	nbIP ln(nbPatent)
Collaboration	ln(WEB_Collab)	dCollab
External financing	WEB_ExtFin	propExtFin_R&D propExtFin_Comm propTotExtFin

Correlation results

All correlation results between the web-based indicators and the questionnaire-based indicators are detailed in Appendix A in Table A-3. For the purpose of this study, the correlations were tested with two-tailed test at the 5% level of significance. Our results show a correlation of 0.306 (*p*-

value < 0.01) between the R&D web-based indicator and whether a firm is likely or not to provide R&D services to third parties (from the questionnaire). Additionally, we find a correlation of 0.306 (*p-value* < 0.01) when we associate the R&D web-based indicator with whether or not a firm has a high percentage of employees working on R&D tasks. Moreover, a correlation of 0.284 (*p-value* < 0.05) is observed between the R&D web-based indicator and whether or not a firm is likely to contract R&D services from external providers. Finally, we find a non-significant correlation of 0.197 (*p-value* = 0.100) when we associate the R&D web-based indicator with whether or not a firm has a long R&D process. The fact that the variable related to the number of R&D projects is not correlated with our R&D web-based indicator (r = 0.002; p-v*alue* > 0.1) strongly suggests that the latter does not properly account for the size dimension of R&D activities.

The IP web-based indicator strongly correlates with the variables from the survey regarding the use of IP protection mechanisms (r=0.368; *p-value* =< 0.01) and patent-related activities (r = 0.396; *p-value* = 0.5). The web content analysis method seems to be able to appropriately capture the importance of the use of IP mechanisms using our sample.

Relations between the web-mining and the questionnaire-based indicators for the other two factors are not as strong. For instance, the collaboration web-based indicator is weakly correlated and non-significant at 5% (r = 0.222; p-v*alue* < 0.1) with the questionnaire-based indicator for the firms that are confirmed as having collaborated. Finally, the external financing web-based indicator is also weakly correlated and non-significant at 5% with the extent of the use of external funding for commercialisation purposes (r = 0.222; p-v*alue* < 0.1).

Consequently, we cannot reach a definite conclusion regarding proposition 1.1, especially in the case of our web-based indicators for collaboration and external financing (a less constraining significance level of 10% would be required to accept the correlations). In the following paragraphs, we will restrict analyses to R&D and IP factors to test propositions 1.1 and 1.2.

MTMM results

All the variables used to measure the R&D and IP factors from the questionnaire-based survey that correlated significantly with our web-based indicators were selected. As explained in section 3.3, the MTMM compares different combinations of methods that measure the same traits (factors). Thus, all possible combinations of traits/methods need to be tested. In the case of the web-based indicators, only one combination of trait and method is used. For the questionnaire-based

indicators, three different variables are correlated with the R&D web-based indicator and two different variables are correlated with the IP web-based indicator. As such, six different combinations are possible and thus six different matrices are built. To ease the reader's comprehension of the interpretation of an MTMM matrix, Table 4-7 shows an example of how the results are displayed, along with a short analysis of all the different validity tests below the table. Furthermore, all traits are considered to be single items, which implies that we cannot calculate the reliability of the measure. Accordingly, the reliability diagonal will be neglected in our analysis.

Table 4-7. MTMM matrix output example

		Web method		**Questionnaire method**	
	Traits	RD1	IP1	RD2	IP2
Web method	RD1	-[a]			
	IP1	$r_{RD1,IP1}$	-[a]		
Questionnaire method	RD2	$r_{RD1,RD2}$	$r_{IP1,RD2}$	-[a]	
	IP2	$r_{RD1,IP2}$	$r_{IP1,IP2}$	$r_{RD2,IP2}$	-[a]

Convergence validity: **Yes** if $r_{RD1,RD2}$ and $r_{IP1,IP2}$ are high and significant;

Discriminant validity: **Yes** if $r_{RD1,RD2}$ and $r_{IP1,IP2}$ are both[b] > $|r_{RD1,IP2}|$ and $|r_{IP1,RD2}|$;

Construct validity without either a combination of trait and method effects or a mono-method bias: **Yes** if $r_{RD1,RD2}$ and $r_{IP2,IP2} > |r_{IP1,RD1}|$ and $|r_{RD2,IP2}|$.

Notes (common to Table 8 to Table 13):

[a] All traits are measured by single items, no reliability statistic can be calculated.

[b] *To increase the readability of the results in each table what we mean by "are both" is:* $r_{RD1,RD2} > |r_{RD1,IP2}|$, $r_{RD1,RD2} > |r_{IP1,RD2}|$, $r_{IP2,IP2} > |r_{RD1,IP2}|$ and $r_{IP2,IP2} > |r_{IP1,RD2}|$

* $p < 0.05$; ** $p < 0.01$

In all matrices, we can observe that the heterotrait-monomethod value is low and not significant for the web-based method (rRD1,IP1 = -0.180; p-value > 0.05) and lower than all corresponding monotrait-heteromethod diagonal values (rRD1,RD2 and rIP1,IP2). This suggests that there is no combination of trait and method effects and that no method bias for the web-based method is present. The complete analysis of all the different matrices results is shown in Table 4-8 to Table 4-13.

Table 4-8. Level of importance of contracting R&D with the number of IP mechanisms used

		Web method		**Questionnaire method**	
	Traits	WEB_RD	WEB_IP	Contract_In_R&D	nbIP
Web method	WEB_RD	-a			
	WEB_IP	-0.18	-a		
Questionnaire method	Contract_In_R&D	0.283*	0.209	-a	
	nbIP	0.007	0.368**	0.433**	-a

Convergence validity: **Yes**, $r_{RD1,RD2} = 0.283^{*}$ and $r_{IP1,IP2} = 0.368^{**}$ are high and significant;

Discriminant validity: **Yes**, $r_{RD1,RD2} = 0.283^{*}$ and $r_{IP1,IP2} = 0.368^{**} >> |r_{RD1,IP2}| = 0.007$ and $|r_{IP1,RD2}| = 0.209$;

Construct validity without either a combination of trait and method effects or a mono-method bias: **No**, $|r_{RD2,IP2}| = 0.433^{**} >> r_{RD1,RD2} = 0.283^{*}$ and $r_{IP1,IP2} = 0.368^{**}$.

Table 4-9. Level of importance of contracting R&D with the number of patents

		Web method		**Questionnaire method**	
	Traits	WEB_RD	WEB_IP	Contract_In_R&D	ln(nbPatent)
Web method	WEB_RD	-[a]			
	WEB_IP	-0.18	-[a]		
Questionnaire method	Contract_In_R&D	0.283*	0.209	-[a]	
	ln(nbPatent)	0.044	0.395*	-0.077	-[a]

Convergence validity: **Yes**, $r_{RD1,RD2} = 0.283^{*}$ and $r_{IP1,IP2} = 0.395^{*}$ are high and significant;

Discriminant validity: **Yes**, $r_{RD1,RD2} = 0.283^{*}$ and $r_{IP1,IP2} = 0.395^{*} >> |r_{RD1,IP2}| = 0.044$ and $|r_{IP1,RD2}| = 0.209$;

Construct validity without either a combination of trait and method effects or a mono-method bias:
Yes, $r_{RD1,RD2} = 0.283^{*}$ and $r_{IP1,IP2} = 0.395^{*} >> |r_{IP1,RD1}| = 0.18$ and $|r_{RD2,IP2}| = 0.077$.

Table 4-10. Level of importance of providing R&D with the number of IP mechanisms used

		Web method		**Questionnaire method**	
	Traits	WEB_RD	WEB_IP	Contract_Out_R&D	nbIP
Web method	WEB_RD	-a			
	WEB_IP	-0.18	-a		
Questionnaire method	Contract_Out_R&D	0.306*	0.068	-a	
	nbIP	0.007	0.368**	-0.066	-a

Convergence validity: Yes, $r_{RD1,RD2}$ = 0.306* and $r_{IP1,IP2}$ = 0.368** are high and significant;

Discriminant validity: Yes, $r_{RD1,RD2}$ = 0.306* and $r_{IP1,IP2}$ = 0.368** >> $|r_{RD1,IP2}|$ = 0.007 and $|r_{IP1,RD2}|$ = 0.068;

Construct validity without either a combination of trait and method effects or a mono-method bias:
Yes, $r_{RD1,RD2}$ = 0.306* and $r_{IP1,IP2}$ = 0.368** >> $|r_{IP1,RD1}|$ = 0.18 and $|r_{RD2,IP2}|$ = 0.066.

Table 4-11. Level of importance of providing R&D with the number of patents

		Web method		**Questionnaire method**	
	Traits	WEB_RD	WEB_IP	Contract_Out_R&D	ln(nbPatent)
Web method	WEB_RD	-[a]			
	WEB_IP	-0.18	-[a]		
Questionnaire method	Contract_Out_R&D	0.306**	0.068	-[a]	
	ln(nbPatent)	0.044	0.396*	-0.134	-[a]

Convergence validity: **Yes**, $r_{RD1,RD2}$ = 0.306** and $r_{IP1,IP2}$ = 0.396* are high and significant;

Discriminant validity: **Yes**, $r_{RD1,RD2}$ = 0.306** and $r_{IP1,IP2}$ = 0.396* >> $|r_{RD1,IP2}|$ = 0.044 and $|r_{IP1,RD2}|$ = 0.068;

Construct validity without either a combination of trait and method effects or a mono-method bias:
Yes, $r_{RD1,RD2}$ = 0.306** and $r_{IP1,IP2}$ = 0.396* >> $|r_{IP1,RD1}|$ = 0.18 and $|r_{RD2,IP2}|$ = 0.134.

Table 4-12. Proportion of employees assigned primarily in R&D with the number of IP mechanisms used

		Web method		**Questionnaire method**	
	Traits	WEB_RD	WEB_IP	propEmp_R&D	nbIP
Web method	WEB_RD	-[a]			
	WEB_IP	-0.18	-[a]		
Questionnaire method	propEmp_R&D	0.306**	0.288*	-[a]	
	nbIP	0.007	0.368**	0.284*	-[a]

Convergence validity: **Yes**, $r_{RD1,RD2} = 0.306^{**}$ and $r_{IP1,IP2} = 0.368^{**}$ are high and significant;

Discriminant validity: **Weak yes**, $r_{RD1,RD2} = 0.306^{*} >> |r_{RD1,IP2}| = 0.007$, but $r_{RD1,RD2} = 0.306^{*} > |r_{IP1,RD2}| = 0.288^{*}$ which is also high and significant; finally, $r_{IP1,IP2} = 0.368^{**} >> |r_{IP1,RD2}| = 0.288^{*}$ and $|r_{RD1,IP2}| = 0.007$;

Construct validity without either a combination of trait and method effects or a mono-method bias:
Weak yes, $r_{RD1,RD2} = 0.306^{*} >> |r_{IP1,RD1}| = 0.18$, but $r_{RD1,RD2} = 0.306^{*} > |r_{RD2,IP2}| = 0.284^{*}$ which is also high and significant; finally, $r_{IP1,IP2} = 0.368^{**} >> |r_{IP1,RD1}| = 0.18$ and $|r_{RD1,IP2}| = 0.284$*.

Table 4-13. Proportion of employees assigned primarily in R&D with the number of patents

		Web method		**Questionnaire method**	
	Traits	WEB_RD	WEB_IP	propEmp_R&D	ln(nbPatent)
Web method	WEB_RD	-[a]			
	WEB_IP	-0.18	-[a]		
Questionnaire method	propEmp_R&D	0.306**	0.455**	-[a]	
	ln(nbPatent)	0.044	0.396*	0.02	-[a]

Convergence validity: **Yes**, $r_{RD1,RD2} = 0.306^{**}$ and $r_{IP1,IP2} = 0.396^{*}$ are high and significant;

Discriminant validity: **No**, $r_{RD1,RD2} = 0.306^{**}$ and $r_{IP1,IP2} = 0.396^{*} << |r_{IP1,RD2}| = 0.455^{**}$;

Construct validity without either a combination of trait and method effects or a mono-method bias:
Yes, $r_{RD1,RD2} = 0.306^{**}$ and $r_{IP1,IP2} = 0.396^{*} >> |r_{IP1,RD1}| = 0.18$ and $|r_{RD2,IP2}| = 0.02$.

Table 4-14. Summary of construct validity results

Table		Convergent validity	Discriminant validity	Construct validity without either a combination of trait and method effects or a mono-method bias
4-8	Level of importance of contracting R&D with the number of IP mechanisms used	Yes	Yes	No
4-9	Level of importance of contracting R&D with the number of patents	Yes	Yes	Yes
4-10	Level of importance of providing R&D with the number of IP mechanisms used	Yes	Yes	Yes
4-11	Level of importance of providing R&D with the number of patents	Yes	Yes	Yes
4-12	Prop. of employees assigned primarily in R&D with the number of IP mechanisms used	Yes	Weak yes	Weak yes
4-13	Prop. of employees assigned primarily in R&D with the number of patents	Yes	No	Yes

Of the six MTMM matrices produced, four matrices yield acceptable results:

- Level of importance of contracting R&D with the number of patents (
- Table 4-9);
- Level of importance of providing R&D with the number of IP mechanisms used (Table 4-10);
- Level of importance of providing R&D with the number of patents (Table 4-11);
- Proportion of employees assigned primarily in R&D with the number of IP mechanisms used (although it is a weak construct) (Table 4-12).

The rejection of the validity of the Level of importance of contracting R&D with the number of IP mechanisms used construct (Table 4-8) suggests the presence of either a combination of trait and method effects or a mono-method bias since the correlation between the two items from the

questionnaire is stronger than the correlation obtained with their corresponding web-based measure. Thus, we cannot effectively discriminate the two web-based indicators from the questionnaire-based indicators, which implies that they cannot be part of the same construct.

The rejection of the discriminant validity between the web-mining method and the questionnaire-based survey method with the Proportion of employees assigned primarily in R&D with the number of patents (Table 4-13) stems from the fact that the correlation is higher with the IP web-based indicator than with the R&D web-based indicator. After all, it can be expected that a high proportion of employees dedicated at R&D activities should correlate strongly with the number of patents. Therefore, these indicators do not discriminate each other, and this construct is thus rejected.

Finally, the Proportion of employees assigned primarily in R&D with number of IP mechanisms used is passing the validity test since the four correlations are lower than the corresponding values found in the validity diagonal which is the accepting criteria even if $|r_{IP1,RD2}| = 0.288^{*}$ and $|r_{RD1,IP2}| = 0.284^{*}$ are reasonably high and significant. Although, it is determined to be a weak construct because of the slight differences in correlation.

While these results are encouraging, the presence of four valid constructs ironically shows that we cannot isolate precise actions with our web-based indicators. In other words, the “real” meaning of these indicators remains unknown and, as a result, proposition 1.1 is not confirmed. However, these measures may need to be put into a broader perspective to be effective, which may provide a better fit for the scope allowed by proposition 1.2. The next steps of this paper test whether the web-based indicators represent factors that are considered important for the firms.

4.5.3 Testing proposition 1.2

Construction of formative indices and new validation construct

Given the vast number of words used to construct the web-based indicators, as we concluded in the previous section, treating each questionnaire-based variable individually may not be appropriate. To illustrate the large lexical field of possible words related to the factors studied, and in order to properly test proposition 1.2, it is conceptually sound to build one single measure, one formative index with all the questions related to R&D and IP. Because the Principal Component Analysis (PCA) performed on all the items related to R&D and to IP presented at the end of Section

4.4.1 did not produce any significant KMO and Cronbach's Alpha measures, the use of a formative index comprising several sub-elements explaining our R&D factor in its broadest sense may be more appropriate.

Partial Least Square (PLS) regressions were estimated to determine whether it is possible to create valid formative indices for R&D and IP. In order to use PLS regressions, the methodology requires only the use of complete data sets (Nelson et al., 1996). Non-response is usually treated either by weight adjustment, that is, deleting incomplete data entry and weighing remaining respondents to compensate for the deletion, or by imputation, that is, adding artificial values based on average by classes and editing methods (Särndal et al., 1992) to replace the missing values (Haziza & Beaumont, 2007). Since our sample size for IP is already low for one of the items (39 for the number of patents), we could not afford to treat the missing data with a weight adjustment. Accordingly, we replaced the missing data with their imputation class based on control variables. We sorted the firms by sector, then by the number of employees and then by revenue. Depending on the situation, we used the mean of the class or the most conservative nearest neighbour, a method commonly used in the literature (Haziza & Beaumont, 2007; Little, 1986; Thomsen, 1973).

Since not all the items shared the same scale, we transformed each variable into a Z-score. PLS regressions were then estimated using WarpPLS 5.0 software with the following settings: MODEL B BASIC Warp3 Stable 3 and MODEL B BASIC Linear Stable 3. The two different settings produced the same conclusions. The details of the construct comparing the web mining technique and the questionnaire-based survey are shown in Table 4-15 (the PLS regressions, and the resulting weights, used to build these indices are provided in Appendix A in Table A-4).

All weights are significant (*p-value* < 0.01); indicator weight-loading signs are all positive; variance inflation factors (VIF) are all very low (< 1.5); and the Effect sizes (ES) are all greater than 0.02. All the criteria are met to indicate that the indices generated are valid (Cenfetelli & Bassellier, 2009; J. Cohen, 1988; Diamantopoulos, 1999; Diamantopoulos & Siguaw, 2006; Diamantopoulos & Winklhofer, 2001; Petter et al., 2007). For each factor, the sum of each weighted variable generated both indicators R&D_INDEX and IP_INDEX.

Table 4-15. The validation construct for proposition 1.2 testing

Concepts	Web Mining Variables	PLS-built indices	Questionnaire Variables
R&D	ln(WEB_R&D)	R&D_INDEX	Z_nb_R&D_proj Z_Int_R&D_source Z_Ext_R&D_source Z_Contract_In_R&D Z_Contract_Out_R&D Z_time_R&D Z_propEmp_R&D
IP	Inv(WEB_IP)	IP_INDEX	Z_nbIP Z_nbPatent

Note: All variables are continuous and normal.

Final MTMM Analysis

This final MTMM matrix includes the two web-based indicators along with the R&D_INDEX and IP_INDEX (see Table 4-16). Once again, the reliability diagonal will be neglected in our analysis since the measures are made with single items from the web method and with formative indices from our questionnaire. The monotrait-heteromethod diagonal shows high and significant correlations for R&D ($r = 0.419$; *p-value* < 0.01) and for IP ($r = 0.520$; *p-value* < 0.01), which hints at strong convergent validity. The heterotrait-monomethod value is low and not significant for the web content analysis method ($r = 0.182$; *p-value* > 0.05), although the questionnaire-based survey method value is high and significant ($r = 0.320$; *p-value* < 0.01). However, the monotrait-heteromethod values (r = 0.419 and r = 0.520) for R&D and IP respectively are much higher than the heterotrait-monomethod values (r = 0.320 and r = 0.182), which indicates no combination of trait and method effects and no mono-method biases. The first heterotrait-heteromethod value is low and not significant ($r = -0.017$; *p-value* > 0.05), while the other is moderate and significant ($r = 0.294$; *p-value* < 0.05). However, and more importantly, the correlations are much lower than the corresponding values found in the validity diagonal, which shows good discriminant validity. Since all the conditions are satisfied under the original guidelines proposed by D. T. Campbell & Fiske (1959), no risk of potential biases is induced within the methods, the traits or a combination of both. The results based on this methodology suggest that our web-based indicators reflect the

importance given to innovation factors such as R&D and IP. It is also worth mentioning the validity of this construct appears to be stronger than all the other attempted in Section 4.2.

Table 4-16. MTMM matrix for R&D_INDEX and IP_INDEX

		Web method		**Questionnaire method**	
	Traits	WEB_RD	WEB_IP	R&D_INDEX	IP_INDEX
Web method	WEB_RD	-[a]			
	WEB_IP	-0.182	-[a]		
Questionnaire method	R&D_INDEX	0.419**	0.294**	-[a]	
	IP_INDEX	-0.017	0.520**	0.320**	-[a]

Convergence validity: **Yes**, $r_{RD1,RD2} = 0.419^{**}$ and $r_{IP1,IP2} = 0.520^{**}$ are really high and significant;

Discriminant validity: **Yes**, $r_{RD1,RD2} = 0.419^{**}$ and $r_{IP1,IP2} = 0.520^{**} >> |r_{RD1,IP2}| = 0.017$ and $|r_{IP1,RD2}| = 0.294^{**}$;

Construct validity without either a combination of trait and method effects or a mono-method bias:

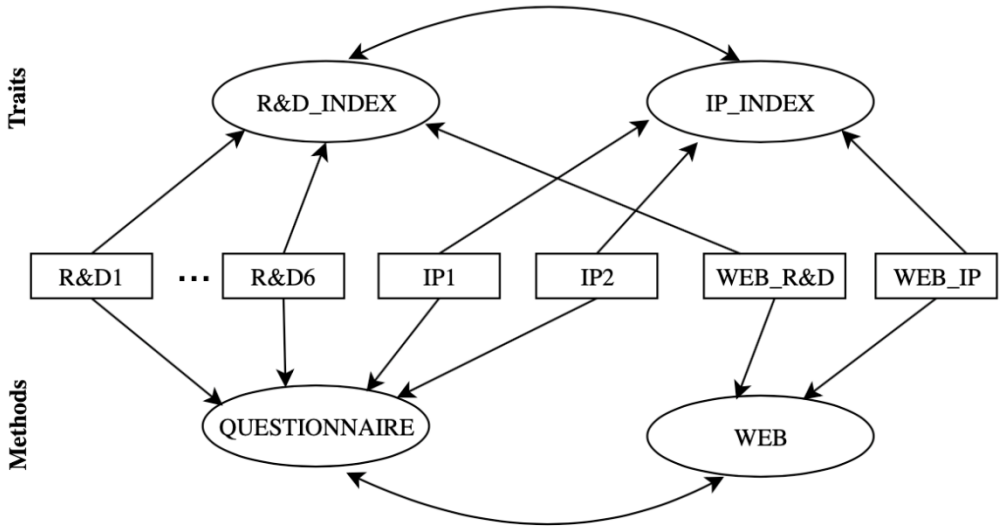
Yes, $r_{RD1,RD2} = 0.419^{**}$ and $r_{IP1,IP2} = 0.520^{**} >> |r_{IP1,RD1}| = 0.182$ and $|r_{RD2,IP2}| = 0.320^{**}$.

Notes: [a] All traits are measured by single items, no reliability statistic can be calculated.

* $p < 0.05$; ** $p < 0.01$

4.5.4 Post hoc analysis: MTMM confirmatory analysis

Some method effects can be induced independently "among the items within or between constructs" (Ortiz de Guinea et al., 2013, p.839) and "based on the nature of the rater, item, construct, and/or context" (Richardson et al., 2009, p.766). This could mean that one or more traits do not contribute or contribute negatively to the construct, which could ultimately affect the correlations found in the MTMM matrix performed in the previous step. Thus, if these method effects are significant, a congeneric bias is present. One way to verify the effect of each item individually is to extend the MTMM matrix with a confirmatory factor analysis (CFA), which allows the estimation of latent factors within the MTMM matrix (Maas et al., 2009). Using two PLS regressions allows traits and method factors to load on the items of the estimated constructs. Convergent validity is obtained when the trait loadings are higher than the method loadings. A possible method effect might thus be induced if a method load is higher than the related trait. The discriminant validity is obtained when correlations between the traits are low or moderate. The model tested for this analysis is presented in Figure 4-2.



RD: Questionnaire-based survey items measuring R&D; IP: Questionnaire-based survey items measuring IP; WEB_R&D: Web mining item measuring R&D; WEB_IP: Web mining items measuring IP;

Figure 4-2. MTMM CFA with formative indexes

MTMM CFA results are presented in Table 4-17. All the traits load on every item with strong values (all above 0.8). Most of the value loadings for the methods are strong but generally lower than the trait loading, indicating convergent validity and no congeneric bias. Possible method effects are observed with three items (R&D2, R&D3 and IP1) which load higher with the method than with the trait. It is possible that our use of formative indices may have had an influence on the measure, thus creating a method bias flagged by the CFA results. However, 7/10 (70%) of our results respect the condition, which is considered a high enough proportion to suggest an overall absence of method bias in the construct (Ortiz de Guinea et al., 2013).

The results from this post-hoc analysis confirm the results previously obtained using the MTMM matrix alone. Our web-based indicators seem to effectively reflect the importance technological firms grant to R&D and IP.

Table 4-17. MTMM CFA results

	Items	Trait Loading	Method Loading
Questionnaire-based survey items measuring R&D	R&D1 (Z_propEmp_R&D)	0.825	0.793
	R&D2 (Z_nb_R&D_proj)	0.921	0.984
	R&D3 (Z_time_R&D)	0.848	0.910
	R&D4 (Z_Ext_R&D_source)	1.000	0.984
	R&D5 (Z_Contract_Out_R&D)	0.983	0.469
	R&D6 (Z_Contract_In_R&D)	0.914	0.905
Questionnaire-based survey items measuring IP	IP1 (Z_nbIP)	0.900	0.993
	IP2 (Z_nbPatent)	1.000	0.978
Web mining item measuring R&D	WEB_R&D (Z_WEB_R&D)	0.996	0.915
Web mining item measuring IP	IP_WEB (Z_WEB_IP)	0.947	0.923

Finally, comparing both traits (as set out in Table 4-18) shows a moderate correlation between the two formative indices, which indicates a good discriminant validity.

Table 4-18. MTMM CFA correlations

	R&D_INDEX	IP_INDEX
R&D_INDEX	1.000	
IP_INDEX	0.273*	1.000

* $p < .05$. ** $p < .01$.

4.6 Discussion and implication for research

We used web content analysis to build innovation indicators from web text content of four factors that are consequential for the success of high-technology innovation, R&D, IP, collaboration and external financing. We then validated the true nature of these indicators, by determining whether they were valid substitutes for specific indicators that would otherwise have required a questionnaire-based survey to be obtained. To better understand the nature of the data extracted, two propositions were extensively tested.

Our first proposition stipulated that our web-based indicators could effectively measure specific actions of a company to perform activities related to a given factor. While significant results were obtained for IP factors and some indicators of R&D, there were no significant correlations with either collaboration or external financing factors.

For the specific case of R&D, we observed that the web-based indicator seems to reflect the promotion needs of various firms in terms of R&D. Our R&D web-based indicator is most highly correlated with the survey-based indicator related to firms more likely to provide R&D services to third parties. Thus, companies use their websites to promote their R&D service offerings. Furthermore, our web-based R&D indicator is highly correlated with the questionnaire-based indicator associated with firms that have a high percentage of employees allocated to R&D tasks. This can be explained by the willingness of firms to attract new R&D talent via their websites. Finally, our R&D web-based indicator correlated significantly with the questionnaire-based indicator tied to whether firms are more likely to contract R&D services from external providers. It is possible that the more important a firm considers R&D, the more likely it is to use both internal and external resources to achieve its R&D-related objectives, which can be described as OI practices (Chesbrough, 2003a). However, we did not find any correlation with an important R&D indicator, the number of R&D projects and our R&D web-based indicator, which at first may seem counter-intuitive, but reflects a limitation of the indicator by not properly accounting for the size dimension of R&D activities.

In contrast, our IP web-based indicator correlates with both the use of IP mechanisms and patent-related activities. Prior to the protection of an intellectual asset, that is, up to the R&D project level, an idea or future invention remains secret. Once an invention is revealed to the world via the disclosure that is generally associated with formal IP protection mechanisms, these IP protection

mechanisms act as a signal for the firm, showing how innovative it is or what it wants the world to see. Given that the firms' websites are used to publicly disclosing such information, it thus appears to be a promising source of data for IP protection-related data.

However promising, these preliminary results required more rigorous testing, beyond the simple correlation test between our indicators. The goal was to identify whether specific variables stand out from the others in terms of validity. To do so, we used Multitrait-Multimethod (MTMM) matrices to verify the validity of all the promising constructs to assess whether our questionnaire-based indicators can effectively discriminate each other when compared to the web-based indicators. Of the six MTMM matrices produced, four passed all the required validity tests, while two failed.

The fact that we obtained four valid constructs with four different combinations of R&D-IP pairs of different variables demonstrates that we cannot isolate specific actions from the questionnaire-based survey from web-based indicators at this point. Accordingly, at this stage we cannot use the web-based indicator built as a direct substitute for action-related indicators, since good convergent and discriminant validity were obtained for three different questionnaire-based indicators with the R&D web-based indicator and two different questionnaire-based indicators with the IP web-based indicator.

The keywords mentioned on websites may be used by companies that contract R&D services offered by third parties, by companies that provide R&D services to other companies, or by companies that have a higher proportion of employees allocated to R&D, or any combination of the three. We cannot separate one action from another.

The same applies to IP protection. Since acceptable results have been found for the number of patents a firm owns and for the total number of mechanisms used to protect IP, it is impossible to know precisely what is being measured by our web-based indicator. It is worth mentioning that neither factor significantly correlates with the other ($r = 0.144$; *p-value* > 0.1), which suggests that both variables represent different traits. Therefore, it is not possible at this point to isolate precise IP-related actions.

The second proposition suggests that our web-based indicators could measure the degree of importance that a company grants to a given factor. To do so, we combined all the elements related to a given factor to create two formative indices from the questionnaire-based indicators:

R&D_INDEX and IP_INDEX. We repeated the MTMM analysis using these two indicators and the two web-based indicators and observed the convergent and discriminant validity of our construct. A post-hoc analysis with a Confirmatory Factor Analysis confirmed our results. Therefore, our proposition 2 is accepted. In other words, we can use our web-based indicators as proxies for the relative importance a company grants to a given factor (such as R&D and IP). It is possible to conclude that the more R&D-related keywords a firm uses, the more important it considers R&D-related activities to be, irrespective of the nature of the activities. The same also applies to the protection of IP. We know that the more IP protection-related terms are used, the more actions are taken in this direction. However, this methodology does not allow us to precisely determine the nature of the actions a firm takes, which echoes previous findings by (Gök et al., 2015). We can only guess the nature of these actions as the indicator gives positive answers in both cases because the use of keywords alone ignores the context in which these words are used.

In a nutshell, our methodology can be used as a valid approach to provide data for future innovation and technology management studies for the relative importance given to factors such as R&D and IP, and to test the validity of the measures thus created. In most questionnaire-based surveys, this information is gathered using 1 to 7 Likert scale questions. If the goal of a study is to determine the degree of importance of core factors such as R&D or IP for a firm, the use of web-based indicators is reasonable. However, if the goal is to gather more precise information, such as the specific actions taken by a firm, these web-based indicators lack the necessary context to behave as expected. While they may provide complementary information, they cannot be used as direct proxies.

The fact that we did not obtain significant correlations with collaboration and external funding suggests erring on the side of caution before using this method on a larger scale. The reasons for the absence of convergent validity of the web-based indicators towards those generated from the questionnaire-based survey can be attributable to the small sample size, the inherent characteristics of the subjects from our sample, the questions used to capture that information within the questionnaire-based survey, and the keywords used to capture the information from the websites. Thus, other tests with different combinations of methods, keywords and traits need to be explored to determine whether it is possible to obtain valid and relevant information from companies' websites.

4.7 Limitations and future research

More data, our research would obviously be more robust, especially in terms of verifying the concept of collaboration and external financing normally addressed by classical methods, which can be appropriately measured using websites. For instance, we were unable to crawl data from all the companies from our survey due to technical limitations, which meant that only 79 out of 89 companies were used in this paper.

Websites are updated from time to time and the information provided changes accordingly, depending on what companies want to make public. It is thus important to note that a single web mining crawl might be insufficient to capture all relevant information given that results are subject to change as websites are updated. Thus, a longitudinal study would be required to more clearly understand how time can influence the traits to be measured, that is, the content available on a website and the nature of the website itself. An analysis of the evolution of web-based indicators over time would provide further validity to our methodology.

The inability to measure specific actions constitutes a limitation in our web mining methodology. The main problem lies in the lack of context tied to the use of keywords alone, possibly leading to multiple false positives. Machine learning and deep learning techniques, such as Recurrent Neural Network, Natural Language Processing or Bag-of-Words models, are promising avenues to explore the necessary context surrounding specific concepts to improve the level of precision of web-based indicators.

Furthermore, we began with theoretical factors for the conceptual framework, then identified the keywords related to these factors, and finally mined the website for these specific keywords. An interesting alternative would be to do this the other way around; in other words, to start with the website content and identify the factors that can be naturally found via unsupervised machine learning algorithms. The term frequency inverse document frequency technique (TF-IDF) could be used to provide insight into the importance of keywords relative to the rest of a document. N-gram low-frequency words clustering could also be tested in order to better isolate specific areas of interest (Price & Thelwall, 2005).

Company websites are purposely structured in a cooperative and agreeable manner for anyone seeking information about products, services, activities, and so on. The self-reporting bias induced by this methodology is inevitable. However, it is important to note that questionnaire-based

surveys and most national official public directories are all also subject to self-reporting biases. Fortunately, the bias induced by the web mining technique is as much a quality as it is a flaw in that it provides insight on how a company wishes to be perceived. In fact, companies post what they care about, what is important to them and who they are as an organisation on their websites. This qualitative information represents the essence of the company. Future research is needed to determine whether such information could, for instance, be used as a proxy to understand a company's culture.

Other qualitative data analyses of the websites' content could be used to reduce the risk of false positives and to gather more accurate data. The use of indicators based on websites' text data will open the door to experimenting with other possible indicators derived from the same websites, such as the use of colours, pictures and illustrations, audio and video content, choices of Web design styles, use of modern Web technologies, frequency of updates, possible interactions with visitors through all of the possible calls to action, and many other sources of data we may not have thought of as yet.

Finding a way to leverage the information from high-tech companies' websites will enable innovation management researchers to access to a whole new source of data that is free to use, accessible at all times and in large quantities, which will ultimately facilitate studies in this field. Finally, the innovation research community is invited to build on this new common ground to create a new systematic way to validate new constructs. In fine, if such new innovation indicators from web-based sources can be validated and clearly understood for what they truly measure, the burden on firms that are increasingly asked to answer questionnaires (from both researchers, industry associations or the government), would be considerably reduced. This paper shows that this is a promising avenue.

4.8 Author contributions

Mikaël Héroux-Vaillancourt: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing—original draft, Writing—review & editing. **Catherine Beaudry**: Conceptualization, Methodology, Project administration, Resources, Funding acquisition, Supervision, Validation, Visualization, Writing—original draft, Writing—review & editing. **Constant Rietsch**: Resources, Software, Visualization, Writing—original draft.

CHAPTER 5 ARTICLE 2: IT PAYS TO SIGNAL – A COMPARISON OF WEB-BASED AND SURVEY-BASED INNOVATION INDICATORS WITH FIRMS' INVESTMENT

Mikaël Héroux-Vaillancourt, Catherine Beaudry, Davide Pulizzotto and Margaret Dalziel

This chapter presents paper 2 submitted in Research Policy journal on March 28th 2023 which presents a research design that allows testing both propositions 2.1 and 2.2. We conducted a 3SLS regression to assess whether signalling is more associated with the evidence of innovation or with the investment received. The results show that web-based indicators are associated with the investment received, but not with the innovation evidence.

5.1 Abstract

This study explores the use of innovation-related language in corporate websites. We identified the online presence of 1,289 firms who participated in 21 questionnaire-based investigations between 2010 and 2016. We empirically investigated whether the presence of innovation-related terms in these websites better explains the evidence of innovation or the investment received by the firm. The content of the corresponding websites was extracted via snapshots hosted on The Wayback Machine and used to build web-based innovation indicators from an extensive document frequency analysis of the keywords related to various innovation factors (innovation culture, collaboration, open innovation, R&D and intellectual property). We then estimated a system of simultaneous equations to show that innovation-related signalling on corporate websites is significantly more related to the investment received than it is with evidence of innovation. The study contributes in understanding the intention behind the use of innovation-related signalling in the corporate world and proposes a new indicator to identify innovation-active firms that are seeking external support.

Keywords: Innovation; Investment; Signalling; Web mining; Document keyword frequency analysis; Simultaneous equations

JEL codes: O310; O320; L250; D820

5.2 Introduction

What are firms signalling on their websites and why? Whether a potential customer, an investor, a competitor, from the media, or someone seeking employment opportunities, the first thing one does to learn more about a firm is to visit its website (Wells et al., 2011). This makes the firm's website one of the most important public relations & communication vehicles it can leverage to scale its core messaging to attain strategic goals (Cornelissen, 2020). As such, a firm's signal retrieved from its website constitute a fascinating object of analysis which encapsulates a firm's assessment of its own market imperatives.

Corporate website signalling can provide valuable contextual insights for quantitative studies, particularly in innovation management. The lack of comprehensive contextual information has been noted as a limiting factor in assessing the success or failure of innovation practices, such as those within the open innovation framework (Bogers et al., 2017). While questionnaire-based surveys can provide precise contextual information, their increased usage can negatively impact response rates, particularly from busy business leaders. As most innovative firms have a website, it is important to take advantage of this abundant and free source of complementary data.

Some innovation studies attempted to assess how innovation-related signalling relates to objective innovation-related assessments, with mixed results (Gök et al., 2015; Héroux-Vaillancourt et al., 2020). The fact that there is no straightforward correspondence between what is being signalled on the websites and the actual objective measures collected by questionnaires raises important questions: What do these web-based indicators really capture? This study is interested in providing further insights into this question by assessing the triangular relationship between the evidence of innovation, the investment received and web-based innovation-related signalling. We focus on the investment received as it enables innovation activities which have consequences on innovation results. Furthermore, we assess how the aggregated innovation-related signals from corporate websites seem to have the strongest relationship with either 1) a direct measure of evidence of innovation or with 2) an indirect measure of evidence of innovation activities through the investment received.

We used the web archive "Wayback Machine" to retrieve the text content from websites associated with 2,379 respondents from 29 surveys conducted by The Evidence Network. After selecting relevant variables, we ended up with a final sample of 1,289 observations. Simultaneous regression

analyses were estimated with both innovation evidence and investment received by firms as a function of various web-based indicators built with the document frequencies of keywords related to innovation culture, collaboration, open innovation (OI), research & development (R&D), and intellectual property (IP). The results show that all web-based indicators have a positive and significant relationship with the investment received by the firms, but none with the evidence of innovation as measured by the questionnaire-based surveys.

This paper contributes to understanding the motivation behind the use of innovation-related language in firms as well as the message conveyed. Therefore, these web-based indicators may provide a unique and complementary measure for evidence of innovation activities, as well as marketing and branding activities that support innovation.

The remainder of the article is organised as follows: Section 5.3 presents the conceptual framework; Section 5.4 describes the methodologies used; Section 5.5 presents the descriptive statistics, Section 5.6 shows our empirical results; Section 5.7 presents a post-hoc analysis; Section 5.8 discusses the results; and lastly, Section 5.9 concludes.

5.3 Conceptual framework

This section establishes the foundation leading to our propositions. First, innovation measurement is defined in the context of this study. Second, factors known to influence innovation that are considered in our analysis are presented. Third, follows a background on web-based studies, to finally introduces the propositions.

5.3.1 Measuring innovation at the firm level

In the most recent edition of the Oslo Manual (OECD & Eurostat, 2018), innovation is distinguished from innovation activities, and both perspectives are therefore taken into consideration.

The concept of novelty is frequently invoked in the evaluation of the impact of innovative outcomes (Garcia & Calantone, 2002). In quantitative studies, the level of novelty is often operationalised by differentiating firms that introduce products or services that are new to the firm, new to the market or the country, and new-to-the-world (Augusto & Coelho, 2009; Brentani, 2001; OECD & Eurostat, 2018; OECD, 1997). The number of innovations is another commonly used indicator in quantitative studies (Flor & Oltra, 2004). To better discriminate firms based on the

radical nature of their products and services (see Amara et al., 2004, 2008), the innovation number may also include the "new-to-the-world" description (Augusto & Coelho, 2009; Lukas & Ferrell, 2000). Moreover, the impact of innovation results is frequently evaluated based on the sales of new products and services (Ebersberger & Herstad, 2012; Matthias Inauen & Andrea Schenker-Wicki, 2011; Szambelan et al., 2020).

Innovation activities constitute the process that enables the generation of innovation results. Thus, the time required between ideation and market launch serves as a measure for the productivity of the innovation process (Ivanov & Avasilcăi, 2014). In general, a shorter time to market allows a company to introduce new products or services quickly, potentially leading to a competitive advantage over rivals (Lieberman & Montgomery, 1988). Additionally, firms operating in the same industry with similar product types but with a faster time to market generally benefit from better processes and greater process innovation performance (Pisano, 1997).

5.3.2 Factors that contribute to innovation

Several factors have been shown to be associated with innovation in this literature. The paragraphs below are by no means intended to provide an exhaustive review but to highlight a few of the usual factors.

The first factor of interest, R&D, and in particular R&D intensity, demonstrates how much a firm invests in its own capability to create inventions (Griliches, 1994, 1998) along with its own absorptive capacity (W. M. Cohen & Levinthal, 1990). The relationship between R&D intensity and innovation performance was shown to be positive in several quantitative studies (Deeds, 2001; Greve, 2003). Due to their risky nature, R&D activities have difficulty to properly meet their financing requirements (K. D. Miller & Bromiley, 1990). The connection between financial considerations and investment choices in R&D has also been extensively studied (Coldbeck & Ozkan, 2018; Fazzari et al., 1988)

External investment in R&D favours innovation performance while financial support during the commercialisation phase allows companies to scale their operations and support their growth (Kor & Mahoney, 2005; Santarelli, 1991). The relationship between innovation and investment is well established in the literature (Cornaggia et al., 2015; Hsu et al., 2014; Kerr & Nanda, 2015; Wellalage & Fernandez, 2019). With publicly traded firms, Aghion et al. (2004) showed that the relationship is also the other way around, the more a firm reports R&D intensity, the more it

receives investment. This relationship between successfully obtaining external financing and product and process innovation is more important for younger small and medium-sized enterprises (SMEs) than for more mature firms (Wellalage & Fernandez, 2019). However, firms pursuing radical innovation strategies experience greater uncertainty, leading to a lower likelihood of securing external financing at founding compared to firms using an incremental strategy, which benefit from established market niches and clearer benchmarks for success (Burton et al., 2002). Public support in the form of grants, subsidies and loans encourages firms to innovate more (Beugelsdijk & Cornet, 2002; Keizer et al., 2002). Furthermore, firms that successfully obtain public funding are also more likely to attract private funding (Meuleman & De Maeseneire, 2012). When it comes to private equity, venture capital investment in young US high-tech firms had a positive and significant impact on their R&D productivity (Brown et al., 2009). Finally, young firms operating in cutting edge immature industries rely on external funding to reach the commercialisation and marketing phases (Kalil, 2005; McNeil et al., 2007).

The second factor of note is based on the fact that firms do not innovate in isolation, but cooperate at various stages of the innovation process. Collaboration has a positive impact on many aspects of the innovation process, from idea generation and research to commercialisation. It can improve a company's competitiveness (Hagedoorn et al., 2000; Roja & Nastase, 2013), notably by increasing the results from innovation, such as the number of patents, sales growth and the return on innovation sales (Belderbos et al., 2004; Carboni, 2013). Collaboration is also essential for the development and deployment of complex emerging technologies, and may contribute to reducing research and development costs, sharing risk, and improving performance (Johnson & Filippini, 2009; Jordan & Lowe, 2004).

Somewhat related to the second factor are "open innovation" (OI) practices. First coined by Chesbrough (2003a), OI aims to improve the innovation system performance by actively performing two key processes: acquiring (inbound) knowledge or technologies from other organisations and distributing (outbound) them or their results to other organisations (Chesbrough, 2003a, 2003b). Multiple quantitative studies showed that acquiring external knowledge or technologies benefits the internal R&D process and affects positively the innovation performance of firms (W. M. Cohen & Levinthal, 1990; Henkel, 2006; Laursen & Salter, 2006b). As rightfully highlighted by scholars, successful companies have been engaging in similar collaborative and knowledge sharing and exchange practices for many years before the term "open innovation" even

existed (Trott and Hartmann, 2009). Other scholars have found decreasing returns to the use of OI practices (Fey and Birkinshaw, 2005; Laursen and Salter, 2004). Interestingly, particularly for this study, OI is often perceived as a buzzword (Hagel & Brown, 2008), and firms will likely display OI initiatives as a way to enhance their brand value (Siqueira et al., 2021).

Before signing collaborative agreements, or opening up its innovation process, firms have a tendency to protect their intellectual property (IP) (W. M. Cohen et al., 2000b). By doing so, companies gain a competitive edge by securing exclusive rights to commercially exploit their technologies, branding, copyright, industrial designs, and so on (Teece, 1986). Patents help companies increase the return on investment for the funds invested in technology development by giving them temporary monopoly power, allowing them to license their technology or sell their patents (Chesbrough, 2003b; Rivette & Kline, 2000).

The literature also suggests a very strong connection between innovativeness and culture (Dobni, 2008). The first evidence came from observing that different orientations allow firms to differentiate themselves in terms of their innovation philosophy, and thus their resulting performance (Aldas-Manzano et al., 2005; O'Cass & Viet Ngo, 2007). These orientations are the result of beliefs, attitudes and actions, key aspects of an organisational culture (Schein, 2004). A culture of innovation has three key traits: shared values that support innovation, established norms for innovation, and visible practices that encourage risk-taking, openness to new ideas, learning, and constructive dissent within a group (Herzog, 2011).

5.3.3 Web signalling, investment and innovation evidence

The previous sections highlighted that firms do not innovate in isolation, need external financing to fund their R&D activities and increasingly rely on finding external path to markets for some of their innovation outcomes. How do they manage that in practice? Firms are agents with their own agenda and objectives within an ecosystem of stakeholders. If they do not already rely on an extensive network of collaborators and funders, or want to expand this network, their corporate website serves as a means to elicit some level of desired reactions from actual and future stakeholders. Sending signals, or signalling, is therefore at the core of a firm's communications.

Corporations have very strong incentives to maintaining an excellent reputation and image, through the careful coherence between a firm's own corporate identity and corporate communications (Chun et al., 2005; Gammoh et al., 2006; E. R. Gray & Balmer, 1998). Disclosure

of sensitive information may be especially costly to reveal, and it is often best for a firm to only disclose information that improves its value (Verrecchia, 1983). This suggests that firms are more likely to disclose information on their website that portrays them in a positive light, rather than revealing true aspects of their operations. As biased as this self-reporting may appear, corporate websites are perceived as centralised signalling platforms where one can decode the firm's intentions and priorities in terms of communication with its key stakeholders. These websites reveal important information, which may be useful for researchers to gather.

Innovation studies have primarily used web content analyses to complement quantitative analyses, providing contextual information such as temporal information, specific product or technology-related details, or triple helix ties (S. K. Arora et al., 2016, 2020; Li et al., 2018; Youtie et al., 2012). However, few studies have attempted to directly assess firms' innovation based on their websites. Some web content analyses that used keyword frequency analysis to assess innovation activities have shown that web-based indicators do not significantly correlate with direct questionnaire-based measures, such as R&D intensity, suggesting that the indicators do not capture the same concept (Gök et al., 2015). Similarly, Héroux-Vaillancourt et al. (2020) arrived at a similar conclusion by demonstrating that web-based variables strongly correlated with various R&D metrics but were unable to effectively differentiate between them. The authors also reported strong discriminant relationships between web-based indicators and sophisticated indices reflecting broader concepts for R&D and IP protection from questionnaire data. However, the true meaning of these broader concepts captured through keyword frequency remains an open question.

The findings from both Gök et al. (2015) and Héroux-Vaillancourt et al. (2020) suggest that web-based innovation indicators built with a keyword method from corporate websites do not provide precise objective measurements. This can be illustrated with an example: a small biotechnology enterprise working on a single product and a large firm with a vast product portfolio may use similar innovation-related language on their websites, despite having vastly different R&D expenditures per unit of income. Innovation signalling may represent a self-reported level of evidence of innovation, similar to how smoke in the sky indicates that a fire is nearby. The smoke does not help to specifically derive the different measurable properties of the fire, such as its temperature, brightness, and combustion rate. Yet, we still know that there is not smoke without fire. Similarly, while web-based innovation-related languages may not correspond directly to specific innovation metrics, we may interpret the terms used as evidence of innovation. This is a

similar epistemic approach as archaeologists that shed light on palaeolithic human societies from cave painting observations in a meaning-making process called evidential paradigm (Bertozzi, 2007; Ginzburg, 1989).

The goal of web-design teams is to showcase the innovative attributes of their firms to their key stakeholders. With that in mind, for this research we need to adopt a much broader definition of innovation evidence that encompasses both process and outcome perspectives of innovation for its measurement. For instance, any direct or indirect already proven innovation indicator may be seen as a potential source of evidence of innovation.

Similarly, a firm may want to convey its innovativeness and innovation culture on its website by hinting towards its use of various concepts related to innovation, and the factors mentioned above: R&D activities, collaboration and OI, how it protects IP, etc. These innovation-related signals are sent with key stakeholders in mind. For example, a firm may benefit from signalling its innovation culture to future and current employees, investors, and clients. On its website, a firm may signal innovation as a way to display how much it supports innovation and considers it to be important. Terms related to creativity and idea generation may be visible on a website to attract new employees or clients. Firms may want to be a beacon for collaboration if they seek and value new partners – the signal would then be intended for providers, future and current employees, investors, academics, and even competitors. To warn off the latter and potential free riders that would consider copying their invention, firms may also highlight that they formally protect their IP.

In addition, firms have a strong incentive to signal their use of formal IP protection and thus their will to appropriate the results of innovation outcomes to potential investors. The literature has indeed found a positive correlation between citation-weighted patents and the stock market valuation of firms (Hall et al., 2000). Both internal and external financing are considered essential for R&D activities that lead to innovation, and are often a costly and long-term investment (Aghion et al., 2004). Just the word “patent” may be perceived as a costly signal (McAndrew, 2018), which provides support for a firm’s ability for innovation. In addition, firms may signal R&D as a way to attract investment, high qualified personal, and high-quality collaborators.

To help better understand what innovation-related signalling may refer to, we explore two possible interpretations. Established innovative firms are defined as firms that reported innovation results (OECD & Eurostat, 2018). To determine whether the web-based innovation-related signalling is

generalisable as a valid unbiased metric of real evidence of innovation activities, it should be able to explain an objective source of evidence of innovation such as classic innovation result indicators measured by questionnaire-based survey data. It is therefore expected that stronger web-based innovation signals are associated with stronger evidence of innovation. This leads to the following proposition:

Proposition 2.1*: A stronger innovation signal on a firm's website is positively associated with greater evidence of innovation.*

Alternatively, innovation-active firms may not yet report innovation results while still be engaging in innovation activities (OECD & Eurostat, 2018). The use of innovation signalling is anchored within a strategy which aims to support corporate objectives. One of these objectives may be related to facilitate the firm's capacity to attract investment. Indeed, as the lack of investment has long been identified as a significant obstacle to innovation (Harhoff & Körting, 1998), it is more than likely that web-based innovation-related signalling may be related to attracting investment or other form of external financing. It is also true that investment enables a firm's capabilities to innovate, and therefore have an impact on what the firm may possibly be reporting. Either way, the stronger the innovation-related signalling is, the more it is expected to be related with investment received. Furthermore, if a firm's website usage of innovation-related language is related with the investment received, then the signal provides indirect evidence for innovation activities. Therefore, we will test the following proposition:

Proposition 2.2*: A stronger innovation signal on a firm's website is positively associated with the firm's receiving more investment.*

These two propositions require the collection of both web-based data (the signal) and more traditional questionnaire-based survey data (the "real" measure) to validate what signals the firms aim to convey, to whom and for what purpose.

5.4 Methodology

5.4.1 Data

This section describes two data gathering methodologies. First, we present the survey data used for the purpose of validating the signal measured in corporate websites. Second, we explain our web-based complementary and text content data collection processes.

The Evidence Network (TEN) surveyed business entities that were served by innovation intermediaries in order to access the impact of intermediary services. TEN provided this study with a sample of data collected between 2010 and 2016, from 29 investigations (also named programs) with 4,740 entries from Canadian and Finnish organisations, including businesses, non-profit organisations and public institutions.

The data stem from the answers to 13 questionnaires which were tailored-made and contextualised to the target respondents in each investigation by TEN representatives. The questionnaires were administered online and designed to be answered in 15 minutes on average. They contain about 30 questions generally divided into four sections:1) preliminary questions on the activities of the respondents in relation to the program, such as the year of the first engagement, the assessment of the importance of the services offered, other specific questions related to a specific program such as specific financial programs and specific project types and so on; 2) impact of the intermediary on the company's resources and capabilities; 3) impact of the intermediary on market performance, and 4) concluding questions on firm characteristics, such as the number of employees, revenue, foundation year and so on. Each investigation was conducted within a month. The close relationship between the companies and the intermediaries combined with follow up thorough calls and email reminders allowed for an average response rate of 39% across all investigations in our sample[23].

The items relevant to this study include foundation year, revenue, sectors, the number of innovations introduced in the past 3 years, the percentage of the company's revenues that can be attributed to innovations, the degree of novelty of innovation, the time to market, and the investment received. Examples of the questions can be seen in Appendix B.

Although the questionnaires are covering the same topics, because they were purposefully designed to suit the needs of the various intermediaries, there are notable differences in their elaboration. Consequently, combining data from various investigations necessitated a meticulous

[23] See Table D-1 in Appendix D.

harmonisation process to ensure that the variables employed in the study were transformed to a common scale[24], along with some degree of imputation to ensure a slightly larger sample size[25].

Furthermore, additional data was needed such as the corporate website address (URL) and the type of organisation. Each entry was individually processed to gather these missing data for our sample. Of the 4,740 observations, we retrieved 2,949 websites, resulting in the exclusion of 1,791 observations from this study.

The various investigations surveyed all types of business entities that interact with the intermediaries. Thus, we identified six mutually exclusive categories of organisations: tech firms, low-tech firms, advanced knowledge service firms, other service firms, non-profit/non-governmental organisations (NGO), public institutions, and undefined. Since this paper is only interested in firms, the responses from non-profit organisations (64), public institutions (75), and from undefined respondents (30) were removed from the sample. The resulting sample includes 2,780 respondents (58.6%) corresponding to 2,483 unique companies with a working website domain.

To obtain the most relevant website data, we minimised the time gap between questionnaire administration and web content retrieval by using the "Wayback Machine" web archive, which stores captured web pages over the past 26 years. Although the service has a limited coverage, previous research has validated its use as a data source for social sciences (S. K. Arora et al., 2016; Murphy et al., 2007).

To constitute our corpus, we extracted the text content of the first two levels of the website, which included the homepage and all pages directly accessible from it. We selected the closest snapshot of a firm's homepage from June 30th of a questionnaire year, and then a breadth-first search

[24] All the items except the sector and the degree of novelty innovation were combined either by taking the middle point in the case of an interval, otherwise by the direct value inputted by the respondent, either the number directly or the number corresponding to a multiple-choice value. The harmonization of the different sectors needed to recategorized some sectors. In the end, we have 6 sectors: Energy & Environment, Digital media & ICT, Life science and health, Materials & Manufacturing, Natural resources and Other. The degree of novelty is presented in the variable section.

[25] Imputation was performed on revenue (4.84%), number of innovations (7.7%), percentage of the company's revenues that can be attributed to innovations (13.8%), time to market (22.6%), and investment received (13.1%) using the nearest neighbor methodology, thus respecting the threshold recommended (Särdnal et al., 1992).

algorithm was applied to crawl the URLs enabling our data collection from the second layer of the website[26].

To detect languages and translate all content from other languages into English, we used LibreTranslate[27] from Argos translate[28]. This is an automatic translation software that is based on OpenNMT[29], an open source neural engine system used for translation (Klein et al., 2017, 2018, 2020). After cleaning, the English textual content accounted for 86% of the initial sample. Following the recommendation of Lachmann et al. (2001), only sentences were used as signalling units for analysis, resulting in 511,815 sentences. Among these, 142,467 sentences contained keywords related to innovation. The detail of the process is shown in Figure 5-1.

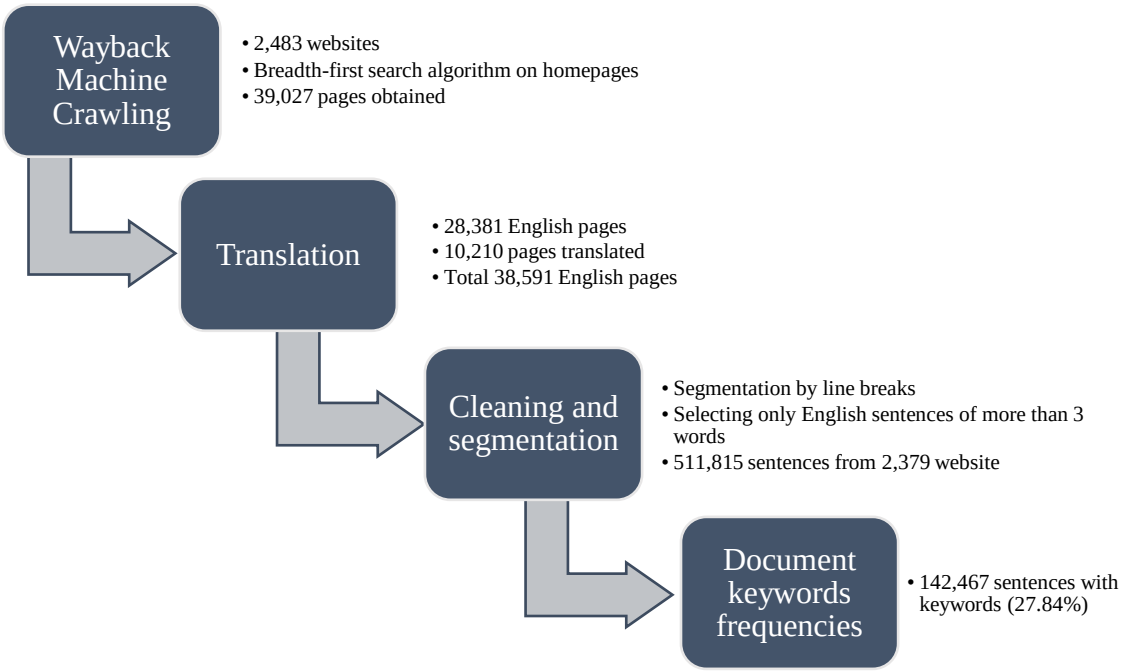


Figure 5-1. Web text content acquisition process

[26] GitHub. (n.d.). Wayback Machine Crawler. Retrieved March 10, 2023, from https://github.com/4point0-ChairInnovation-Polymtl/Wayback_Machine_Crawler
[27] LibreTranslate. (n.d.). Retrieved March 10, 2023, from https://libretranslate.com
[28] Argos Open Tech. (n.d.). Retrieved March 10, 2023, from https://www.argosopentech.com
[29] OpenNMT. (n.d.). Retrieved March 10, 2023, from https://opennmt.net

5.4.2 Variable construction

Dependent variables

The two propositions presented at the end of the conceptual framework necessitate two dependent variables. The first pertains to the level of evidence of innovation, while the second is related to the investment obtained.

We retrieved various indications that a firm innovates, which show evidence of innovation from the best objective knowledge at our disposal i.e., various innovation measures obtained from the different surveys. Then, using an inductive approach, specific measurements are combined to derive to the broader concept of "evidence of innovation" (*InnovEvidence*). This broader concept is built from three classic innovation measures: the degree of innovation novelty, the number of innovations and the percentage of revenue generated from the latest innovation (OECD & Eurostat, 2018). These specific indicators point to different kinds of innovation results. When transformed to be on the same scale, their combination provides an objective measure of the broader concept of "evidence of innovation".

An index to assess the "intensity" of evidence of innovation can be obtained by standardising the measurement scale of these three indicators. First, the degree of novelty of innovation in terms of international leadership was measured using 5-point and 7-point ordinal scales. To harmonise these two scales, a new 3-point ordinal variable (*DegNoveltyInnov)*[30] was created. Second, some of the surveys asked the respondents for the number of innovations (*nbInnov*) as simple counts of the number of product or service innovations introduced by the firm in the three years prior to the survey. Third, the variable *PropRevInnov* represents the proportion of a company's revenues that can be attributed to product, process, and service innovations.

To combine these three variables, we divided them all by the maximum, effectively standardising the measurement scale over the range of 0 to 1. *InnovEvidence* is then obtained from the corresponding value of the three transformed variables *DegNoveltyInnov*, *nbInnov*, and *PropRevInnov*. The resulting continuous variable is a measure of the level of evidence of innovation, where 0 represents the absence of innovation evidence and 1 is the strongest innovation

[30] *DegNoveltyInnov* is 0 for no innovation developed, 1 for national/regional leader, and 2 for international leader.

evidence. The second dependent variable is a direct measure of the amount of funding or investments received by the firm[31] (*Investment*).

Independent variables

Section 5.3.2 highlighted five well-established factors directly related to firms' innovation: innovation culture, collaboration, OI, R&D and IP protection. Associated with these factors are the signals to key stakeholders through the use of innovation-related language on firms' websites. The paragraphs below explain how the text content of corporate websites was transformed into measurable signals to compare with the two dependent variables described above.

Corporate websites may use a variety of synonyms and connected terms to signal innovation-related activities and culture. We postulate that the more a firm uses terms associated with innovation, the more it is signalling innovation. The web-based methodology used in this study consists of web content analysis based on document keyword frequency.

The lexical field related to the five factors was examined from peer-reviewed publications and important concepts from seminal books such as the Oslo Manual in the case of innovation (OECD & Eurostat, 2018) and the Frascati Manual (OECD, 2015) for R&D. Keywords related to R&D (Gök et al., 2015; Héroux-Vaillancourt et al., 2020; Meyer, 2000; OECD, 2015; Schmoch, 2007), Collaboration (Arvanitis, 2012; Belderbos et al., 2004; Carboni, 2013; Ramdani, 2014) Innovation culture (Herzog, 2011; OECD & Eurostat, 2018), Open innovation (Bogers et al., 2017; Chesbrough, 2003a, 2006b; Liyanage et al., 1999) and intellectual property protection (Rietsch, 2015; Teece, 1986) were inspired from the literature. Multiple tests were performed to arrive to a final list of keywords: using the original list; adding and removing keywords that were either consistently not used by firms (e.g. "risk taking", "market introduction") or more likely to be leading to false positives (e.g. "experience", "test") or that would give too much exposure to terms frequently repeated within the legal pages, such as the terms of services or privacy policy or cookie policy (e.g. "new product", "new service", "intellectual property"). The textual content was processed using stemmed versions of the keywords to capture most of the declination of a given concept (e.g., using "innov" instead of "innovation", "innovative", "innovation culture",

[31] Half of the external financing is equity financing only, while the other half includes equity, grant, and loan.

"innovator", "innovate", and so on). The keyword construct is described in Table C-1 in Appendix C.

In the natural language processing (NLP) terminology, document frequency is the number of documents containing a particular term. In this study, each sentence is considered as a document. To reduce the impact of the variance attributable to differences in website structures and sizes or to the exposure to the Wayback Machine crawler, each frequency was standardised by dividing it by the total number of sentences.

The standardised document keywords frequencies[32] appear to follow a Pareto distribution where the top seven (19.44%) out of 36 keywords represent 80.6% of the cumulated frequencies. Most of these terms are related to either R&D (67%), Innovation culture (18%) or Collaboration (10%), while the keywords with the lowest frequencies are related to OI (3%) and IP (2%). Within each concept, the top keywords yield the majority of the relative document frequencies. For Innovation culture, 69% of the cumulated frequencies stem from two keywords: "innov" (39%) and "create" (30%). For Collaboration, "partner" (68%) and "collab" (20%) yield 87% of the cumulated frequencies. As for OI, "community" alone is gathering 76% of all the frequencies, followed by "startup" (12%), which leaves the remaining 13% to the other 14 keywords. The first three keywords for the R&D concept represent 82.03% of the frequencies: "tech" (34%), "develop" (30%) and "lab" (18.5%). The very low presence of "r&d" in corporate websites came at a surprise with only 845 documents mentioning the exact term, representing less than 1% of the occurrences. Not surprisingly, the IP concept was overwhelmingly dominated by the keyword "patent" (81%).

Each keyword has its own meaning, is used for a specific context, and is thus a different element of a larger signal. Conceptually, these indicators are formative indices, i.e., a signal that is made up of the sum of its parts that are all independent of each other (Diamantopoulos & Siguaw, 2006)[33]. Therefore, the sum of the different standardised frequencies of all the keywords will be interpreted as the signal for a given factor as it was done in previous research (Gök et al., 2015; Héroux-Vaillancourt et al., 2020). A web-based index (*webIndex*) is built with the sum of the standardised frequencies from the keywords of all factors taken as a whole. The same process, simply adding the standardised frequencies from the factor's corresponding keywords, was

[32] See Table C-3 and Table C-4 in Appendix C.

[33] We conducted tests with reducing techniques and found no empirical evidence supporting the existence of any underlying reflective latent construct.

repeated for each factor and yielded five more continuous variables: *webInnov*, *webCollab*, *webOpen*, *webRD* and *webIP*. The construct is also empirically supported with Partial Least Squares regression. These results are presented in Appendix D.

The time to market (*TimetoMarket*) corresponds to the time in months between conceptualisation and the market launch. It is used as a productivity measure of the innovation process that also explains evidence of more innovation results.

Technological companies require more financial resources in general to engage in their innovation activities (Kor & Mahoney, 2005). As a result, they also need external financing (Kalil, 2005; McNeil et al., 2007). Furthermore, they are the ones that attract more private equity investments (Brown et al., 2009). Thus, they are expected to contribute to explaining the investment received. Therefore, a binary variable identifies the tech firms (*TechFirm*).

In addition to the twelve web-based variables presented above, a number of control variables were added. Other firm characteristics such as its location, age, size and sector are known to influence innovation (Becheikh et al., 2006). A dummy variable identifies young firms (between 4 and 8 years old) at the time of the questionnaire (*dYoung*)[34]. Three dummy variables account for the year of the investigation (*d2011, d2013,* and *d2015*)[35]. The annual revenue reported by the firm at the time of the investigation (*Revenue*) provides some indication of its size[36]. In addition, three sectoral dummy variables are added for each of the following industries: Digital Media & ICT (*dDigMediaICT*), Life science health (*dLifeSciHealth*), Materials & Manufacturing (*dMaterialMnfg*)[37]. Finally, as our sample comprises Canadian and Finish firms, a dummy variable that takes the value 1 if the firm is located in Finland (*dFinland*) was added. A description of all the variables used is presented in Appendix D.

5.4.3 Regression model

To test propositions 2.1 and 2.2, a system of equations is proposed to model the triangular relationship between the innovation evidence, the investment received and the web-based

[34] The other age groups constitute the omitted group (*dEarly* and *dMature*).
[35] Other investigation years (d2010, d2012, d2014 and d2016) were also considered, but were omitted because of either too small representation or too high correlations with other independent variables.
[36] The number of employees is associated significantly with other independent variables.
[37] The remaining sectors constitute the omitted group as they are less expected to be associated with both innovation and investment.

innovation signal. Web-based innovation signalling may have an incidence on both the evidence of innovation outcomes (proposition 2.1) and the investment received (proposition 2.2).

The first equation (5.1) models the influence of the web-based innovation signal, as well as the time to market and the investment received as factors contributing to the evidence of innovation. The second equation (5.2) models the effect of the web-based innovation signal and whether firm is technology-based entity on the investment received[38].

As the investment is both a factor explaining innovation evidence and a dependable variable, it is a system of simultaneous equations. This system of equation allows to assess how much the web-based innovation signalling explains the variation in both the evidence of innovation and the investment received at the same time, and thus to test whether the web-based innovation signalling has a greater impact on explaining the variation in either one of the dependent variables.

$$InnovEvidence = \beta_0 + \beta_1 Investment + \beta_2 webSignal + \beta_3 TimetoMarket + \sum_{i=1}^{n} \beta_i Control_i + \varepsilon_1 \quad (5.1)$$

$$Investment = \gamma_0 + \gamma_1 TechFirm + \gamma_2 webSignal + \sum_{i=1}^{n} \beta_i Control_i + \varepsilon_2 \quad (5.2)$$

where $control_i$ include the age of the firm (*dYoung*), Investigation year (*d2011*, *d2013*, *d2015*), *Revenue*, sectors (*dDigMediaICT*, *dLifeSciHealth*, *dMaterialMnfg*) and location (*dFinland*).

Three-stage least squares (3SLS) regressions were estimated using the reg3[39] procedure in Stata to produce consistent estimates from the simultaneous equation models.

5.5 Descriptive statistics

The intersection between the observations from the questionnaires with responses for the variables of interest and control and the web-based sample yields a final sample of 1,289 observations, 1,209 of which are unique firms, and 69 repeated observations (5.35%), including one firm with four

[38] Innovation results usually contribute to the investment received as well. However, since the sample represents predominantly small firms served by intermediaries to improve their innovation outcomes, it is considered more likely that these firms seek to obtaining investment which would lead to more innovation results.

[39] The reg3 procedure in Stata 16 uses an instrumental-variables approach and generalized least squares (GLS) to account for the correlation structure in the errors across the equations (StataCorp., 2019)

observations, nine firms with three observations, and 59 firms with two observations. The repeated measures stem mainly from responses from different programs in different years.

The 1,289 respondents represent a wide range of Canadian (53.5%) and Finnish (46.5%) companies of different sizes (63.3% employ fewer than 10 employees, 30.8%, have between 11 and 99 employees, 3.9% employ between 100 and 499 employees, and 2% have more than 500 employees). Their activities are in multiple sectors such as Energy & Environment (4.1%), ICT (29.3%), Health and Biotechnology (12.3%), Materials & Manufacturing (15.8%), Natural resources (3.7%) and other sectors (34.9%). Different organisation types are also included in the sample: Tech companies (61.8%), Low-tech companies (27.1%), advanced knowledge services (10.1%) and other services (2%). The data was also gathered in different years: 2010 (2%), 2011 (8%), 2012 (18%), 2013 (14%), 2014 (38%), 2015 (16%) and 2016 (4%). All descriptive statistics and correlation results are presented in Appendix D.

5.6 Results

5.6.1 3SLS regression results

Table 5-1 presents the 3SLS regression results. Each indicator enters the model separately and sequentially which results in six regressions. A final regression combines all distinct concepts to assess their relative contribution. The models can be interpreted as follows: Model 1 examines the impact of the web-based innovation evidence, then Models 2-6 evaluate the impact of web-based evidence of each concept separately, and finally, Model 7 shows the combined effects of all the web-based evidence concepts together. Robustness check is available in Appendix D.

The results reveal a consistent story across the board. Nearly all web-based indicators yield positive and significant results in explaining the investment received (eq. 5.2), yet they do not demonstrate significant positive impacts on the evidence of innovation (eq. 5.1). Only Models 5 and 7 revealed significant results between web-based indicators and *InnovEvidence*, with Model 5 showing a surprising negative impact of the web-based R&D indicator and Model 7 showing a significant positive coefficient of the web-based innovation culture indicator and a significant, but still negative impact of the web-based R&D signalling.

With the investment received, the web-based innovation signalling in Model 1´ is positive and significant, and so are all the individual web-based indicators (Models 2´ to 6´). When all web-

based indicators are combined, the innovation culture indicator (*webInnov*) is no longer significant, but all the remaining web-based indicators remain positive and significant, demonstrating the robustness of the results. These results seem to support proposition 2.2, indicating that these web-based indicators are more interpretable as signals related to investment rather than direct innovation evidence.

Next, as predicted with the correlation results, the investment received does not have an impact on the evidence of innovation. This is unexpected, considering the established relationship between innovation and investment from the literature. There is also a significant negative relationship between *InnovEvidence* and the time to market which means the most productive firms at innovating seem to show more evidence of innovation results. Additionally, the results suggest that being a technological firm does indeed have a significant positive impact on the *Investment* received.

Finally, with *InnovEvidence,* there is a significant positive impact from firms that answered a questionnaire in 2011, firms from the digital media & ICT and materials & manufacturing sectors while having significant negative impact for firms located in Finland and firms that answered to a 2015 questionnaire. For *Investment*, there is a positive and significant impact of being a young firm at the moment of the questionnaire, firms from the digital media & ICT and life science and health sectors, and from firms located in Finland while seeing significant negative relationship from firms that answered in 2011 and 2013.

Table 5-1. 3SLS regression results

Variables	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1. InnovEvidence							
Investment	0.038	0.024	0.030	0.027	0.044	0.030	0.043
	(0.043)	(0.038)	(0.038)	(0.036)	(0.044)	(0.041)	(0.047)
TimetoMarket	-0.213***	-0.214***	-0.215***	-0.217***	-0.208***	-0.216***	-0.203***
	(0.073)	(0.073)	(0.073)	(0.073)	(0.072)	(0.073)	(0.073)
TechFirm							
dYoung	-0.016	-0.014	-0.014	-0.014	-0.017	-0.015	-0.017
	(0.019)	(0.018)	(0.018)	(0.018)	(0.019)	(0.019)	(0.019)
d2011	0.211***	0.208***	0.208***	0.207***	0.212***	0.208***	0.209***
	(0.038)	(0.037)	(0.037)	(0.037)	(0.038)	(0.037)	(0.037)
d2013	0.043	0.045	0.043	0.044	0.044	0.044	0.044
	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)
d2015	-0.051*	-0.046*	-0.049*	-0.047*	-0.053*	-0.048*	-0.054*
	(0.033)	(0.032)	(0.033)	(0.032)	(0.033)	(0.033)	(0.034)
ln(*Revenue*)	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
dDigMediaICT	0.116***	0.114***	0.117***	0.117***	0.114***	0.115***	0.112***
	(0.023)	(0.023)	(0.023)	(0.023)	(0.024)	(0.024)	(0.024)
dLifeSciHealth	0.045	0.054	0.051	0.051	0.042	0.050	0.044
	(0.058)	(0.056)	(0.055)	(0.055)	(0.058)	(0.057)	(0.060)
dMaterialMnfg	0.041*	0.039*	0.040*	0.039*	0.041*	0.04*	0.040*
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
dFinland	-0.232***	-0.218***	-0.225***	-0.226***	-0.235***	-0.225***	-0.236***
	(0.029)	(0.026)	(0.026)	(0.027)	(0.029)	(0.028)	(0.033)
ln(webIndex)	-0.021						
	(0.019)						
ln(webInnov)		0.008					0.016**
		(0.009)					(0.009)

Table 5-1. 3SLS regression results (continued)

Variables	**1**	**2**	**3**	**4**	**5**	**6**	**7**
ln(webCollab)			-0.009				-0.007
			(0.012)				(0.012)
ln(webOpen)				-0.003			-0.003
				(0.004)			(0.004)
ln(webRD)					-0.030*		-0.033**
					(0.021)		(0.019)
ln(webIP)						-0.001	-0.001
						(0.003)	(0.003)
Constant	0.442***	0.423***	0.433***	0.435***	0.438***	0.431***	0.433***
	(0.043)	(0.044)	(0.044)	(0.044)	(0.044)	(0.045)	(0.042)
Nb obs.	1,289	1,289	1,289	1,289	1,289	1,289	1,289
R2	0.199	0.213	0.208	0.21	0.192	0.207	0.196
F	351.76***	358.87***	355.27***	356.05***	349.81***	354.73***	354.92***
Variables	**1´**	**2´**	**3´**	**4´**	**5´**	**6´**	**7´**
2. Investment							
Investment							
TimetoMarket							
TechFirm	0.440***	0.485***	0.481***	0.507***	0.430***	0.448***	0.402***
	(0.089)	(0.088)	(0.087)	(0.088)	(0.090)	(0.088)	(0.089)
dYoung	0.153**	0.148**	0.144**	0.149**	0.150**	0.163**	0.154**
	(0.084)	(0.084)	(0.084)	(0.084)	(0.084)	(0.084)	(0.083)
d2011	-0.264*	-0.270*	-0.256*	-0.243*	-0.261*	-0.224*	-0.197
	(0.171)	(0.171)	(0.170)	(0.171)	(0.170)	(0.170)	(0.169)
d2013	-0.342**	-0.379***	-0.365***	-0.401***	-0.339**	-0.362***	-0.305**
	(0.151)	(0.151)	(0.151)	(0.151)	(0.151)	(0.150)	(0.150)
d2015	0.180	0.156	0.179	0.146	0.177	0.169	0.201*
	(0.152)	(0.153)	(0.152)	(0.153)	(0.152)	(0.152)	(0.151)

Table 5-1. 3SLS regression results (continued)

Variables	**1´**	**2´**	**3´**	**4´**	**5´**	**6´**	**7´**
ln(Revenue)	0.019	0.019	0.012	0.015	0.020	0.021*	0.015
	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
dDigMediaICT	0.161*	0.155*	0.157*	0.160*	0.177**	0.193**	0.150*
	(0.100)	(0.101)	(0.099)	(0.100)	(0.099)	(0.099)	(0.100)
dLifeSciHealth	1.068***	1.125***	1.090***	1.154***	1.051***	1.075***	1.020***
	(0.142)	(0.141)	(0.141)	(0.141)	(0.142)	(0.140)	(0.141)
dMaterialMnfg	-0.042	-0.030	-0.031	-0.011	-0.043	-0.031	-0.034
	(0.123)	(0.123)	(0.123)	(0.124)	(0.123)	(0.123)	(0.122)
dFinland	0.408***	0.376***	0.368***	0.412***	0.401***	0.404***	0.485***
	(0.106)	(0.106)	(0.105)	(0.109)	(0.106)	(0.105)	(0.108)
ln(webIndex)	0.242***						
	(0.067)						
ln(webInnov)		0.089**					0.008
		(0.040)					(0.043)
ln(webCollab)			0.183***				0.125***
			(0.045)				(0.048)
ln(webOpen)				0.049***			0.027*
				(0.020)			(0.020)
ln(webRD)					0.266***		0.141**
					(0.072)		(0.078)
ln(webIP)						0.042***	0.036***
						(0.008)	(0.009)
Constant	0.590***	0.735***	0.741***	0.770***	0.639***	0.717***	0.509***
	(0.158)	(0.150)	(0.143)	(0.144)	(0.151)	(0.143)	(0.155)
Nb obs.	1,289	1,289	1,289	1,289	1,289	1,289	1,289
R2	0.129	0.123	0.131	0.124	0.129	0.137	0.149
F	190.50***	181.46***	194.61***	182.93***	191.56***	204.26***	226.43***

Notes: Standard errors in parentheses. Unilateral level of significance test. ***, **, * show significance at the 1%, 5% and 10% level respectively.

5.6.2 Marginal effect analysis

To illustrate how web-based innovation signalling relates to attracting investment, different marginal effects were calculated[40]. They show the change in investment resulting from a standardised unit change in a web-based indicator, while holding all other independent variables constant[41]. Model 1´ measures the impact of the real usage of web-based innovation signalling on the investment received, while Model 7´ measures the relative impact of each web-based indicator, ceteris paribus.

All innovation-signalling impact

Expected values of investment in relation with the *WebIndex* indicator (as presented Model 1´ in) were generated. Results are presented in mostly[42] standard increments of a web-based indicator unit (x) by using its corresponding mean ($\overline{x}$) and standard deviation (δ). Table 5-2 presents the corresponding investment predicted values, the differences in variation with the standard deviation increments. The marginal effect calculations suggest that without any web-based innovation signalling, firms raise investments of $152,780. Firms that reach the mean *WebIndex* signalling increase their investment by $75,750 on average, which represents a 50% augmentation. With every standard deviation increment, firms obtain more investment, but with a diminishing return as illustrated in the differences column.

Table 5-2. Marginal increase in investment by marginal increase in web-based signalling

x =WebIndex	Investment	Differences	
0	$152,780		
$\overline{x}$ (mean)	$228,540	$75,750	(49.58%)
$\overline{x} + \delta$	$267,930	$39,400	(17.24%)
$\overline{x} + 2\delta$	$296,140	$28,210	(10.53%)

[40] Performed after an OLS regression from Model 1´. The OLS results are available in the supplementary materials.
[41] We used standard values from web-based indicators and a regression model to estimate the value of investment (y) based on the transformed value of x and other independent variables. We then converted the value of y back to its original scale in dollars.
[42] Due to heavily skewed distributions towards zero, the standard deviation is often close to or larger than the distance between zero and the mean. Thus, the results start from zero up to the mean and then in multiples of the standard deviation.

Table 5-2. Marginal increase in investment by marginal increase in web-based signalling (continued)

x =WebIndex	Investment	Differences	x =WebIndex
$\overline{x} + 3\delta$	$318,560	$22,420	(7.57%)
$\overline{x} + 4\delta$	$337,370	$18,810	(5.90%)
$\overline{x} + 5\delta$	$353,690	$16,310	(4.83%)
$\overline{x} + 6\delta$	$368,160	$14,470	(4.09%)
$\overline{x} + 7\delta$	$381,210	$13,050	(3.54%)
$\overline{x} + 8\delta$	$393,130	$11,920	(3.13%)
$\overline{x} + 9\delta$	$404,120	$10,990	(2.80%)

Note: $\overline{x}$ = 0.2719 and δ= 0.2711.

All innovation-signalling impact

When individually considered in Models 2´ to 6´, all web-based signals have a significant and positive relationship with investment. Model 7´ allows for the cumulativeness and relative performance of each web-based indicator[43]. Table 5-3 presents the predicted investment values associated with varying a given indicator in standard deviation increments[44], while holding all other web-based indicators to zero.

Table 5-3. Marginal increase in investment by marginal increase in web-based signalling

x =	*WebInnov*	Differences	*WebCollab*	Differences	*WebOpen*	Differences
0	$132,965		$132,965		$132,965	
$\overline{x}$ (mean)	$136,215	$3,250	$166,724	$33,760	$155,161	$22,196
$\overline{x} + \delta$	$137,453	$1,238	$192,210	$25,485	$163,625	$8,464
$\overline{x} + 2\delta$	$138,107	$654	$206,623	$14,413	$166,929	$3,304
$\overline{x} + 3\delta$	$138,556	$448	$216,918	$10,295	$169,037	$2,108

[43] Comparing web-based indicators can be challenging due to the lack of a common scale and variations in popularity of different keywords. Relying solely on raw values can be misleading, but using means, standard deviations, and analyzing them based on their distribution can help evaluate their relative impact.

[44] Marginal effect analysis was performed in Stata 16.1 with OLS regression from Model 7´ from the simple web-based construct.

Table 5-3. Marginal increase in investment by marginal increase in web-based signalling (continued)

x =	*WebRD*	Differences	*WebIP*	Differences
0	$132,965		$132,965	
$\overline{x}$ (mean)	$162,583	$29,618	$220,622	$87,657
$\overline{x} + \delta$	$179,694	$17,111	$236,512	$15,890
$\overline{x} + 2\delta$	$191,395	$11,700	$241,337	$4,826
$\overline{x} + 3\delta$	$200,409	$9,015	$245,635	$4,298

At $x = \overline{x}$, Table 5-3 reports the opportunity cost of not signalling a concept on a website. The value of the predicted investment is at its maximum point with Web-based IP signalling is noticed ($220,622), ceteris paribus. The differential impact of signalling at the mean value $\overline{x}$ of given concept compared with no signalling of that same concept is highlighted in the Differences columns. For instance, the importance of web-based IP signalling increases by $87,657 (66%) by signalling at the mean value i.e., more than double the impact of the next better signalling factor which is collaboration ($33,760; 25%), closely followed by R&D ($29,618; 22%) and OI ($22,196; 16.5%). The innovation culture concept does not seem to have a differential impact on the predicted investment received, only $3,250 difference (about 2.4%) is observed.

As was observed with the *WebIndex*, notable diminishing returns are observed for all the web-based signalling once the mean value is reached. This appears to be especially important with *WebIP* seeing only an $8,464 (5.5%) increase at $\overline{x} + \delta$. Collaboration seems the most resistant to diminishing returns with the highest amount at $\overline{x} + 3\delta$.

The figures below show the distribution of investment received as a function of the sum of the standardised keywords frequencies by factor. In Figure 5-2. Investment received by the sum of standardised keyword frequencies by concept – short scale, the scale is truncated to better appreciate the mere presence of a factor on the investment received. The sharp increase of the impact of standardised keywords related to IP is especially notable.

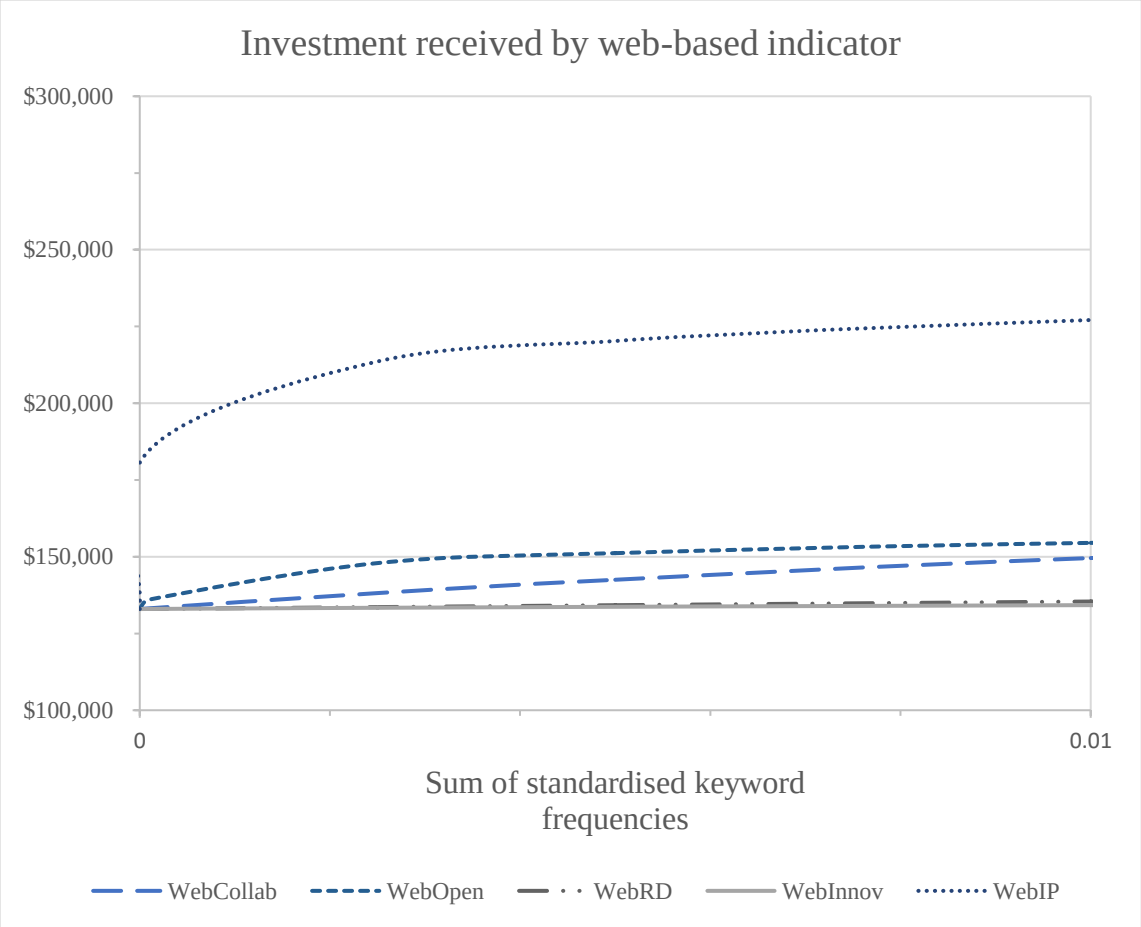


Figure 5-2. Investment received by the sum of standardised keyword frequencies by concept – short scale

Figure 5-3 presents a larger scale that shows the cumulative effects of increased relative frequencies per factor. As we show previously, collaboration-related signal seems to perform well in relation to the other terms on the investment received. In contrast, the innovation culture concept distribution seems considerably flat compared to the other distributions.

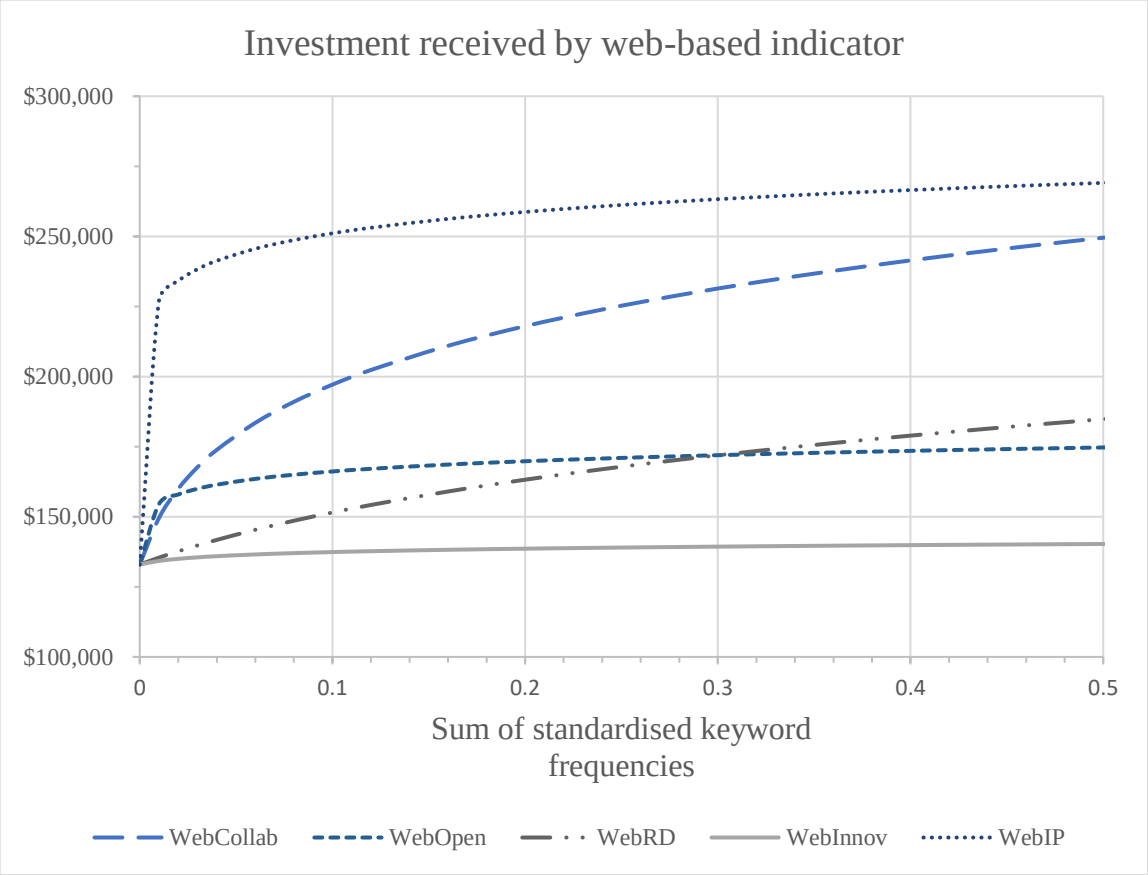


Figure 5-3. Investment received by the sum of standardised keyword frequencies by concept – long scale

5.7 Post-hoc analysis

To conclude the exploration of our triangular relationship, we added innovation evidence as an independent variable in Model 1' to segment the marginal effects based on innovative profiles. We generated an ordinal index of innovation evidence using a 1-to-3 scale based on a combination of *DegNoveltyInnov*, *nbInnov*, and *PropRevInnov*. The resulting categorical variable (*oInnovEvidence)* takes the value 1 when no evidence of innovation is measured (21.57%), the value 2 when some evidence of innovation is detected (54.85%) and 3 when they exhibit strong

evidence of innovation (23.58%). An OLS regression was then estimated[45] on a modified equation 5.2, where we added *oInnovEvidence* as an independent variable[46].

Marginal effects associated with different levels of innovation evidence (using *oInnovEvidence*) and web-based signalling on the investment received were first segmented by sector (see Table 5-4). The results suggest that web-based signalling firms are always associated with higher investment, even if they do not seem to have any evidence of innovation. Firms in digital media and ICT that have strong evidence of innovation but do signal on their corporate website are outperformed by firms without any evidence of innovation but that do signal. For each sector, the firms with the strongest evidence of innovation that use web-based signalling systematically outperform any other non-signalling firms from the same sector in our sample.

Table 5-4. Investment received by web-based & evidence of innovation and sectors

Sectors	**Innovation evidence**	**Signalling**	**Not signalling**	Δ (***Abs***)	Δ (%)
Digital media ICT	Strong	$236,682	$158,642	$78,040	49.19%
	Some	$222,926	$148,631	$74,296	49.99%
	None	$209,831	$139,100	$70,731	50.85%
Life sciences & health	Strong	$668,759	$473,099	$195,659	41.36%
	Some	$634,271	$447,999	$186,271	41.58%
	None	$601,437	$424,104	$177,333	41.81%
Materials & manufacturing	Strong	$186,625	$122,211	$64,414	52.71%
	Some	$175,271	$113,948	$61,323	53.82%
	None	$164,462	$106,081	$58,381	55.00%
All sectors ($\overline{x}$)	Strong	$364,022	$251,318	$112,704	44.85%
	Some	$344,156	$236,859	$107,297	45.30%
	None	$325,243	$223,095	$102,148	45.79%

Signalling corresponds to *WebIndex* at $x = \overline{x}$, no signalling corresponds to *WebIndex* at $x = 0$.

[45] See Appendix D.

[46] Same as our previous results, there is no significant relationship found between innovation evidence and the investment received. Tests available in Appendix D confirmed their disturbances were independent.

5.8 Discussion

This exploratory paper presents a unique perspective on how firms utilise language related to innovation on their corporate websites. It attempts to understand the meaning behind the measures obtained from web-based indicators built using a document keyword frequency analysis of innovation-related factors such as innovation culture, collaboration, OI, R&D and IP protection. This study hence builds on previous research that found little evidence of correlation between web-based R&D indicators and direct measures of R&D intensity (Gök et al., 2015), but good correspondence of the former with sophisticated indices reflecting the broader importance allocated to the concept of R&D (Héroux-Vaillancourt et al., 2020).

Taking a market signalling perspective, this study investigates the web-based information emitted by a sample of 1,289 firms involved with innovation intermediaries between 2010 and 2016. It builds on previous research that found little evidence of correlation between web-based R&D indicators and direct measures of R&D intensity (Gök et al., 2015), but good correspondence of the former with sophisticated indices reflecting the broader importance allocated to the concept of R&D (Héroux-Vaillancourt et al., 2020).

Two propositions linking the web-based signal to a broad measure of evidence of innovation and to the investment received by the firm were offered. While the results suggest that the use of web-based innovation-related language is not related to the former, it is positively related to investment, which may be interpreted as an indirect measure for evidence of innovation activities. For simplicity, the term 'signalling' is used to refer to web-based indicators.

One possible explanation for the results of the first equation is that the measured evidence of innovation may correspond with radical innovations, which have been found to be less appealing for investment due to their inherently riskier nature (Burton et al., 2002). However, since radical innovations also require more external financing (Kalil, 2005; McNeil et al., 2007), firms should be more inclined to signal, which was not observed. Instead, the lack of significance of the investment received as a factor contributing to the evidence of innovation suggests that either the funds received were associated with poor innovation results or, more likely, that the ensuing innovation results are yet to come. Nevertheless, considering the extensive literature that establishes a positive correlation between external financing and innovation (Cornaggia et al., 2015; Hsu et al., 2014; Kerr & Nanda, 2015), it seems plausible that signalling played a role in

securing investment, which has yet to yield demonstrable innovation outcomes. After all, the firms in our sample consulted with innovation intermediaries to improve their innovation outcomes.

Two possible mechanisms may be at play to explain the association between the investment and the signal. First, the use of innovation-related language signals a company's innovation-related attributes to the market, making it easier to access investment (Aghion et al., 2004). Second, obtaining investment provides firms with more resources to develop their innovation capabilities (Wellalage & Fernandez, 2019), improve marketing support for innovation (Kerr & Nanda, 2015), or simply enabling a firm's culture that has the confidence to showcase its innovation-related attributes (Herzog, 2011).

This suggests that within the context of this research design, firms that engaged the most in innovation signalling does not seem to be more innovative than those with less signalling. In other words, firms with actual evidence of innovation results that could use innovation-related signalling as a badge of honour on their website do not seem to do it as much as those in need of more external resources, despite the social prestige often associated with perceptions of innovativeness. However, firms that engaged the most in innovation signalling should be considered as innovation-active firms as they are associated with the investment received (OECD & Eurostat, 2018). Therefore, this suggests that innovation-related signalling may not help to identify innovative firms, but it may contribute to identifying innovation-active firms that are seeking external support.

When evaluating the relationship between investment received and both innovation evidence and innovation signalling, the highest levels of investment were observed when the signalling aligned with the strongest evidence of innovation. This indicates that these two indicators are measuring distinct but complementary elements that support innovation financing. This innovation signalling may be a novel way to assess marketing and brand equity activities supporting innovation (see OECD & Eurostat, 2018 section 4.2.3). Simply put, firms with marketing and branding support for innovation with tangible innovation evidence seem to be receiving more investments.

The marginal effects analysis highlighted that the IP signalling has the strongest initial influence on the investment received. Pulled mainly by the term "patent", the IP protection signal appears to be a strong hint that the firm engages in real IP protection, as it is relatively simple to verify and exceedingly difficult to counterfeit. The web-based collaboration signal was not as quickly impactful as the IP indicator, but shows the most cumulative impact as it suffered the less from

diminishing returns. Firms may prove their collaborative strength by showing instances of collaboration that imply the participation of other firms, which again is something difficult to fake. It is possible that multiple displays of collaboration form a collaborative reputation, a signal valued by investors. R&D signalling behaves very similarly to collaboration signalling although it appears to be slightly less impactful. A key difference may be the collaborative language used unwittingly, which may unconsciously predispose investors to want to work with firms that display a strong collaborative culture; they would thus feel more secure to get involved, and thus mitigating investment risk. OI-related signalling follows a similar pattern as Collaboration and R&D, but with a lower relation with the investment received, and a lower cumulative impact than IP signalling.

One exception is the signal related to the innovation culture concept which clearly does not seem to convey as much value as the other innovation concepts. Innovation being considered a buzzword (Bensaude Vincent, 2014; Mirończuk & Protasiewicz, 2020), it is a very obvious term to signal, anyone may claim to be innovative. The regression results, however, showed that when considering only innovation signalling, it did have a significant positive association with the investment received. Of course, that is an incomplete story, but it does suggest that it holds some value. The better explanation may be the following: when other more fine-grained innovation factors are present on a firm's website, they seem to be stealing "the signalling show". Mentioning more specific innovation concepts carries, a more fine-grained innovation signal, rendering the broader innovation concept unnecessary in comparison. Using more precise language is more valuable, a result in line with Lachmann et al. (2001) language signalling theory model prediction.

5.9 Conclusion

An interpretation of our results suggests that firms seeking external financing may improve the process by having a website that communicates genuine innovation-related qualities. Although, the model exhibits diminishing returns, no apparent drawbacks to displaying higher amounts of innovation-related topics were observed. Furthermore, the innovation-related language should be as specific as possible pointing to factors such as IP, Collaboration, R&D and OI. If a firm possesses patents or uses any other formal IP protection mechanisms, it should mention the fact on its website as it seems to yield almost immediate benefits. The use of collaborative language should be especially encouraged as it may demonstrate the ability and willingness of a firm to work with others, which may be a potential key attractor for investors and other stakeholders.

R&D-related language may demonstrate technological know-how, and OI-related with community-based keywords may signal status within its network or an altruistic demonstration towards the local community.

This paper demonstrated that corporate websites may be used to obtain relevant information about an innovation ecosystem as it may help to identify innovation-active firms that are seeking external support. This implies that such web-based indicators may be used to complement other classical types of indicators. Additionally, the data acquisition process presented and using the Wayback Machine is a useful tool to complement existing databases and surveys-based data. Innovation agencies could easily collect data from this source to enhance their existing knowledge base.

The surveys administered by TEN were designed to assess the impact of innovation intermediaries. Assessments of the innovativeness of surveyed firms were made in the service of this primary objective. TEN surveys the entire population of firms served by an intermediary. Not all surveyed firms respond, but the average response rate of 39% is high. Times series and more recent data, as well as additional explanatory variables to ensure better robustness, are needed to explore any possible causality inferred from the propositions.

The Wayback Machine service does have limitations which may have resulted in non-selection and selection biases through the online data acquisition. Additionally, the online translation service used is not perfect and it is possible that some Finnish words were not properly translated which may create a noisy web-based signal. Furthermore, some important keywords may have been unintentionally left out as words used by the industry to relate to a specific factor may differ from those used in academia. As language evolves over time, the lag that exists between now and the survey data implies that some terms may not be as popular at that time as they are nowadays. Maintaining processes that would perform punctual web scraping systematically on selected firms paired with punctual questionnaire-based survey investigations with these same firms would be a much better long-term data acquisition process, allowing panel data collection and longitudinal studies to validate our findings and improve our understanding from web-based signal and their meanings.

Moreover, the signal indicator built with document keyword frequency analysis does not account for the context in which the words are used, which may lead to false positive and false negative results. Rule-based techniques or more advanced methods such as topic modelling with Latent

Dirichlet Allocation (LDA), or supervised machine learning methods, or even using large language model (LLM)-based technologies may improve our way of measuring innovation signalling from websites. Further research in this area would be especially important as LLM technologies have an untapped potential to obtain key insights from large corpora.

Finally, this study focused on analysing the textual content retrieved from corporate websites, but firms use various other communication channels such as images, photos, videos, web design elements, reports, and social media. These can also be examined from a signalling perspective to gain a deeper understanding of a firm's intentions, motivations, which ultimately, point to their needs. How to turn these data into valid interpretable signals at scale may give us unprecedented knowledge about our business ecosystems, allowing to address what is needed for innovation to thrive.

5.10 Author contributions

Mikaël Héroux-Vaillancourt: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, resources, software, validation, visualisation, writing—original draft, writing—review and editing. **Catherine Beaudry**: project administration, resources, funding acquisition, supervision, validation, visualisation, writing –review & editing. **Davide Pulizzotto**: methodology, software, validation. **Margaret Dalziel**: data curation, resources, validation, writing - review & editing.

CHAPTER 6 ARTICLE 3: WHERE THERE IS SMOKE, IS THERE FIRE? EXPLORING TO WHAT EXTENT INNOVATION SIGNALS FROM CORPORATE WEBSITES MEET EXPECTATIONS

Mikaël Héroux-Vaillancourt, Catherine Beaudry, Davide Pulizzotto and Margaret Dalziel

This chapter presents a research paper submitted to the Technological Forecasting and Social Change journal on March 28th, 2023. It describes how the signal obtained with document frequency analysis compares with what is expected from the literature. The results show that firms with some level of innovation evidence that are high knowledge intensive and operate in a sector that is expected to be more innovation have a stronger signal.

6.1 Abstract

This study explores how innovation-related dimensions retrieved from corporate websites correspond with objective assessments. Starting with a dataset of 1,594 respondents who answered questionnaire-based investigations between 2010 and 2016, we extracted the content of the corresponding websites via The Wayback Machine. Several innovation-related dimensions such as innovation culture, collaboration, open innovation, R&D and intellectual property are examined using Document keyword frequency analysis. Zero-inflated negative binomial regressions estimate different associations between innovation-related signalling and firms' innovation characteristics. The results suggest that innovation signalling may have potential to identify firms based on their profile in terms of innovativeness, knowledge intensity and their sector. This paper contributes our better understanding of what to expect from web-based innovation indicators derived from corporate websites.

Keywords: Innovation; Signalling; Web mining; Document keyword frequency analysis

JEL codes: O320, L250

6.2 Introduction

Most quantitative research in the field of innovation and technology management relies on public databases and questionnaire-based surveys. Both of these data sources have undeniable benefits, but also limitations. For instance, the former may provide many different data, but may end up lacking important items for properly measuring specific innovation aspects intended for a given

study. The questionnaire-based method typically addresses this shortcoming allowing researchers to gather more precise information tailored for their studies. However, that increased data precision comes with significant extra cost as its growing popularity leads to participants being over-solicited causing a decline in response rates. This is a concern particularly in studies relying on the input of business leaders who always have to deal with many pressing issues. In light of this, researchers are exploring ways to lighten, and even alleviate, the burden on respondents by reducing the length and complexity of questionnaires, and also exploring alternative sources of data that can provide further insights into factors that affect innovation within firms. Given that questionnaires are typically designed to optimise response rates, researchers often face the challenge of selecting only the most salient questions, resulting in the loss of potentially valuable and nuanced information that could be used to contextualise the findings. Indeed, an increasing number of innovation researchers acknowledge the importance of obtaining more detailed contextual information in quantitative studies to gain a deeper understanding of the factors that contribute to the success or failure of certain innovation practices (Bogers et al., 2017).

In the late 90s, researchers in the field of information science introduced the idea of using the internet as a data source for research, creating a new discipline called webometrics (Almind & Ingwersen, 1997). The primary areas of webometrics research included link analysis, blog searching, and web impact evaluation. Link analysis entails examining the network of hyperlinks found on web pages in the hopes that they may reveal the existing relationships between different organisations (Hyun Kim, 2012; Katz & Cothey, 2006; Vaughan, 2004). This approach has worked well to shed light to existing collaborative network structures with the importance of actors within them (Minguillo & Thelwall, 2012; Stuart & Thelwall, 2006). Blog searching became an area of research as a result of the vast diversity and amount of publicly generated information pushed by individuals at any given time. Blog searching is more common these days on social media platforms such as Twitter (Thelwall et al., 2011), while Google Trends is a useful tool for keeping track of topics' relative importance at any given moment (Choi & Varian, 2012). Web impact assessment is based on the idea that online documents or ideas with greater influence will generally be cited or discussed more often, assuming no other major factors are involved (Thelwall, 2009).

Nowadays, the digital realm continues to expand at a rapid pace. Online information has indeed grown exponentially in volume, variety, and velocity a phenomenon commonly referred to as "big data". In this complex information ecosystem, organisations such as companies, public institutions,

and civil societies have all invested heavily. These organisations have at least a website, and some have also established a presence on platforms such as YouTube, Twitter, Facebook, and LinkedIn. Through these new media channels, firms actively signal vast amounts of relevant information to their stakeholders, including information about their mission, activities, areas of interest, and partnerships, and so on. This rich information ecosystem holds enormous potential for providing valuable insights at large scale. Indubitably, the successful identification and translation of online signals related to innovation into credible and dependable innovation-related metrics presents a potentially valuable opportunity for augmenting the level of intricacy in quantitative models.

Nearly all innovation-active firms maintain a current website that features key communication priorities with stakeholders. By performing web impact assessment on the dimensions related to innovation presented on these websites, researchers can gain insight into the emphasis that firms place on these dimensions. Based on classical content analysis (Krippendorff, 1980), web content analysis attempts to connect textual information with its context by converting unstructured text data from web pages into structured data for quantitative research. Given this potential, the use of content analysis methods to analyse web content and extract relevant insights is an enticing opportunity for researchers. In the last decade, we have seen a significant interest from scholars to explore and use web-based textual content to build indicators based on keyword frequencies for innovation management quantitative studies.

While most of these studies employed these web-based indicators as ways to add contextual information to their models (see for example Arora et al., 2013; Li et al., 2018; Youtie et al., 2012), few studies targeted firms' website as a potential data source to assess their innovation profile. For example, Al-Hassan et al. (2013) attempted to use corporate websites to cluster firms into the same self-reported North American Industry Classification System code (NAICS), but was only marginally successful. Gök et al. (2015) found no correlations between R&D web-based indicator built with a frequency analysis method from this data source and a traditional R&D questionnaire-based indicator, hence suggesting these two measures were measuring different concepts. Héroux-Vaillancourt et al. (2020) continued this analysis comparing web-based with questionnaire-based data and reported strong significant correlations with broader concepts related to R&D and Intellectual Property (IP) protection, thus hinting that the same concepts were in fact being measured. Taking a signalling perspective, it was demonstrated that the firms exhibiting the highest level of innovation evidence did not necessarily engage in the most signalling, whereas

those signalling the most tended to be more closely associated with larger investment received (Authors, forthcoming). The author then proposed that these indicators may reflect marketing and branding equity activities that innovation-active firms undertake to support their innovation efforts and attract external support.

Following in their footsteps, this exploratory study aims to investigate which firm attributes are most closely associated with innovation-related signalling and to evaluate whether these signals align with what one would expect based on the existing literature.

The Evidence Network (TEN) provided data stemming from 29 investigations between 2010 and 2016. We used the Wayback Machine to collect the textual data from the websites of 1,594 organisations, essentially small firms. Indicators were built using Document frequency analysis based on the occurrences of sentences containing innovation-related keywords, i.e., linked to classic factors known to contribute to innovation that may be advantageous from the firms' perspectives to signal. Zero-inflated negative binomial regression results show that web-based indicators correspond mostly with expectations regarding overall innovation and R&D, while showing some promises with innovation culture, collaboration and IP. This exploratory paper contributes to understanding further the opportunities and limitations associated with using corporate websites to assess firms' innovation profiles.

The remainder of the article is organised as follows: Section 6.3 presents the theoretical development; Section 6.4 describes the various methodologies used to collect data and construct the necessary variables; Section 6.5 presents our empirical results; Section 6.6 discusses the implications of results; and lastly, Section 6.7 concludes.

6.3 Theoretical development

This section first outlines our approach to measuring innovation from a signalling perspective. Then, how this signalling may be linked to key innovation elements that make up the innovation profiles of firms is explored based on our current understanding of the literature.

6.3.1 Innovation signal and innovation

Everything firms mention on their websites serves as a means to achieve strategic objectives such as selling its products or services, or enhancing reputation and brand identity (Chun et al., 2005; Gammoh et al., 2006; E. R. Gray & Balmer, 1998; Hall et al., 2000). Firms highlight innovation-

related dimensions on their websites for the same reason - to achieve strategic goals, such as attracting external financing (Authors, forthcoming). Some of these dimensions can be identified from the innovation management literature and are expected to be signalled from the firms' perspective. The dimensions considered are innovation culture, collaboration, open innovation (OI), research and development (R&D), and intellectual property (IP).

The positive connection between innovation and a supporting culture has been established in the literature (Dobni, 2008). Early research indicated that there exist various market orientations that distinguish companies in terms of their approach to innovation, leading to different outcomes in terms of their performance (Aldas-Manzano et al., 2005; Hurley & Hult, 1998; O'Cass & Viet Ngo, 2007). These orientations stem from the beliefs, attitudes, and actions that are fundamental components of an organisation's culture (Schein, 2004). Herzog (2011) showed there is internal environment of a firm or business unit that better supports innovation.

As a result, organisations are increasingly recognising the importance of fostering a supportive culture for innovation, which has significant implications for human resource development professionals as they strive to build teams that can generate innovative work (Folkestad & Gonzalez, 2010). To achieve this, firms must attract employees who share a belief system compatible with an innovation culture (Schein, 2004). The collection of data on activities that build a supportive culture, as suggested by OECD & Eurostat (2018), includes communication about the importance of innovation, which encompasses vision and strategy. Hence, firms that prioritise innovation are expected to communicate the significance of their innovation efforts on their websites. Furthermore, there are external incentives for firms to signal their innovation prowess, as it is highly valued by investors (Sood & Tellis, 2009). In summary, firms have strong incentives to signal their innovation culture to both future and current employees and investors.

Collaboration plays a positive role in numerous stages of the innovation process. First, it may improve idea generation, evaluation and validation (Koen et al., 2014). Then, collaboration may enhance a firm's competitiveness (Hagedoorn et al., 2000; Roja & Nastase, 2013), and stronger innovation performance which then results in patent numbers, sales growth, and return on innovation sales (Arvanitis, 2012; Belderbos et al., 2004; Carboni, 2013). Second, collaboration is especially needed to lower R&D expenses, sharing risk, and enhancing performance while

developing new complicated technologies (Johnson & Filippini, 2009; Jordan & Lowe, 2004; Parker, 2000).

Thus, firms need to find quality partners to innovate. Additionally, firms that participate in collaborative alliances may enhance their reputation and brand value (Cooke & Ryan, 2000). Moreover, collaboration and cooperation are highly valued by employees, as they can improve teamwork and satisfaction in the workplace (Driskell et al., 2018). Therefore, it would be reasonable to expect that firms seeking to attract potential partners and employees would use collaboration-related language on their websites.

OI describes a distributed process in which organisations exchange intellectual assets such as technologies, know-how and knowledge to improve the innovation performance (Chesbrough, 2003a; Chesbrough & Bogers, 2014). Numerous studies have shown that acquiring external knowledge or technologies to enhance internal R&D processes has a positive impact on a firm's innovation performance (W. M. Cohen & Levinthal, 1990; Henkel, 2006; Laursen & Salter, 2006b) and distributing knowledge and technologies may have positive effects on a firm's innovation performance (Héroux-Vaillancourt & Beaudry, 2018). It is important to note that the relationship between OI practices and firm performance varies among studies (Bogers et al., 2017). The benefits of OI seem to reach a maximum after a certain level of effort before leading to diminishing returns (Fey & Birkinshaw, 2005; Laursen & Salter, 2004, 2006b).

The concept of OI gained so much popularity in the corporate world in the recent years (Chesbrough, 2012; Chesbrough & Brunswicker, 2014; Dodgson et al., 2006) that it has been identified as a buzzword to signal its "openness" which may have branding value (Hagel & Brown, 2008). Similar to cultivating an innovative culture, OI is positively related to human resources considerations, such as hiring and training employees who hold favourable attitudes towards OI practices (Podmetina et al., 2013). Therefore, it is reasonable to expect that firms would display open innovation-related concepts on their websites as a way to enhance their brand value and attract potential employees who are compatible with their innovation practices (Siqueira et al., 2021; van Dijk et al., 2014).

R&D intensity has long been considered a traditional indirect measure for innovation (Becheikh et al., 2006). It measures the level of investment a firm puts into developing its own capabilities for invention (Griliches, 1990, 1994, 1998; Hausman et al., 1984; Hitt et al., 1997). Several studies

have shown a positive relationship between R&D intensity and innovation performance (Baysinger & Hoskisson, 1989; W. M. Cohen & Levinthal, 1990; Deeds, 2001; Greve, 2003; Hall, 1990; Parthasarthy & Hammond, 2002). While R&D intensity used to be considered as a proxy for innovation, it has been argued to be flawed because innovations can also come from sources other than the R&D department (Michie, 1998; OECD & Eurostat, 2018). This is increasingly the case as firms adopt more OI practices. However, R&D remains a significant component in understanding innovation performance, especially for technological companies (OECD, 2015; OECD & Eurostat, 2018).

Furthermore, financing the requirements of R&D activities is often challenging due to their risky nature (K. D. Miller & Bromiley, 1990). The funding of R&D is viewed as crucial for innovation, especially for smaller businesses that have fewer resources available (Aghion et al., 2004; Wellalage & Fernandez, 2019). Despite the financial challenges, R&D is valued to some extent by the investment market (Bae & Kim, 2003; Gharbi et al., 2014). Moreover, engaging in R&D requires firms to have the absorptive capacity to develop and apply new knowledge (W. M. Cohen & Levinthal, 1990), which in turn necessitates the recruitment and retention of highly qualified personnel. As high-qualified employees have more employment options than ever before with the emergence of globalisation, nowadays firms compete with each other to hire the best candidates (Chesbrough, 2003a). Thus, there are strong market imperatives that encourage firms to signal R&D as a means of attracting investment, highly qualified personnel, and high-quality collaborators, as mentioned previously in relation to the significant ties between collaboration and R&D.

IP protection is a way for firms to gain a competitive advantage by securing exclusive rights over their technology, branding, copyright, industrial designs, and other intellectual property (Teece, 1986). Patents, which are often a result of R&D activities, are a commonly used to measure a company's inventiveness (Coombs et al., 1996; Flor & Oltra, 2004; OECD & Eurostat, 2018). Patents enable firms to increase the return on their technology development investment by granting them temporary monopoly power, which they can also use to license their technology or sell their patents (A. Arora, 1995; Chesbrough, 2003a; Feldman & Florida, 1994; Fosfuri, 2006; Mazzoleni & Nelson, 1998; Merges, 1999; Rivette & Kline, 2000). Authors have found a positive relationship between appropriability and innovation performance that eventually reaches a point of diminishing returns (Laursen & Salter, 2006a).

Firms may highlight that they formally protect their intellectual property to deter the potential imitators. Evidence suggests that patenting activities are positively related to stock market valuation (Hall et al., 2000, 2005). Furthermore, patents are viewed as a powerful signal from venture capital (VC) investors (Colombo et al., 2023). Therefore, it is reasonable to expect that firms would signal their formal IP protection on their websites as a means of attracting potential investors and deterring competitors.

All the evidence presented so far highlight the strong incentive for firms to signal to attract and retain the means to innovate. The next section dives into the potential relationships between innovation-related signalling and the dimensions described above, and firms' innovation attributes based on what would be expected from the literature.

6.3.2 Innovation signal and the innovation literature

Given that innovation-related signalling has been shown to correspond to innovation activities (Authors, forthcoming), and that innovation activities are expected to lead to some level of innovation outcomes (OECD & Eurostat, 2018), it is reasonable to expect that higher levels of innovation-related signalling will also correspond to higher levels of innovation outcomes. We therefore propose the following:

Proposition 3.1: There is a positive association between a firm's innovation evidence level and the strength of its innovation-related signal, including a) innovation culture, b) collaboration, c) open innovation, d) R&D and e) IP-related signals.

Generally, a shorter time to market provides companies with an opportunity to rapidly launch new products or services, potentially granting a competitive edge over their rivals (Afonso et al., 2008; Lieberman & Montgomery, 1988). However, rushing to market may also result in lower-quality products or services, which may then negatively impact a company's reputation and ultimately its innovation performance (Bayus, 1997; M. A. Cohen et al., 1996). Furthermore, companies operating in the same industry and offering similar products but with a shorter time to market suggest superior processes and, consequently, higher innovation process performance (Pisano, 1997). Because of the mixed evidence supporting the impact of both shorter and longer time-to-market on innovation performance, this relationship may be bidirectional. Although similar arguments can be made for most innovation-related dimensions, firms that depend on patents

typically spend more time in the R&D phase (Roin, 2013). As a result, we suggest these two propositions:

Proposition 3.2: There is an association between a firm's time to market and the strength of its innovation-related signal, including a) innovation culture, b) collaboration, c) open innovation, d) R&D-related signals.

Proposition 3.2e: There is a positive association between a firm's time to market and the strength of its IP-related signal.

The ability to leverage knowledge into implemented ideas is fundamental to any innovation. Indeed, an actor possessing sufficient knowledge can add, alter, and combine previously assimilated information from a new angle, making it usable for potential exploitation (Zahra & George, 2002). This transformation involves a process of bissociation, i.e., a process of applied creativity that involves the association or recombination of two ideas or two pieces of knowledge that have nothing visibly in common, to create a third, fundamentally novel element (Koestler, 1964). Furthermore, firms with more knowledge are more effective to absorb internally external knowledge which has a crucial impact on the ability of firms to innovate (W. M. Cohen & Levinthal, 1990), which is needed to benefit from open innovation opportunities (Bogers et al., 2017). Firms that operate in the development and deployment of complex emerging technologies, as well as those operating in complex value chains, require a multitude of expertise that can only be brought together through collaboration (Armellini et al., 2014, 2017; Johnson & Filippini, 2009). Research has shown that university research, which generates new knowledge for the industry, has a positive impact on the development of scientists, skilled labour, patents, and innovation in high-tech industries (Kim et al., 2014). Thus, it is expected that firms with high knowledge-intensity are more associated with a stronger innovation-related signal, hence, the following proposition:

Proposition 3.3: There is a positive association between a firm's level of knowledge intensity and the strength of its innovation-related signal, including a) innovation culture, b) collaboration, c) open innovation, d) R&D and e) IP-related signals.

An industrial sector can be analysed through the lenses of technological regimes which can be decomposed in terms of opportunities, properties of the knowledge base, cumulativeness and appropriability conditions (Breschi et al., 2000; Breschi & Malerba, 1997). The characteristics of

a firm's knowledge base are linked to the type of innovation endeavours it entertains. Technological knowledge varies significantly across different technologies, exhibiting varying levels of specificity, tacitness, complexity, and independence (Winter, 1997). Technological opportunities can be understood as the probability of achieving innovation based on the amount of money invested in research and development. Sectors with higher opportunities are ripe for innovative conditions. The cumulative nature of technological advancements can be attributed to the fact that present-day knowledge and innovation serve as the foundation for future advancements. Each innovation leads to a chain of subsequent innovations, which either incrementally improve upon the original or introduce novel knowledge that can be applied to related areas. Environments with high levels of cumulativeness are often characterised by a steady flow of innovative activities and increasing returns, as each new advancement builds upon the knowledge gained from past innovations. Finally, appropriability refers to the ability to safeguard innovations from imitation through the use of intellectual property mechanisms which allow the firms to capitalise on the profits generated by innovative activities (Levin et al., 1987; Teece, 1986).

As the sectors related with Life science and health require an advanced knowledge base which follows closely the latest scientific research, they have a diverse set of cumulative technologies, they offer high opportunities and strong appropriation mechanisms through patents activities (Castellacci, 2008; Coriat et al., 2003). Therefore, firms that are operating in these sectors are expected to be strongly associated with innovation activities. Furthermore, Digital media and other ICT sectors are expected to generate high levels of innovation activity due to their significant opportunities, advanced yet widely accessible knowledge base, and ongoing flow of cumulative advancements from both hardware and software fronts, but appropriation is often limited to know-how and copyright for software (C. Antonelli, 1998). The Manufacturing sector innovation characteristics depends on the scale of the operations and the product characteristics (Pavitt, 1984). SMEs with higher knowledge intensity in sectors such as manufacturing or energy and green tech typically specialised their production which yields high opportunities and strong appropriability through patents (Castellacci, 2008; Pavitt, 1984). Other sectors such as Natural resources, or energy or environmental services are mostly process oriented with lower opportunities (Baldwin & Rafiquzzaman, 1995). Thus, the following propositions are described as:

Proposition 3.4: A firm operating in either Life science and health or Digital media and other ICT have a higher association with the strength of its innovation-related signal, including a) innovation culture, b) collaboration, c) open innovation, and d) R&D signals than firms from Natural resources, or energy or environmental services or other sectors.

Proposition 3.4e: A firm operating in either Life science and health or Manufacturing sector have a higher association with the strength of its IP-related signal than firms from digital media and other ICT Natural resources, or energy or environmental services or other sectors.

Proposition 3.5: There is a positive interaction between a firm's sector and a high knowledge intensity level on the strength of its innovation-related signal, including a) innovation culture, b) collaboration, c) open innovation, d) R&D and e) IP-related signals.

Pavitt (1984) sectorial taxonomy allows to differentiate sectors based on how they relate to each other on the innovation value chain with four main categories: supplier-dominated, scale-intensive, specialised suppliers, and science-based industries. Each category describes innovation-related patterns. For instance, the supplier-dominated sectors are mainly SMEs that have low internal innovation capacity and focus on process innovation to enable cost reduction. The scale-intensive label is attributed to large firms with high capacity for process innovation and vertical technological diversification that optimise for economies of scale. Specialised suppliers are small size firms that have high innovation capacity focusing on very specific product innovation. Finally, science-based industries are also described as large with high internal innovation capacity geared toward both product and process innovation. Statistics Canada made an adaptation of this taxonomy by creating a dichotomy based on the level of wages. Scale-based industries and the addition of value-added knowledge-based industries (which regroup specialised suppliers and science-based categories) are considering providing higher paying jobs compared to the supplier-dominated which was decomposed in resource-based and labour-intensive industries to better reflect the particularities of the Canadian economy (Baldwin, 1995; Baldwin & Rafiquzzaman, 1995). These categories not only distinguish industries with higher-paying jobs, but also happen to discriminate innovative industries better from lesser innovative ones. Castellacci (2008) also adapted the taxonomy with distinguishing advanced knowledge providers (value-added knowledge-based without science-based) from mass production goods (scale-based and science-based) from another category called "Supporting infrastructure services" and the supplier-

dominated goods and services. From European Community Innovation Survey (CIS) of 2002-2004, advanced knowledge providers who are represented with knowledge-intensive business services (such as software, R&D, engineering consultancy) and specialised provider (such as SMEs in Life science and health or specialised manufacturing) firms shown to have higher levels of innovativeness than supporting infrastructure services and the supplier-dominated goods and services and similar level of innovativeness as mass production goods. It is therefore expected that all factors contributing to innovation are corresponding with higher innovation-related signalling for firms operating in either mass production or advanced knowledge-based industries. Thus, the following proposition is defined as:

Proposition 3.6: A firm operating in either mass production or advanced knowledge-based industries have a higher association with the strength of its innovation-related signal, including a) innovation culture, b) collaboration, c) open innovation, d) R&D and e) IP-related signals than firms from supplier-dominated industries.

Furthermore, Castellacci (2008) showed significant differences between advanced knowledge providers, with Knowledge-intensive business services (KIBS) showing higher level of opportunities, higher acquisition of other external knowledge, lower appropriability through patent and closer ties with universities. Thus, the following propositions are elaborated:

Proposition 3.7: KIBS firms have a higher association with the strength of its innovation-related signal, including a) innovation culture, b) collaboration, c) open innovation, and d) R&D-related signals than firms operating as specialised-suppliers.

Proposition 3.7e: A firm operating as specialised-suppliers have a higher association with the strength of its IP-related signal than firms operating as knowledge-intensive business services.

Other contextual factors are known to influence innovation such as the firm's size, age and country (Becheikh et al., 2006). Larger firms have more resources to dedicate to R&D and are less exposed to market volatility (Van Dijk et al., 1997). Indeed, the comparative advantage of large firms are believed to be a mainly in terms of access to resources, while small companies' strengths are often attributed to behavioural traits (Vossen, 1998). Furthermore, there is a comprehensive literature on SMEs, especially on tech firms, which have been identified as important participants in the process of introducing new ideas and technology to the market (Acs & Audretsch, 1990; Baumol, 2005; Hicks & Hegde, 2005).

The relationship between a firm's age and innovation is also not trivial. Older companies are likely to accumulate experience and knowledge over time building absorptive capacity (W. M. Cohen & Levinthal, 1990), which may lead to more innovation than younger firms that did not have yet built that foundation (Sørensen & Stuart, 2000). However, other research showed that established procedures and routines in older firms may create some resistance to the integration of significant external advances, hindering innovation (Freel, 2003). This is further supported by a recommendation to mature firms to circumvent the innovation dilemma by allowing disruptive innovations to be developed outside of the main core business operations (Christensen & Raynor, 2013).

The innovation landscape in Canada over the past few years has given rise to concerns. Notably, the country has exhibited a growth rate that is among the lowest when compared to other developed nations while having a highly educated population with strengths in critical areas of research and science. One aspect that was especially noteworthy was a decline in financial support for R&D in all sectors, including the investment by firms. Meanwhile, Finland is often viewed as an innovation champion due to the success of his economy where it showed to be more competitive than the US and other EU countries on some key technological fronts (Oinas, 2005). The importance the Finnish government gave on national innovation systems early on and along with key public investments seems to have favoured a strong support for innovation (Fagerberg, 2017; Oinas, 2005; Torregrosa-Hetland et al., 2019). Nowadays, the Canada ranked 15th compared with 9th with Finland on global innovation index. Furthermore, differences in cultures and even the use of language-related terms may affect signalling events on corporate websites[47].

6.4 Methodology

6.4.1 Data

This section outlines two distinct methodologies for data collection. Initially, the survey data used to identify our subjects and their different characteristics will be presented. Subsequently, the web-based data collection processes, including the collection of text content data, will be explained.

[47] Although translation programs are increasingly getting better, they are not perfect and may lead to a lack of representation of Finnish innovation-related signals.

The Evidence Network (TEN) provides impact assessments and program evaluations for organisations offering intermediary services. The data provided for this study comes from a sample collected between 2010 and 2016, from 29 investigations (also named programs) with 4,740 observations from Canadian and Finnish business entities, including firms, non-profits and public institutions.

Questionnaires were distributed online and were intended to be completed in around 15 minutes. Data collection for each survey spanned a maximum of one month and typically included 30 questions, separated into four sections: 1) Preliminary questions about the activities of the respondents in relation to the intermediary program, such as the year of the first engagement, the importance of the services provided, and other particular inquiries regarding a particular intermediary program, such as specific financial programs and project types 2) the intermediary's impact on the organisation's resources and capabilities; 3) the intermediary's impact on market performance; and 4) closing questions about the organisation's characteristics (e.g., the number of employees, revenue, year of foundation, and so on).

An average response rate of 39% was achieved across all investigations in our sample due to the close ties between the organisations and the intermediaries, as well as diligent calls and emails sent in a follow-up.

From the different surveys, we identified all the items that corresponds to establish innovative profiles and other general characteristics of the firms: the number of innovations introduced in the recent past, the percentage of the firm's revenues that can be attributed to innovations, the degree of novelty of innovation, the time to market, year of foundation, sectors and employment[48].

As the questionnaires are tailored-made to meet the distinct requirements of various intermediaries, this led to many differences in measurement scales. Consequently, merging the data obtained from

[48] Examples of the questions used is provided in Appendix B

the various surveys necessitated a meticulous process to harmonise the variables used in the study[49]. Additionally, imputation was performed to slightly augment the sample size[50].

To obtain the necessary web-based data, we needed to retrieve the corporate websites of respondents who interact with intermediaries. Each entry was individually processed to gather this missing data, resulting in 2,949 respondent websites retrieved. These organisations were categorised into tech firms, low-tech firms, advanced knowledge service firms, other service firms, non-profit/non-governmental organisations (NGOs), public institutions, and undefined. Since this study focuses on firms, we removed the non-profit (64), public institutions (75), and undefined respondents (30). Our final sample includes 2,780 respondents (58.6%) from 2,483 unique companies with functional website domains.

To make sure our web-based data matches the temporality of the questionnaire-based data, we downloaded the corporate website content from the online archive Wayback Machine[51]. The closest snapshot of a firm's homepage from June 30th of a questionnaire year was selected. From there, we extracted the text content of the homepage and all pages directly accessible from it[52]. Then, to identify the language of the web content and translate it into English, we utilised

[49] For example, the number of employees was measured by asking respondents directly how many full-time and part-time employees they had, or by presenting several options at intervals. In order to harmonize the scale, we took the midpoint of each category for both full-time and part-time employees. Then to calculate the total number of employees we summed the full-time and part-time employees. Most items were combined like this example, excepting for the sector and the degree of novelty of innovation. The harmonization of the different sectors required special attention since the different questionnaires use different categories that were not necessarily mutually exclusive. In the study, various questionnaires focused on different levels of specificity within certain sectors. Some questionnaires targeted specific sub-sectors, such as 1) Green Energy Technologies, 2) Environmental - Clean Technologies, and 3) Environmental - Water Technologies. Other questionnaires had a broader scope, including categories like 1) Energy and 2) Environment. There were also some questionnaires with even more general classifications, such as 1) Energy or Environment. To maintain consistency in the analysis, we grouped these categories into their most general form, i.e., Energy or Environment. The process resulted in 6 main sectors: Materials & manufacturing, Life science and health, Energy & environment, Digital media & ICT, Natural resources and Other. The group "Other" represents a wide range of different sectors such as financial services, food and beverage, construction, retail, entertainment and so on.

50 Imputation was performed on the number of innovations (7.7%), the percentage of company revenues attributed to innovations (13.8%), and time to market (22.6%) to enhance the sample size. The nearest neighbor methodology was applied on classes obtained from the combinations of sectors, countries, sizes, program, and organization types (either tech firm or not). The average value was computed for each class and were attributed to the corresponding missing values within each class. The imputed values constitute less than 30% of the original sample, adhering to the guidelines recommended by (Särdnal et al., 1992).

[51] Unfortunately, the Wayback Machine is not perfect as it crawls the web punctually without really known rules which has for consequences that the snapshots selection is not evenly distributed across and within websites. Despite its drawbacks, the use of the Wayback machine remains a valid data source for social sciences (S. K. Arora et al., 2016; Murphy et al., 2007).

[52] The code developed to execute our data acquisition strategy is available at: https://github.com/4point0-ChairInnovation-Polymtl/Wayback_Machine_Crawler

LibreTranslate[53] from Argos translate[54]. This is an automatic translation software based on OpenNMT[55], an open-source neural engine system for translation and sequence learning (Klein et al., 2017, 2018, 2020). We successfully obtained cleaned English text content for 86% of the initial website list.

Section 6.3 described five established dimensions related to firms' innovation that are expected to be signalled to key stakeholders through the corporate websites: innovation culture, collaboration, OI, R&D, and IP. Sentences containing various synonyms and terms related to a dimension were identified. We expect that the more a firm displays sentences with terms associated with innovation, the more it is signalling innovation. Web content analysis with document keywords frequency analysis in sentences related to innovation is therefore an appropriate methodology to quantify these signals.

All the keywords were derived from the lexical field related to the five factors described by peer-reviewed publications and seminal books such as the Oslo Manual for innovation (OECD & Eurostat, 2018) and the Frascati Manual for R&D (OECD, 2015). Some of the keywords considered related to R&D (Gök et al., 2015; Héroux-Vaillancourt et al., 2020), collaboration (Ramdani, 2014) and intellectual property protection (Héroux-Vaillancourt et al., 2020; Rietsch, 2015) were also used in previous publications. The text content was processed using a stemmed version of the keywords to ensure that all variations of a dimension were captured (e.g., using "innov" instead of "innovation", "innovative", "innovation culture", "innovator", etc.). The final list of keywords used is obtained after numerous tests to remove keywords that were either not used by firms, or that were more likely to lead to false positives, or appeared too frequently in legal pages (such as terms of services or privacy policy)[56].

Based on the suggestion of Lachmann et al., (2001), we limited the scope of our analysis to sentences as signalling units, resulting in a total of 511,815 sentences. Of these, 142,467 sentences contained keywords related to innovation. The process is illustrated in Figure 6-1.

---

[53] LibreTranslate. (n.d.). LibreTranslate [Repository]. GitHub. Retrieved March 15, 2023, from https://github.com/LibreTranslate/LibreTranslate

[54] Argos Open Tech. (n.d.). Argos Translate [Repository]. GitHub. Retrieved March 15, 2023, from https://github.com/argosopentech/argos-translate/

55 OpenNMT. (n.d.). OpenNMT [Repository]. GitHub. Retrieved March 15, 2023, from https://github.com/OpenNMT

56 A summary of the keyword construct is provided in Appendix C.

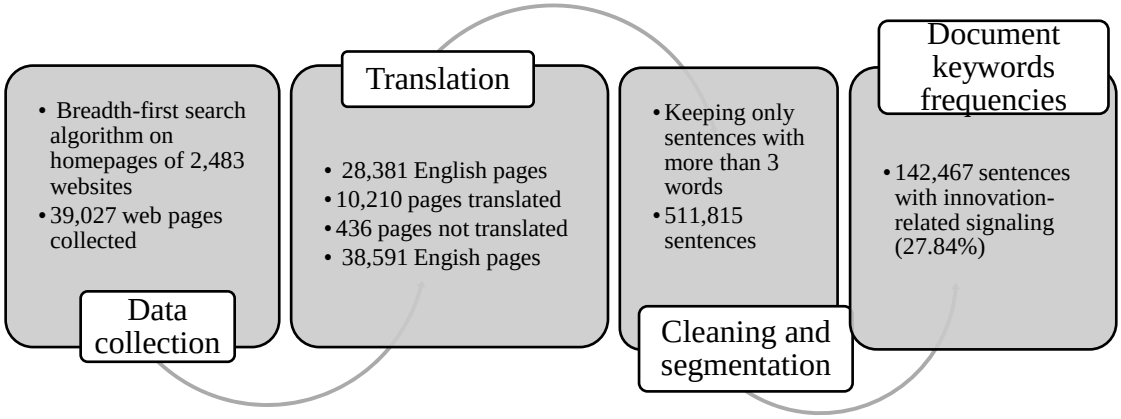


Figure 6-1. Text content collection process

A total of 1,889 observations were excluded due to the absence of data for at least one variable of interest whether dependent, independent or control. Thus, our final sample consists of 1,594 observations, of which 1,500 are unique firms. Seventy-eight firms are responsible for 94 (5.9%) repeated observations across the whole sample. Finally, there are 21 firms that generated a total of 25 observations within a same year (1.6%) across the sample.

6.4.2 Dependent variables

We observe that three R&D-related keywords ("tech", "develop", and "lab") made up 53% of the document frequencies[57]. The frequencies seemed to follow a Pareto distribution, with the top eight out of 36 keywords accounting for 82% of the cumulative frequencies. Four of these terms were related to R&D, three are related to innovation culture, and 1 to collaboration.

The R&D had the highest relative occurrence at 65%, followed by innovation culture (19%), collaboration (10%), open innovation (4%), and intellectual property (1.6%). Within each dimension, the majority of the document frequencies came from the top one or two keywords. For innovation culture, the two keywords "innov" (36%) and "create" (31%) accounted for 67% of the

[57] These frequencies are provided in Appendix C.

cumulative frequencies. In collaboration, "partner" (63%) and "collab" (24%) made up 87% of the frequencies. For open innovation, the sole term "community" yield 76% of the frequencies. The first three R&D-related keywords represented 81% of the frequencies, with "tech" (32%), "develop" (28%), and "lab" (20%). Lastly, for the intellectual property factor, "patent" accounted for 71% of the frequencies.

Web-based indicators are generated by aggregating the innovation-related signals by dimension. Therefore, the sum of all the document frequencies is aggregated into a discrete variable that quantifies the overall innovation signal (*CountAll*). The same process is repeated for each dimension resulting in an innovation culture signal (*CountInnov*), collaboration signal (*CountCollab*), OI signal (*CountOpen*), R&D signal (*CountRD*), and IP signal (*CountIP*).

6.4.3 Independent variables

To investigate which attributes of the firm have the strongest association with the different innovation-related signals, several categorical variables were created to describe different types of firm attributes. A summary of all the independent variables is available in Appendix E.

To measure evidence of innovation, a nominal index was generated by combining all the items related to innovation, including the number of innovations introduced in the last 3 years, the percentage of company revenue attributed to innovations, and the degree of innovation novelty (OECD & Eurostat, 2018)[58]. The resulting index ranges from 1 to 3, with 1 indicating no evidence of innovation (*dNoInnovEvidence*), 2 indicating some evidence of innovation (*dSomeInnovEvidence*), and 3 indicating strong evidence of innovation (*dStrongInnovEvidence*)[59].

The second group of independent variables relate to the time it takes to bring an innovation to the market (*TimetoMarket*). The time to market was segmented into three groups: firms with a longer

[58] Firms in the first group have not developed any new or improved products, processes, or services or did not declare any innovations that are new-to-the-world or did not have any revenue that can be attributed to product, process, or service innovation. Firms in the second category exhibit moderate regional or national innovation novelty or have either introduced one or two innovations considered to be new-to-the-world or generate less than 50% of their revenue from product, process, and service innovation. The third group of firms represents firms with a strong international leadership in terms of innovation novelty or that have introduced three or more innovations considered to be new-to-the-world or generate more than 50% of their revenue from product, process, and service innovation.

[59] Because our transformation of the variable cannot guaranty a uniform distribution within each of the three categories, the three corresponding dummy variables (presented in parentheses in the text) will be used in the regression analysis.

time to market (*dLongTMkt*), firms with an average time to market (*dAverageTMkt*) and firms with a shorter time to market(*dShortTMkt*)[60].

Two dummy variables were created to reflect the knowledge intensity of the firm: advanced knowledge services and tech firms are categorised as high knowledge intensity firms[61] (*dHighKI*), while low knowledge intensity firms correspond to low-tech industries and other services (*dLowKI*).

Sectors were categorised into 5 dummy variables: Digital Media & ICT (*dDigMediaICT*), Life science health (*dLifeSciHealth*), Materials & Manufacturing (*dMaterialMnfg*), Natural resources, Energy and Environment (*dEnerEnvironNat),* and Other (*dOther).* Natural resources and energy & environment were grouped together given the small number of firms in each sector, the relative similarity of their innovation-related characteristics.

As an alternative to the sectors above, an adaptation of the sectorial taxonomy presented in Section 6.3.2 (Baldwin & Rafiquzzaman, 1995; Castellacci, 2008; Pavitt, 1984) is operationalised with a combination of the sector, the organisation type and the number of employees[62]. First, all large and medium firms were manually processed and classified as either scale-based (*dScaleManuf*), science-based (*dScience*), KIBS (*dKIBS*), specialised-suppliers (*dSpecializedSupplier*) or supplier-dominated (*dSupplierDominated*). All small enterprises form Digital media and ICT[63] and the firms identified as advanced knowledge services were categorised as KIBS. All small firms that are identified as tech firms which are not operating in the Digital media and ICT sector are

[60] We generated the 20th percentile value and the 80th percentile of each sector. We then attribute the value 1 to the firms included in the interval that was the closest to the 80th percentile of their corresponding sector, i.e. with the longer time to market (dLongTMkt), and the value 3 to the firms that were the closest to the 20th percentile of their corresponding sector, i.e. with shorter time to market (dShortTMkt). The value 2 was given to the remaining firms, i.e. those with an average time to market (dAverageTMkt) Since the responses obtained in the surveys had very strong focal points (6, 12, 18, 24 months for instance), the resulting distribution is almost uniform with 30% 36% and 34%. For this reason, we only consider the two extremes (short and long time to market) compared to the average in the regression analysis.

[61] Both tech firms and advance knowledge services distinguish themselves from other firms by the involvement of Science, Technology, Engineering and Mathematics (STEM) employees to operate.

[62] It is important to note that the adaptation of the taxonomy presented in this study is not expected to be perfect. Rather, it is meant to be a useful framework for exploration and to provide a general indication of the relationships among the different signaling dimensions and sectorial attributes as described by this adaptation of the taxonomy. While further refinement may improve the robustness of the results, it is believed to be fit for purpose to test the propositions.

[63] While it is possible that the Digital media and ICT group includes also a few firms in telecommunications, the predominant composition of this group appears to consist of firms engaged in software development, internet-based businesses, and consulting services.

categorised as specialised-suppliers. Finally, all other small firms are listed as supplier-dominated. Because the number of scale-based and science-based firms is very low, we also combined them into another category called *dMassProd*.

Five more groups of variables were added as control variables. First, to generate firms' profiles based on their age (*Age*), we created three dummy variables for the very young early stage firms (*dEarlyStage*) that were 3 years old or younger at the *questionnaire* year, those aged between 4 and 8 years old (*dYoung*), and those at least 9 years old (*dMature*)[64].

Second, from the information about the number of employees, four size categories of the firm were generated. Following Statistic Canada's categorisation (Government of Canada, 2020), large firms employ 500 and more employees (*dLarge*), medium firms have between 100 and 499 employees (*dMedium*). Due to the fact that the sample consists of over 92% of small firms, of which approximately 2/3 had fewer than 10 employees, we distinguish very small firms with 1 to 10 employees (*dVerySmall*) from small firms (*dSmall*) with 11 to 99 employees.

Third, as the sample includes firms from Finland and Canada, we created one dummy variable for Canadian country (*dCanada*). Fourth, the investigation's time period is controlled by generating four dummy variables, each representing distinct periods: 2010-2011 (*d2010_2011*), 2012-2013 (*d2012_2013*), 2014 (*d2014*) [65], and 2015-2016 (*d2015_2016*). Finally, the last control variable is built to consider the variance in website size, structure, and exposure to the Wayback Machine crawler. The variable *SentenceRatio* corresponds to the total number of sentences by firm divided by the maximum value by size and by sector.

6.4.4 Model

To test the propositions, two regression models are estimated to explain the variation of signalling unit for each dimension (i.e., overall innovation, innovation culture, collaboration, OI, R&D and IP). Negative binomial (NB) and the zero-inflated negative binomial[66] version of these regression models are all legitimate models used with count data (Foltz et al., 2000; Greene, 2008, 2012;

[64] Each of the intervals represents approximately 33% of the age distribution of the company at the time of the questionnaire.
[65] The 2014 category represents about half of the respondents (51.8%).
[66] More specifically, these models fit an overdispersed count data with an excessive number of zeros.

Hausman et al., 1984; Long, 1997). The negative binomial model for N observations is presented in (Eq. 1) where i=1, ….N and X_i is a vector of explanatory variables.

$$Prob[Y = y_i \,|X_i\,] = \frac{exp\,(-\lambda_i)\lambda_i^{y_i}}{\Gamma(1+y_i)}$$

$$\mathrm{E[y_i|X_i]} = \lambda_\mathrm{i} = \exp(x_i x\,\beta), \quad \mathrm{y_i} = 0, 1, \ldots \mathrm{i}, \quad \mathrm{i} = 1, \ldots \mathrm{N} \tag{6.1}$$

and $V[y_i\,|X] = \lambda_i\,(1 + \,(\lambda_i\,/\,\nu_i))$

If $\nu_i = \alpha^{-1}$ for $\alpha > 0$ then $V[y_i\,|X] = \lambda_i\,(1 + (\alpha\lambda_i\,))$

When alpha is significantly different from zero, the data is over-dispersed (Var $[y_i\,|X_i] \neq \lambda_i$). The negative binomial (NB) regression is thus preferred to a Poisson model as it introduces latent heterogeneity in the Poisson model's conditional mean, with the error following a Gamma distribution (Greene, 2008; Long, 1997). The negative binomial model (Eq. 6.2) is specified as follows:

The zero-inflated is composed of a count model (Poisson or NB) that is shown in Eq. 6.2 and a logit model that models the excessive zeros as shown in Eq. 6.3.

$$\xi_i^\beta = x_i\beta + \,offset_i^\beta \tag{6.2}$$

$$\xi_i^\gamma = z_i\gamma + \,offset_i^\gamma \tag{6.3}$$

The expected value (λ_i) of the count model[67] is defined as Eq. 6.4.

$$\lambda_i = e^{\xi_i^\beta} \text{ and } p_i = \frac{1}{(1+\alpha\lambda_{i\,})} \text{ where } \;\; \nu_i = \frac{1}{\alpha} \tag{6.4}$$

Equation (6.5) and (6.6) show the simplified version of our models where *countSignals*$_i$ represents the number of signals retrieved from the corporate website of observations i (1 to 1594*)*. The first model allows to test propositions 3.1, 3.2, 3.3, 3.4 and 3.5 while propositions 3.6 and 3.7 are tested with model 2.

[67] For simplicity, we provide the equation of the zero-inflated binomial model

$$countSignals_{ij} = f\begin{pmatrix} dSomeInnovEvidence_i, dStrongInnovEvidence_i, \\ dShortTMkt_i, dLongTMkt_i, \\ dHighKI_i, dMaterialMnfg_i, \\ dLifeSciHealth_i, dDigMediaICT_i, dOther_i, \\ Control\ variables_i \end{pmatrix} \quad (6.5)$$

$$countSignals_{ij} = f\begin{pmatrix} dSomeInnovEvidence_i, dSomeInnovEvidence_i, \\ dShortTMkt_i, dLongTMkt_i, \\ dSupplierDominated_i, dKIBS_i, dMassProd_i, \\ Control\ variables_i \end{pmatrix} \quad (6.6)$$

where $countSignals_j$ is *CountAll, CountInnov, CountCollab, CountOpen, CountRD and CountIP*. Both equations share the following control variables for: age (*dYoung* and *dMature)*, firm size *(dVerySmall, dMedium* and *dLarge)*, country *(dCanada)*, period *(d2010_2011, d2012_2013* and *d2015_2016)*, as well as the size of the website *(SentenceRatio).*

These models were first tested estimating with NB regressions. In all cases, the NB models proved to be a better fit[68]. Zero-inflated negative binomial (ZINB)[69] regressions were then estimated. The logit part of the model, also called inflate, contains the variables that may contribute to explaining the excessive number of zeros. Contextual variables were used to explain the excess of zeros such as knowledge intensity, age, sector or sectorial category, size, country, period and the size of the website. To assess if ZINB is a better fit for the data the deviance statistics test was performed[70] where:

$$D = -2(LL_{ZINB} - LL_{NB}) \quad (6.7)$$

where D follows a χ^2 distribution at *p* degree of freedom where *p* is the number of coefficients in the inflate model.

[68] The alpha is always significantly different from zero (Hilbe, 2011; Long, 1997).

[69] The model is composed of two parts: a logit model that fits the excess of zeros distribution and a negative binomial model that fits the remainder of the count data distribution. A zero can be interpreted as an absence of signaling. This absence may be due because a firm did not signal any innovation-related terms at all (true negative) or because we did not detect it (false negative). As the web-based data collection process is dependent on the Wayback Machine which induces selection and non-selection biases, some pages were possibly not retrieved that may have contained innovation-related signaling. The zinb command from Stata 16 makes a model containing both a negative binomial and a logit model, to iterate over the likelihood (LL) of the function until it converges to a maximum value.

[70] Test performed with Fitstat (see Long & Freese, 2001)

For each regression, the test was significant in favour of the ZINB model. As a consequence, only the ZINB regressions are presented in the results section.

6.5 Results

6.5.1 Descriptive statistics

Appendix E presents the descriptive statistics of the sample of 1,594 observations. We have analysed the innovation-related signals from websites by counting the occurrences of these signals in sentences.

The overall innovation (*countAll*) measures the total number of innovation-related signals found in the sentences in corporate websites. The mean count is 50.040 implies that on average, each website counts about 50 signals (words or combination of words). The substantial standard deviation (92.466) and the wide range between the minimum and maximum values (0 to 1133) suggest that there is significant variation in innovation signalling across organisations. For instance, we find no innovation (the value is 0) signals in 9.16% of firms. The variable *countInnov* attempts to capture the presence of innovation culture within the websites. Understandably as we explore a subsample of signals, the mean count of innovation culture is lower (10.103), and the proportion of websites with no signal (27.85%) increases of cases. The same occurs for the other types of signals. The mean count of collaboration-related signals on the websites (*countCollab*) is 5.242 (41.59% do not signal at all). The mean presence of OI-related signals (*countOpen*) is much lower (2.065 with 71.52% of 0s). In contrast, the mean of R&D signals (*countRD*) has the highest mean count (22.181 with only 17.52% not signalling), suggesting that it is the most commonly emphasised topic across websites. Finally, the extent of IP signals (*countIP*) exhibits a very low mean count of 0.788 with 78.86% of firms not mentioning IP-related topics. When IP is mentioned, rarely, it does not dominate the overall message highlighted on corporate websites.

Turning to the independent variables of the model, we find that the respondents present different innovation profiles (23% exhibit no innovation evidence, 54% present some innovation evidence, and 23% are associated with strong innovation evidence). On average the firms in our sample need a little more than 20 months to bring an innovation to market (30% take a longer time to market, 36% an average time to market, and 34% take a shorter time) represent a diverse range of Canadian (48%) and Finnish (52%) firms, mostly small (60% have fewer than 10 employees, 33% host between 11 and 99 employees, 3.64% employ between 100 and 499 employees and 3.20% have

more than 500 employees), from multiple sectors such as Materials & Manufacturing (16%), Life science and health (11%), Natural resources (3.64%), Digital media & ICT (28%), Energy & Environment (3.76%), and other (38%). About 70% are in the high knowledge intensity level. The different age groups are relatively well distributed (33% are younger than 4 years old, 39% are aged between 4 and 8 years old, 28% are older than 8 years old). Finally, the time spread of the data gathered covers different time periods not uniformly distributed (7% in 2010-2011, 24% in 2012-2013, 52% in 2014, and 17% in 2015-2016).

Finally, the last column of Table E-5 shows the average overall count for each group of independent variables[71]. Some intriguing patterns seem to already emerge from this simple one-on-one assessment as firms with some innovation evidence (55.58) have higher average signals than those with no innovation evidence (40.01) and strong innovation evidence (46.97). Moreover, companies with a longer time-to-market duration (63.02 vs 41.41 for average and 47.87 for a shorter time to market) and higher knowledge levels (59.94 vs 27.40) have stronger average signals. A closer look at the sectors shows that the Life Science and Health sector emerges with stronger average signal (92.34), followed with Digital media and ICT (63.48) while Materials and Manufacturing lags behind with 30.56 signals on average. In terms of sectorial categories, Mass Production has the largest average value with 161.81, followed with KIBS (62.29), specialised-suppliers (47.47) and Supplier Dominated firms experiencing the weakest signals (25.15). Considering the ongoing digital transformation of manufacturing firms (Jones et al., 2021), it is plausible that these firms (both Materials and Manufacturing and specialised-suppliers) were investing less in their website during the study period or had fewer innovation-related requirements from external stakeholders.

Firm size and age seem to play a role in the average signal strength. Larger, more mature companies showcase higher signals than their smaller, younger counterparts. Finally, a geographical perspective highlights that Canadian firms have a much larger average signal than Finnish firms, which may be a reflection of selection and translation biases.

[71] An ANOVA analysis is also provided in Appendix E.

6.5.2 Regressions results

The full zero-inflated negative binomial regression results are presented in Appendix E in Table E-15 to Table E-17. The regression results show the variations in incidence rate ratio, in square brackets above the coefficient, compared with the omitted group of each category.

The regression results suggest that firms with some evidence of innovation (*dSomeInnovEvidence*) seem to trigger a stronger signal for overall innovation (regressions 1a, 1b), innovation culture (weakly significant in 2a and 2b), and R&D (5a, 5b), indicating that these signals may be useful in detecting moderately innovative firms. For instance, firms presenting some innovation evidence exhibit about 17% more signal[72] than firms with no innovation evidence. For R&D signals, the increase for the same two groups reaches 20%. Although the coefficient of *dStrongInnovEvidence* is never significant compared to *dNoInnovEvidence* (the omitted group), predicted marginal effects[73] (presented in Table 6-1[74]) highlight that firms that present some evidence of innovation signal more than the most innovative firms (strong evidence of innovation)[75], hence casting doubts on the methodology's ability to discriminate between moderately and highly innovative firms. This finding is consistent with the observation that the signal is related to innovation-active firms (Authors, forthcoming). It also contributes to explaining why Gök et al. (2015), with a similar methodology, found that R&D signalling does not reflect an objective R&D measure, as well as supports Héroux-Vaillancourt's (2020) demonstration that the signal aligns with broader conceptual measures. The results for collaboration, OI, and IP signalling were not conclusive with regards to the association of innovation results. As a consequence, proposition 3.1 is at best partially supported (overall, innovation culture, R&D-related signals), or not supported (collaboration, OI and IP-related signals).

[72] The increases in percentage are derived directly from the incidence rate ratio (IRR) in square brackets.

[73]The target group for *countAll*, *countInnov*, *countCollab*, *countOpen*, *countRD* are all the same with *dSomeInnovEvidence* =1, *dAverageTMkt* =1, *dHighKI* =1, *dDigMediaICT* =1, *dVerySmall*=1, *dYoung* = 1, *dCanada*=1 and *d2014* =1. For *countIP*, the group is *dSomeInnovEvidence* =1, *dLongTMkt* =1, *dHighKI* =1, *dLifeSciHealth* =1, *dSmall* =1, *dMature* = 1 *dCanada*=1 and d2014 =1 as it optimizes the number of nonzero observations.

[74] The complete results of the marginal effects of each variable for each regression are provided as online supplementary material.

[75] The marginal effect presented in the cell corresponding to Δ2 for *dStrongInnovEvidence* (3) is negative and significant implying that firms with some innovation evidence signal more (about 10 "signals") than those with strong innovation evidence, *ceteris paribus*.

Table 6-1. Marginal effects related to innovation evidence

Variable	countAll (reg. 1a)			countOpen (reg. 4a)		
	Signal	**Δ1**	**Δ2**	**Signal**	**Δ1**	**Δ2**
dNoInnovEvidence **1**	62.17 ***			1.36 **		
dSomeInnovEvidence **2**	72.82 ***	10.65 ***		1.50 ***	0.13	
dStrongInnovEvidence **3**	62.84 ***	0.67	-9.98 ***	1.36 ***	0.00	-0.14
	countInnov (reg. 2a)			**countRD (reg. 5a)**		
	Signal	**Δ1**	**Δ2**	**Signal**	**Δ1**	**Δ2**
dNoInnovEvidence **1**	11.97 ***			23.51 ***		
dSomeInnovEvidence **2**	14.06 ***	2.09 *		28.05 ***	4.54 ***	
dStrongInnovEvidence **3**	13.92 ***	1.95 *	-0.14	22.65 ***	-0.86	-5.39 ***
	countCollab (reg. 3a)			**countIP (reg. 6a)**		
	Signal	**Δ1**	**Δ2**	**Signal**	**Δ1**	**Δ2**
dNoInnovEvidence **1**	8.38 ***			1.89 ***		
dSomeInnovEvidence **2**	8.86 ***	0.48		1.82 ***	-0.06	
dStrongInnovEvidence **3**	6.76 ***	-1.62 *	-2.10 ***	1.95 ***	0.07	0.13

Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

The regression results suggest that time to market is not associated with overall innovation (1a, 1b), innovation culture (2a, 2b), collaboration (3a, 3b), OI (4a, 4b), R&D (5a, 5b) and IP (6a, 6b) signalling. Only when comparing the marginal effect (in Table 6-2) between the two extreme categories, short versus long time to market, do we note a significant difference for OI and IP signalling. Stronger OI signal appears to be associated with shorter, rather than longer, time to market, while firms with a longer time to market are associated with more IP-related terms. The latter further supports the notion that IP signalling is aligned with its corresponding activity, an observation consistent with the findings from Authors (forthcoming). Proposition 3.2 is therefore only supported for OI and IP-related signals.

Table 6-2. Marginal effects related to time to market

Variable	countAll (reg. 1a)			countOpen (reg. 4a)		
	Signal	**Δ1**	**Δ2**	**Signal**	**Δ1**	**Δ2**
dLongTMkt **1**	73.66 ***			1.25 ***		
dAverageTMkt **2**	72.82 ***	-0.84		1.50 ***	0.24	
dShortTMkt **3**	68.79 ***	-4.86	-4.03	1.92 ***	0.66 ***	0.42
	countInnov (reg. 2a)			**countRD (reg. 5a)**		
	Signal	**Δ1**	**Δ2**	**Signal**	**Δ1**	**Δ2**
dLongTMkt **1**	14.41 ***			28.04 ***		
dAverageTMkt **2**	14.06 ***	-0.35		28.05 ***	0.01	
dShortTMkt **3**	14.81 ***	0.40	0.75	25.34 ***	-2.70	-2.71
	countCollab (reg. 3a)			**countIP (reg. 6a)**		
	Signal	**Δ1**	**Δ2**	**Signal**	**Δ1**	**Δ2**
dLongTMkt **1**	7.23 ***			1.82 ***		
dAverageTMkt **2**	8.37 ***	1.15		1.41 ***	-0.41 *	
dShortTMkt **3**	8.12 ***	0.89	-0.25	1.16 ***	-0.66 ***	-0.25

Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

As anticipated, our analysis suggests that high knowledge intensity is strongly associated with all signalling dimensions. Compared to low knowledge intensity firms, high knowledge intensity exhibit 74% more innovation signals of any kind on their websites, 31% more innovation culture signals, 26% more collaboration signals, 77% more R&D signals, 94% more IP signals, but 29% less open innovation signals. For the latter, the low knowledge intensity firms who are sending the majority of signals on their corporate websites (the marginal effects are shown in Table 6-3). Hence, only for OI-related signals is the third proposition not supported, for all other types of signals, it is strongly supported.

Table 6-3. Marginal effects related to knowledge intensity

Variable		countAll (reg. 1a)		countInnov (reg. 2a)		countCollab (reg. 3a)	
		Signal	**Δ1**	**Signal**	**Δ1**	**Signal**	**Δ1**
dLowKI	1	41.72 ***		10.57 ***		4.64 **	
dHighKI	2	72.82 ***	31.10 ***	14.06 ***	3.49 ***	8.37 ***	3.73 **
		countOpen (reg. 4a)		**countRD (reg. 5a)**		**countIP (reg. 6a)**	
		Signal	**Δ1**	**Signal**	**Δ1**	**Signal**	**Δ1**
dLowKI	1	2.02 **		15.83 ***		0.95 ***	
dHighKI	2	1.50 ***	-0.52 **	28.05 ***	12.21 ***	1.82 ***	0.88 ***

Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Turning now to the sectors, it was anticipated that firms operating within the Digital media and ICT as well as Life science and health sectors would exhibit a stronger correlation with innovation-related signals, the results in Table E-15 and Table E-16 present the results compared to the omitted category (*dEnerEnvironNat*). Our research results are aligned with our expectation, especially for the former (*dDigMediaICT*) in terms signals of innovation (239% more signal), innovation culture (321%), collaboration (336%), OI (581%), R&D (262%), and IP (327%). Understandably, when the signal is relatively small in a given sector, as it is in the Energy, Environment and natural resources sector, even an increase in a small number of words is associated with a large proportional increase (in terms or IRR and percentage). The marginal effects for this particular example (in column Δ3 and line *dEnergyEnv.Nat* in Table 6-4) present the predicted number of signals of each type that the Energy, Environment and Natural resources sector exhibits compared with the Digital Media and ICT sector, *ceteris paribus*.

The signals related to open innovation (4a, 4b) are principally associated with the Other sector, hinting that the terms selected may be too general to properly discriminate these dimensions within an innovation-related context. Furthermore, the granularity and precision of the sector definitions limit our analysis, and the impact of *dOther*, which represents a diverse assortment of sectors, was underestimated. Furthermore, Materials and manufacturing firms exhibit a weaker association (i.e., not statistically significantly different from the Energy, Environment and natural resources sector) with the different signals than expected, except for OI-related signals. Materials and manufacturing firms were also on average those with the least number of sentences, and therefore

signals, measured compared with all other sectors[76]. Proposition 3.4 is therefore supported for innovation culture and R&D-related signals and partially supported for the other categories.

In the "b" models of Table E-15 and Table E-16, interactions between knowledge intensity (KI) and sectors were added to the regression analysis as some sectors may be associated with a specific level of knowledge intensity. Nearly all firms displayed stronger signals than the omitted group (*dEnerEnvironNat x dLowKI*). For instance, high knowledge intensity firms in Life science and health yield almost 5 times more overall signals than low knowledge intensive firms in energy, environment and natural resources. As expected, our findings suggest that firms operating within the same sector but with higher levels of knowledge intensity tend to exhibit a greater overall innovation signal compared to those with lower levels of knowledge intensity. The diagonal in Table 6-5 (for countAll and Δ1, Δ3, Δ5, Δ7 and Δ9 columns) clearly highlights this additional impact from higher knowledge intensity, for overall and R&D-related signals (with the exception of materials and manufacturing for the latter). However, the systematic impact of knowledge intensity within each sector is not replicated across all the subcategories of signals. Only the higher knowledge firms in Other sectors are associated with stronger innovation culture-related signals while higher knowledge firms in Other and Digital Media and ICT are associated with a stronger collaboration-related signal. Stronger IP-related signals are observed for Life science and Health and Other sectors. Finally, OI-related signals seem to have a stronger association with lower knowledge intensive firms in manufacturing and Other sectors. It appears the knowledge intensity interaction is especially useful to discriminate the firms among themselves in the Other sector as it is a heterogeneous group. Proposition 3.5 is thus only supported by the overall innovation signal.

[76] These statistics are available in Appendix E.

Table 6-4. Marginal effects related to sector

Variable		countAll (reg. 1a)					countOpen (reg. 4a)				
		Signal	**Δ1**	**Δ2**	**Δ3**	**Δ4**	**Signal**	**Δ1**	**Δ2**	**Δ3**	**Δ4**
dMaterialMnfg	**1**	24.71 ***					0.38 **				
dLifeSciHealth	**2**	54.2 ***	29.49 ***				0.51 ***	0.13			
dDigMediaICT	**3**	72.82 ***	48.11 ***	18.62 ***			1.50 ***	1.11 **	0.98 ***		
dEnerEnvironNat	**4**	21.39 ***	-3.32	-32.81 ***	-51.43 ***		0.21 ***	-0.18 ***	-0.31 ***	-1.29 ***	
dOther	**5**	50.22 ***	25.51 ***	-3.98	-22.60 ***	28.83 ***	1.30 ***	0.91 **	0.78 ***	-0.20	1.09 ***
		countInnov (reg. 2a)					**countRD (reg. 5a)**				
		Signal	**Δ1**	**Δ2**	**Δ3**	**Δ4**	**Signal**	**Δ1**	**Δ2**	**Δ3**	**Δ4**
dMaterialMnfg	**1**	3.85 ***					10.33 ***				
dLifeSciHealth	**2**	8.37 ***	4.52 ***				28.25 ***	17.92 ***			
dDigMediaICT	**3**	14.06 ***	10.21 ***	5.69 ***			28.05 ***	17.72 ***	-0.21		
dEnerEnvironNat	**4**	3.28 ***	-0.57	-5.09 ***	-10.78 ***		10.70 ***	0.37	-17.55 ***	-17.35 ***	
dOther	**5**	9.45 ***	5.60 ***	1.08	-4.61 ***	6.17 ***	20.87 ***	10.54 ***	-7.38 ***	-7.17 ***	10.17 ***
		countCollab (reg. 3a)					**countIP (reg. 6a)**				
		Signal	**Δ1**	**Δ2**	**Δ3**	**Δ4**	**Signal**	**Δ1**	**Δ2**	**Δ3**	**Δ4**
dMaterialMnfg	**1**	1.96 **					0.88 ***				
dLifeSciHealth	**2**	4.86 ***	2.90 *				1.82 ***	0.95 **			
dDigMediaICT	**3**	8.37 ***	6.41 ***	3.51 ***			1.20	0.33 ***	-0.62		
dEnerEnvironNat	**4**	1.66 ***	-0.30 *	-3.20 ***	-6.71 ***		0.64 ***	-0.23	-1.18 ***	-0.56 ***	
dOther	**5**	5.63 ***	3.67 ***	0.77 **	-2.74 ***	3.97 ***	1.54 **	0.66 ***	-0.29	0.33 *	0.89 ***

Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table 6-5. Marginal effects related to sector x Knowledge intensity

		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
countAll (reg. 1b)											
dMaterialMnfg x *dLowKI*	**1**	16.88 ***									
dMaterialMnfg x *dHighKI*	**2**	22.88 ***	6.00 **								
dLifeSciHealth x *dLowKI*	**3**	24.20 ***	7.32 *	1.32							
dLifeSciHealth x *dHighKI*	**4**	55.64 ***	38.75 ***	32.76 ***	31.44 ***						
dDigMediaICT x *dLowKI*	**5**	46.13 ***	29.25 ***	23.25 ***	21.93 ***	-9.51 **					
dDigMediaICT x *dHighKI*	**6**	68.25 ***	51.37 ***	45.38 ***	44.06 ***	12.62 **	22.13 ***				
dEnerEnvironNat x *dLowhKI*	**7**	11.04	-5.84 ***	-11.84 ***	-13.16 ***	-44.59 ***	-35.09 ***	-57.21 ***			
dEnerEnvironNat x *dHighKI*	**8**	21.40 ***	4.52	-1.48	-2.80	-34.23 ***	-24.73 ***	-46.85 ***	10.36 ***		
dOther x *dLowKI*	**9**	26.08 ***	9.20 ***	3.20	1.88	-29.56 ***	-20.05 ***	-42.18 ***	15.04 ***	4.68	
dOther x *dHighKI*	**10**	50.74 ***	33.86 ***	27.86 ***	26.54 ***	-4.90	4.61	-17.52 ***	39.70 ***	29.34 ***	24.66 ***

Table 6-5. Marginal effects related to sector x Knowledge intensity (continued)

		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
countInnov (reg. 2b)											
dMaterialMnfg x *dLowKI*	**1**	3.09 ***									
dMaterialMnfg x *dHighKI*	**2**	3.78 **	0.69								
dLifeSciHealth x *dLowKI*	**3**	6.54 ***	3.45 *	2.76							
dLifeSciHealth x *dHighKI*	**4**	8.32 ***	5.23 ***	4.55 ***	1.78						
dDigMediaICT x *dLowKI*	**5**	11.71 ***	8.62 ***	7.93	5.17	3.39 ***					
dDigMediaICT x *dHighKI*	**6**	13.92 ***	10.83 ***	10.14 ***	7.38 *	5.60 ***	2.21				
dEnerEnvironNat x *dLowhKI*	**7**	2.43	-0.66	-1.35	-4.11 **	-5.89 ***	-9.28 ***	-11.49 ***			
dEnerEnvironNat x *dHighKI*	**8**	3.28 ***	0.19	-0.50	-3.27	-5.05 ***	-8.44 ***	-10.65 ***	0.84		
dOther x *dLowKI*	**9**	6.41 ***	3.32 ***	2.63 ***	-0.14	-1.92	-5.31	-7.52 ***	3.97 ***	3.13 ***	
dOther x *dHighKI*	**10**	10.02 ***	6.93 ***	6.24 ***	3.48	1.70	-1.69	-3.90 ***	7.59 ***	6.75 ***	3.62 ***

Table 6-5. Marginal effects related to sector x Knowledge intensity (continued)

		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
countCollab (reg. 3b)											
dMaterialMnfg x dLowKI	**1**	1.35									
dMaterialMnfg x dHighKI	**2**	1.86 **	0.51								
dLifeSciHealth x dLowKI	**3**	3.72 **	2.38	1.86							
dLifeSciHealth x dHighKI	**4**	4.80 ***	3.45	2.94 *	1.07						
dDigMediaICT x dLowKI	**5**	4.25 **	2.91 *	2.39 ***	0.53	-0.54 **					
dDigMediaICT x dHighKI	**6**	8.86 ***	7.52 ***	7.00 ***	5.14 **	4.07 ***	4.61 **				
dEnerEnvironNat x dLowhKI	**7**	0.90 *	-0.45 *	-0.96 *	-2.82 *	-3.89 ***	-3.35 ***	-7.96 ***			
dEnerEnvironNat x dHighKI	**8**	1.71 ***	0.36	-0.16	-2.02	-3.09 ***	-2.55 ***	-7.16 ***	0.80		
dOther x dLowKI	**9**	3.19 **	1.85	1.33 ***	-0.53	-1.60	-1.06 *	-5.67 ***	2.29 ***	1.49 ***	
dOther x dHighKI	**10**	5.92 ***	4.57 *	4.05 ***	2.19	1.12 *	1.66	-2.95 ***	5.01 ***	4.21 ***	2.72 *

Table 6-5. Marginal effects related to sector x Knowledge intensity (continued)

		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
countOpen (reg. 4b)											
dMaterialMnfg x dLowKI	**1**	1.18									
dMaterialMnfg x dHighKI	**2**	0.27 **	-0.91 ***								
dLifeSciHealth x dLowKI	**3**	0.73 *	-0.45 **	0.46							
dLifeSciHealth x dHighKI	**4**	0.50 ***	-0.67 ***	0.23	-0.22						
dDigMediaICT x dLowKI	**5**	1.45 **	0.28 *	1.19 **	0.73	0.95 **					
dDigMediaICT x dHighKI	**6**	1.52 ***	0.34 *	1.25 ***	0.79 *	1.01 ***	0.06				
dEnerEnvironNat x dLowhKI	**7**	0.20 **	-0.98 ***	-0.07 *	-0.53 ***	-0.30 **	-1.25 ***	-1.32 ***			
dEnerEnvironNat x dHighKI	**8**	0.22 ***	-0.96 ***	-0.05 **	-0.51 ***	-0.28 ***	-1.23 ***	-1.30 ***	0.02		
dOther x dLowKI	**9**	1.59 ***	0.42	1.33 ***	0.87 **	1.09 ***	0.14 ***	0.08 *	1.39 ***	1.37 ***	
dOther x dHighKI	**10**	1.20 ***	0.02 *	0.93 ***	0.47	0.70 ***	-0.25	-0.32	1.00 ***	0.98 ***	-0.39 *

Table 6-5. Marginal effects related to sector x Knowledge intensity (continued)

		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
countRD (reg. 5b)											
dMaterialMnfg x dLowKI	**1**	8.03 ***									
dMaterialMnfg x dHighKI	**2**	9.48 ***	1.45								
dLifeSciHealth x dLowKI	**3**	9.38 ***	1.34	-0.10							
dLifeSciHealth x dHighKI	**4**	31.00 ***	22.96 ***	21.52 ***	21.62 ***						
dDigMediaICT x dLowKI	**5**	17.96 ***	9.93 ***	8.48 ***	8.58 ***	-13.03 ***					
dDigMediaICT x dHighKI	**6**	26.83 ***	18.80 ***	17.35 ***	17.45 ***	-4.16	8.87 ***				
dEnerEnvironNat x dLowhKI	**7**	4.89 ***	-3.15 **	-4.59 ***	-4.49 ***	-26.11 ***	-13.07 ***	-21.94 ***			
dEnerEnvironNat x dHighKI	**8**	10.53 ***	2.49	1.05	1.15	-20.47 ***	-7.43 ***	-16.30 ***	5.64 ***		
dOther x dLowKI	**9**	10.74 ***	2.71 *	1.26	1.37	-20.25 ***	-7.22 ***	-16.09 ***	5.86 ***	0.22	
dOther x dHighKI	**10**	21.38 ***	13.34 ***	11.90 ***	12.00 ***	-9.62 ***	3.42 *	-5.45 ***	16.49 ***	10.85 ***	10.63 ***

Table 6-5. Marginal effects related to sector x Knowledge intensity (continued)

		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
countIP (reg. 6b)											
dMaterialMnfg x dLowKI	**1**	0.41 *									
dMaterialMnfg x dHighKI	**2**	0.80 ***	0.39								
dLifeSciHealth x dLowKI	**3**	0.43 ***	0.02	-0.37							
dLifeSciHealth x dHighKI	**4**	1.85 ***	1.44 ***	1.05 ***	1.42 **						
dDigMediaICT x dLowKI	**5**	0.90	0.49 **	0.10 **	0.48 *	-0.95					
dDigMediaICT x dHighKI	**6**	1.09	0.68 ***	0.29 ***	0.66 **	-0.76	0.18				
dEnerEnvironNat x dLowhKI	**7**	0.34 *	-0.07	-0.46 **	-0.09	-1.51 ***	-0.57 ***	-0.75 ***			
dEnerEnvironNat x dHighKI	**8**	0.58 ***	0.17	-0.22	0.16	-1.27 ***	-0.32 ***	-0.50 ***	0.25		
dOther x dLowKI	**9**	0.70 ***	0.29	-0.10	0.27	-1.15 ***	-0.20 **	-0.39 ***	0.36 *	0.12	
dOther x dHighKI	**10**	1.47 ***	1.06 **	0.67 ***	1.04 *	-0.38	0.57	0.38	1.13 ***	0.89 ***	0.77 ***

Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

With the sectorial categories, our findings indicate that firms operating within advanced knowledge-based industries or mass production exhibited a stronger overall innovation signal (65% more – reg. 1c), as well as in all other signalling categories (51% to 231% more – reg. 2c to 6c) when compared to those associated with specialised supplier industries (see Table E-17). The predicted marginal effects in Table 6-6 provide a more nuanced picture. For instance, compared to supplier dominated firms, KIBS firms do not present statistically more OI-related signals, even though both present statistically significant results in the regression analysis (reg. 4c), i.e., compared to the omitted sector. That is the only exception for KIBS; for all other categories of signals, these firms publish stronger messages than supplier dominated firms (see all Δ1 columns in Table 6-6). Innovation culture and collaboration-related signals do not show meaningful differences between specialised-suppliers and supplier-dominated firms, and for OI-related signals, the relationship is negative, the former signalling more than the latter. Propositions 3.6 and 3.7 are therefore those for which we found the best support, only for IP, is proposition 3.7 not supported.

The regression analysis revealed some unexpected distinctions between a firm's signal strength and its size, which may be attributed to the limited number of medium and large firms in our sample, potentially skewing the regression results. Nevertheless, the regression findings suggest that medium-sized firms exhibit a weaker signal compared to other firms in terms of overall innovation, innovation culture, and R&D dimensions. When examining medium size firms, only 17/58 were in either Digital media and ICT as well as Life science and health, suggesting a strong sectorial effect. Conversely, small firms displayed a stronger signal for collaboration and OI, whereas very small firms surprisingly exhibited a stronger IP signal compared to small and large firms. Regarding the age of the firm, our analysis suggests that overall, there are no meaningful distinctions, with a few notable exceptions. Early-stage firms appear to emit a weaker R&D signal, while mature firms exhibit a stronger IP-related signal compared to other firms.

Table 6-6. Marginal effects related to sectorial category

Variable		countAll (reg. 1c)				countOpen (reg. 4c)			
		Signal	**Δ1**	**Δ2**	**Δ3**	**Signal**	**Δ1**	**Δ2**	**Δ3**
dSupplierDominated	**1**	25.00 ***				1.53 ***			
dSpecializedSupplier	**2**	41.74 ***	16.75 ***			0.57 ***	-0.96 ***		
dKIBS	**3**	68.80 ***	43.81 ***	27.06 ***		2.22 ***	0.69	1.65 ***	
dMassProd	**4**	45.42 ***	20.42 ***	3.68	-23.38 **	0.64 ***	-0.89 ***	0.07	-1.58 ***
		countInnov (reg. 2c)				**countRD (reg. 5c)**			
		Signal	**Δ1**	**Δ2**	**Δ3**	**Signal**	**Δ1**	**Δ2**	**Δ3**
dSupplierDominated	**1**	5.68 ***				12.87 ***			
dSpecializedSupplier	**2**	6.71 ***	1.03			22.44 ***	9.57 ***		
dKIBS	**3**	14.43 ***	8.75 ***	7.72 ***		34.25 ***	21.38 ***	11.80 ***	
dMassProd	**4**	9.24 ***	3.56 *	2.53 *	-5.19 **	23.86 ***	10.99 ***	1.42	-10.38 *
		countCollab (reg. 3c)				**countIP (reg. 6c)**			
		Signal	**Δ1**	**Δ2**	**Δ3**	**Signal**	**Δ1**	**Δ2**	**Δ3**
dSupplierDominated	**1**	3.21 ***				0.43			
dSpecializedSupplier	**2**	4.29 ***	1.09			1.44 ***	1.01 **		
dKIBS	**3**	9.34 ***	6.13 ***	5.04 ***		0.84	0.41 ***	-0.59	
dMassProd	**4**	3.68 ***	0.47 **	-0.62 *	-5.66 ***	0.97	0.54	-0.47 ***	0.12 ***

Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Our analysis indicates that Canadian firms compared to Finnish firms exhibit a stronger signal for all the dimensions, except for OI which do not seem to be different. However, we obtained considerably more sentences for Canadian firms (about 31% more)[77]. It is possible that this difference is due to a methodological bias resulting from either the Wayback Machine exposure or the translation process, which may have led to false negative conversions.

In light of the results discussed above, a number of the propositions made in the theoretical framework are validated. Table 6-7 summarises the validation matrix of the propositions.

Table 6-7. All results summary

Propositions	**Overall**	**Innovation culture (a)**	**Collaboration (b)**	**OI (c)**	**R&D (d)**	**IP (e)**
Innovation evidence						
Proposition 3.1	Partially	Partially	No	No	Partially	No
Time to market						
Proposition 3.2	No	No	No	Yes	No	Yes
Knowledge intensity						
Proposition 3.3	Yes	Yes	Yes	No	Yes	Yes
Sector						
Proposition 3.4	Partially	Yes	Partially	Partially	Yes	Partially
Proposition 3.5	Yes	No	Partially	No	Partially	Partially
Sectorial category						
Proposition 3.6	Yes	Partially	Partially	Partially	Yes	Yes
Proposition 3.7	Yes	Yes	Yes	Yes	Yes	No
Control effects						
Size	Yes	No	Yes	No	No	No
Age	No	No	No	No	Yes	Yes
Country	Yes	Yes	Yes	No	Yes	Yes

[77] These statistics are available in Appendix E.

6.6 Discussion

Regarding the overall innovation signal strength, the results align reasonably well with expectations derived from the literature. Specifically, firms with a stronger overall innovation signal are associated with some level of innovation results, are operating at a higher level of knowledge intensity, and are working in sectors characterised by either a technological regime or a sectorial category that would predict higher innovativeness. Furthermore, our findings revealed a consistent trend in the signal strength from firms operating in Digital media and ICT or identified as KIBS. These firms demonstrated a stronger signal in almost every dimension. This may indicate the greater comfort these industries have with online communication. It is also no accident that these firms are those for which we obtained the most sentences. In contrast, specialised suppliers and Materials and manufacturing firms with higher knowledge intensity displayed consistently a weaker signal than anticipated. This could be attributed to factors such as a lack of familiarity with web-based technologies or a lower need for external resources. It is also possible they do use their websites for other needs that were not captured. Additionally, as these firms may be more engaged in manufacturing process innovation, it is possible that this type of innovation is less visible on their corporate websites. Finally, the time to market, the size and the age of the firms do not seem to have a significant impact on the overall innovation signal strength.

The R&D signal corresponds relatively well with what may be expected based on the literature, which is not surprising given that it is the biggest contributor to the overall innovation signal. Furthermore, it seems the youngest firms do not signal as much R&D-related terms which may correspond well to where the firms are in their innovation cycle, as most early-stage firms are primarily focused on launching their operations around a novel value proposition, while more established firms need to engage more actively in R&D activities, to bring to market new products and services (Christensen, 1997). Furthermore, the relative success of this web-based indicator suggests there is a strong semantic alignment between the lexical field used in academia and industry to relate with the R&D dimension. Moreover, the language used to describe the R&D dimension seems to be mostly consistent across different sectors.

The propositions related with the strength of the IP signal were also providing interesting insights. As the indicator is pulled by the term “patent”, it was interesting to see that the more established firms with a longer time to market were more associated with stronger IP signal which suggests a

strong connection between the term and reality. However, the differences in sectorial characteristics did not lead to conclusive results. Our findings suggest that KIBS industries tend to exhibit a stronger association with the IP indicator than specialised providers. This could be due to specialised providers in our sample being less engaged in patenting activities than expected or signalling less IP-related activities than their actual activities compared to KIBS firms. Further research is needed to better understand the relationship between sectorial characteristics and IP signalling.

Both innovation culture and collaboration have been corresponding to expectations in about half of the cases. Attempting to measure something as elusive as a culture is not a trivial task, and reducing something as complex into keywords comes with foreseeable limitations. However, it was interesting to see that the innovation culture dimension measured, led by the terms “innov” and followed closely by “create”, are positively associated with at least some innovation evidence. This suggests that beyond the buzzword, there is something tangible behind the use of the “innovation” lexical field. Hence, this suggests that the signal may potentially contribute to differentiating innovative firms from non-innovative firms as defined by the Oslo Manual (OECD & Eurostat, 2018).

However, it should be noted that the keyword method used in our study may not effectively differentiate between internal and external collaboration, which could impact the lack of association between the signal and the evidence of innovation observed. Moreover, as the keyword “partner” used in our study is not exclusive to the collaboration dimension, it is likely that it may be used in other contexts than innovation.

Firms with a shorter time to market appear to signal more open innovation (OI)-related terms, with this signal being led by the keyword ‘community’. Within the OI framework, firms can build and sustain relationships with external stakeholders by engaging with communities, which can provide valuable feedback, insights, and expertise to identify opportunities for innovation and improve existing products or services, ultimately resulting in lower time to market (West & Bogers, 2014; West & Lakhani, 2008). The association between time to market and OI-related signalling may shed light on the role of community engagement in the innovation process. It is also possible that the term 'community' is used more frequently by firms that are engaged in social innovation, and may be relatively quicker to implement (Toivonen, 2016).

However, most of the propositions related to OI signalling have been inconclusive for the most part. Despite having the largest number of keywords aimed at capturing an OI signal, it was overshadowed by the term ‘community’ by a wide margin. This suggests that there may be a considerable gap between the terms commonly used in the research community and those used in industry. Consequently, assessing the OI dimension using keyword-related methods may prove to be particularly challenging.

6.7 Conclusion

This paper explores how different innovation-related signals from corporate websites correspond to what would be expected considering the established literature related to innovation management with a sample of 1,594 firms. The overall strength of an innovation signal is assessed from the occurrences of keywords related to innovation culture, collaboration, OI, R&D and IP using Document frequency analysis. Doing so, it builds on the related literature that aims to understand what can be expected from corporate websites as a potential data source for assessing firms’ innovation characteristics (Gök et al., 2015; Héroux-Vaillancourt et al., 2020; Authors, in press). This paper contributes to this emerging literature by assessing how innovation evidence, time to market, knowledge intensity, sector and other sectorial categories, size, age and country influence innovation signalling.

While this study has shed light on the differences in strength of different innovation-related signals across sectors, there is still much to be explored in terms of the nuances and underlying reasons for these differences. Further research is necessary to delve deeper into the relationship between sectors and innovation-related signalling patterns, and to identify the factors that may be contributing to these patterns. This could involve examining the role of different types of innovation in different sectors, as well as investigating how other factors such as organisational culture, market conditions, and regulatory environment may be influencing innovation-related signalling on corporate websites.

The acquisition of online data using the Wayback Machine service has inherent limitations that may result in non-selection and selection biases. Additionally, the use of an online translation service is not perfect, and potential mistranslation may lead to a noisy web-based signal, particularly if Finnish words were not accurately translated. Despite these challenges, the keyword frequency methodology employed in this study may mitigate some of these concerns, as the quality

of translation is not a critical factor, unlike in bottom-up techniques such as topic modelling. Furthermore, as language evolves over time, some terms may not have been as prevalent during the survey data collection period as they are today. To address these challenges, a periodic web scraping of selected firms paired with questionnaire-based surveys, could be implemented. This would enable panel data collection and longitudinal studies, allowing for a more thorough validation of our findings and an improved understanding of web-based signals and their implications for innovation management studies.

Moreover this research is showing the limitation of a keyword frequency method for capturing a signal related to an innovation dimension, especially from dimensions that are either less tangible or that do not share the same language used in both academia and the industry. As a result, there is the possibility that some critical keywords have been inadvertently excluded, as the industry may use different terminology than academia to describe certain activities related to a dimension such as OI. Furthermore, the signal indicator generated through document keyword frequency fails to consider the context in which words are used, which could result in erroneous positive or negative outcomes, such as it may have been the case with collaboration. Using more sophisticated bottom-up approaches, such as topic modelling with Latent Dirichlet Allocation, supervised machine learning models, or even cutting-edge LLM-based technologies, could potentially enhance our capacity to precisely gauge innovation signals from websites.

Nevertheless, this exploratory paper presents a promising methodology for deriving innovation-related characteristics of firms at scale. By examining innovation-related signalling emitted by innovation-active or moderately innovative firms seeking external support, we can gain insights into the needs of firms to foster innovation. Moving forward, further research in this area could lead to a better understanding of how to use web-based sources of data to inform innovation management practices. Overall, this paper highlights the potential for using web-based data to enrich innovation management quantitative research and invite the research community to continue the exploration and validation of new constructs in this area.

6.8 Author contributions

Mikaël Héroux-Vaillancourt: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, resources, software, validation, visualisation, writing—original draft, writing—review and editing. **Catherine Beaudry**: project administration,

resources, funding acquisition, supervision, validation, visualisation, writing –review & editing. **Davide Pulizzotto**: methodology, software, validation. **Margaret Dalziel**: data curation, resources, validation, writing - review & editing.

CHAPTER 7 GENERAL DISCUSSION

Given that the research presented in this thesis is exploratory in nature, we have tested a series of propositions, which have been grouped into three distinct categories to attempt to answer our research objectives. The first set of propositions assessed how web-based measures correspond with several questionnaire-based measures using a sample of Canadian high-tech firms, mostly SMEs. The second group of propositions attempt to understand whether web-based measures are associated with evidence of innovation results and/or with the investment received, an indirect measure for evidence of innovation activities. The remaining propositions tested in Paper 3 assessed how the web-based measures compared with what would be expected from the literature if they are reflecting real evidence of activities corresponding to the factors they represent. The propositions assess the association between the web-based indicators and the innovation evidence level, the time to market, the knowledge intensity, the sector and other sectorial categories. These two latest groups of propositions are tested with a large sample of predominantly small Canadian and Finnish technological firms served by innovation intermediaries to improve their innovation capabilities. Furthermore, all the firms involved in our three papers are interesting from a policy-making perspective. The first group develops cutting-edge emerging technologies, which are important to properly support to ensure their success. The second group represents a wide range of firms, mostly SMEs, that are seeking to improve their innovation outcomes. What follows discusses our findings and highlights each paper's main contributions.

7.1 Web-based indicators vs questionnaire-based indicators on innovation factors

The initial intuition from the literature was to examine whether these web-based indicators could directly correspond to objective proxy measurements that would reflect established innovation-based measurements, such as the number of publications, the number of patents, R&D expenditure, and grants. However, this has not been the case (Gök et al., 2015). The first paper presented aimed to validate and complement the findings reported by Gök et al. (2015) with proposition 1.1, by assessing the correlation results of various innovation activities between web-based indicators and questionnaire-based direct measures for these activities. Secondly, with proposition 1.2 we explored an alternative correspondence by adopting an innovation culture perspective to assess whether these web-based indicators can be regarded as artefacts reflecting the level of importance

attributed to innovation activities by the firms. Several activities related to R&D, collaboration, external financing, and IP were obtained using a classic questionnaire-based method on a sample of 79 high-tech firms. These activities were then compared with web-based indicators constructed through a classic keyword frequency analysis, standardised by the total number of words. Initially, both external financing and collaboration factors did not exhibit significant correlations between web-based and questionnaire-based measures. This already suggested that proposition 1.1 might not be corroborated, at least for these factors. Since both R&D and IP activities were assessed with different questions, we investigated for each factor if one of these activities would be significantly more associated with the corresponding web-based indicator. To do so, we introduced six MTMM matrices to assess the construct validity to determine if there is empirical support to associate a web-based indicator with a specific activity rather than another. This test was inconclusive, ironically, because we identified three strong valid constructs, which means that we were unable to discriminate between activities.

After all, this intuition that the same words used from different actors in different contexts referred to a meaning that would point to the same measurable quantities in the real world may have been, respectfully, philosophically naive. The French philosopher Jacques Derrida sought to challenge the idea that words have fixed, inherent meanings, and proposed instead the meaning of words is always shifting and dependent on the context in which they are used. He argued that this instability of meaning creates a gap or difference between the words that are spoken or written, and the concepts they are meant to represent (Derrida, 2016). For example, even for words that signify concepts that have a physical existence, such as “tree” or “mountain”, although there is no confusion about what is a tree or a mountain, no one has the exact same tree or mountain in their own mind. This gap is even greater with words that refer to concepts that rely on other words to exist, such as innovation or collaboration or any other “imagined reality” as coined by Harari (2014). As mentioned previously, a biotech SME or a big pharmaceutical enterprise are more than likely to have very different objective measures when it comes to their activities, while using similar innovation-related language on their websites.

The observation that many activities related to a factor measured by the questionnaire correlated with the corresponding web-based indicator implies that these activities are implicitly considered by the indicator. As a keyword frequency method does not take any context into account, this suggests that the web-based indicators were representing many different activities. This further

hints that proposition 1.2 is more likely to be representative of reality. To support this proposition, all the items corresponding to R&D and IP were combined into formative indices, which would reflect an overall level of importance for these factors. This time, the MTMM matrix provided convergent and discriminant construct validity, and the results were robust as the same conclusion was obtained by performing a CFA.

This study presents the first empirical evidence that web-based indicators based on a keyword method can properly assess broader conceptual associations between a firm and factors related to innovation, rather than specific measurable quantities. These associations may reveal to us, as external observers, artefacts of a corporate culture supporting these factors.

Finally, construct validation has seldom been performed in this field. Thus, the paper also contributes by demonstrating how to use the methodology presented to explore and validate new data and methods.

7.2 Web-based indicators as evidence of innovation

The second paper starts from the same premise as proposition 1.2, which is that the keyword methodology captures broader concept rather than specific activities. Moreover, the research design was heavily inspired from both the semiotics and the signalling theories literature. The index sign concept from semiotics led to the initial intuition behind the idea that web-based indicators may be seen as evidence of innovation. The semiotics literature uses the example of smoke that points to a fire as an index sign. Further related readings introduced the concept of an evidential paradigm, which provided the example of cave paintings – cultural artefacts that offer evidence for characteristics of Palaeolithic societies.

Drawing on these concepts, words related to innovation or factors related to innovation can be interpreted as the "smoke" that should lead to an "innovation fire". However, as there are innovation results and innovation activities, we turned towards signalling theory as inspiration to distinguish between these two types of “fires”. Typically, researchers in the field of signalling theory address information asymmetry problems by interpreting the actions of a trusted source that takes on the signalling cost. We identified an investor as a third party that has more information than we do and make a decision based on that information.

As our sample is constituted of smaller firms that are collaborating with innovation intermediaries seeking to improve their innovation capabilities, it provides our analysis with the data to assess whether the innovation-related language is associated with the investment received, thereby validating to some extent that the language is representative of reality as investors are much more likely to be associated with tangible activities rather than "empty" words. Besides, the relationship between investment and innovation is well established in the literature (Aghion et al., 2004; Kerr & Nanda, 2015; Wellalage & Fernandez, 2019).

This led to an innovative research design that uses both innovation results from questionnaire data as objective evidence of innovation and the investment received as an indirect measure of evidence of innovation activities into a simultaneous equation system that enables the assessment of the contribution of the web-based indicators in both equations at the same time. Since we expected both propositions to be validated to some extent, the simultaneous equation would allow to assess the stronger association with the web-based variables.

To estimate this simultaneous model, two sets of web-based indicators were built based on the standardised document frequency of keywords and one based and validated with a PLS regression method. Due to the need to adhere to publication guidelines in terms of length, the PLS constructs were omitted from the paper. However, they are provided as supplementary materials in Appendix D. These supplementary materials offer additional insights into the development and validation of the web-based indicators, which strengthen the overall findings of this study. Our results suggest that the use of web-based innovation-related language does not explain innovation evidence from survey data, while it does have a strong association with the investment received. Furthermore, all the 3SLS and OLS regression models corresponding to the two sets of web-based indicators were performed and yield the same results which is a testimony of the robustness of both the web-based measurements and the results from our equation models.

The first equation result could suggest that radical innovations, which are riskier and less appealing for investment, were predominantly measured in the study (Brown et al., 2009). However, these firms paradoxically require more external financing (Kalil, 2005; McNeil et al., 2007) which should lead to a stronger signal as it is the case with the second equation. Alternatively, it is possible the funds were either associated with poor innovation results or that the innovation outcomes are coming. Considering the extensive literature supporting a positive correlation

between external financing and innovation (Cornaggia et al., 2015; Hsu et al., 2014; Kerr & Nanda, 2015; Wellalage & Fernandez, 2019), it is plausible that signalling played a role in securing investment, which has yet to yield tangible innovation outcomes.

The paper discusses the two possible causal links with the association between the web-based signal and the investment received i.e., receiving funds enables innovation activities which enables the creation of a web-based signal or innovation activities enabled web-based signal which is instrumental to obtain the investment. Although we do not possess the time series data required to make definitive claims, the evidence gathered from the needs of the respondents (SMEs served by intermediaries to improve their innovation outcomes and/or access resources), the lack of correspondence between both the investment and the web-based variable and the evidence of innovation results, the marginal effects analysis that shows diminishing returns between the signal and the investment received, and the post-hoc analysis that shows a stronger association between signalling and strong innovation evidence on the investment all seem to support the idea that the signal is facilitating the acquisition of external financing.

Ultimately, the evidence of innovation indicator corresponds to a company's recent innovation success. For investors, the task at hand is to seek out current and pertinent information that enables them to make informed decisions about the prospects of potential investment candidates. Although past innovation success can help mitigate some of the risks associated with investment (as it is suggested by the post hoc analysis), it does not translate automatically into new high-yield investment opportunities. Indeed, it is not rare to see firms taking less risk as they grow more mature, which translate into sustainable or efficiency innovation which may not require as much investment (Christensen, 1997).

Conversely, a company with current lower levels of evidence of innovation provides little insight into its potential for upcoming innovations. Hence, such firms may potentially represent a greater chance to achieve higher-yield innovations, particularly in the case of small younger smaller firms (Breitzman & Hicks, 2008). As a result, it may be plausible that web-based indicators, which appear to be closely associated with investment which then supports innovation activities, could potentially serve as a predictor of a firm's upcoming innovation success. More research is needed to assess these relationships.

We then performed a series of marginal effects analysis after our regressions to better understand how the web-based indicators seem to impact the investment received. Most web-based indicators are associated with an increase in investment, with diminishing returns, but not to a point where the use of innovation-related language starts to have a negative impact. Therefore, words may be cheap and convenient signals (Lachmann et al., 2001), but it appears that when innovation-related language is associated with investment, it has a correspondence with reality.

Furthermore, this analysis showed the different impacts related to signalling factors. The IP web-based indicator has the highest initial impact on the investment received. First of all, this indicator is primarily driven by the term "patent." The process of obtaining a patent is costly, and when enforced, it comes with significant penalties. Patent information is public, making it easy to validate and nearly impossible to fake. Thus, it seems that the web-based IP indicator is capturing a costly signal. However, it appears that once you use that signalling card, there is a significant drop in investment value upon reusing it.

Firms may display their collaborative strength by showcasing instances of collaborative projects, which implies the participation of other firms, something that would be difficult to fake. It is possible that multiple displays of collaboration constitute a collaborative reputation, a costly signal valued by investors. We then notice web-based R&D signalling to behave very similarly to web-based collaboration signalling, although it appears to be slightly less impactful. It is possible that a key difference is the use of collaborative language may be sent unwittingly, which may make investors unconsciously predisposed to want to work with firms that display a collaborative culture as they would feel more secure in their involvement, thus mitigating the risk of an investment. OI web-based signalling follows a similar pattern as Collaboration and R&D, but with a lower impact on the investment received, and a lower impact than IP signalling at the highest frequencies. It is mainly driven by the word "community," which may either be a virtuous signal to show a commitment to the local community or a costly signal referring to their community of users, both shown to be valuable in their own ways. Finally, innovation culture-related terms pulled with “innov”, which is described as a “buzzword” (Bensaude Vincent, 2014; Bergfeld & Weber, 2011; Bock B. B., 2012; Brem, 2009; Kuratko et al., 2014; Mirończuk & Protasiewicz, 2020), may not convey as much value as other innovation factors, but our results show that they still hold value in attracting investment when considered alone. However, we observed that when more specific

innovation factors are present, they overshadow innovation culture, emphasising the importance of using precise language for effective signalling, in line with Lachmann et al. (2001) discussion.

Furthermore, the post-hoc analysis revealed that higher web-based innovation signalling, when aligned with the strongest evidence of innovation, is associated with the highest investments. This suggests that our web-based and questionnaire-based indicators are indeed measuring different albeit complementary, key elements that appear valuable from the investment market perspective. As the signal originates from the website, which serves as a showcase for a firm's most valuable attributes, it is proposed that the signal actually reflects the firm's engagement in marketing and branding equity activities that support their innovation endeavours (see OECD & Eurostat, 2018 section 4.2.3). This suggests that there are synergies between stronger marketing and branding equity activities, stronger innovation evidence and greater investments received.

The primary contribution of this paper is that our research design, using this particular sample, appears to identify innovation-active firms seeking external financial resources, rather than innovative firms. This suggests that the most innovative firms do not seem to use as much their innovative attributes for signalling purpose which means they may not need as much external resources for innovating. This supports the proposition that these web-based indicators represent marketing and branding equity activities supporting innovation for firms that are actively seeking external resources, which are essential for their innovation efforts. This may serve as an important metric to monitor commercial support for innovation. For example, Canada is a world leader in science and research, but struggles to convert these strengths into commercial success (Council of Canadian Academies, 2018). Using these indicators at scale on a population of firms could potentially help monitor innovation-active firms that are seeking external support, and establish policies accordingly. These indicators can be paired with current and past innovation policies and programs to better evaluate their impact on the national innovation system.

This paper also makes significant methodological contributions. In collaboration with Davide Pulizzotto, a process was developed to obtain the text content of a website from the past using the Wayback Machine, which has been made open source. This allows researchers to enrich existing databases with new data. Additionally, formative indicators were created using document keyword frequencies, and the paper demonstrates how to assess construct validity with a PLS regression method.

7.3 Innovation signaling vs the innovation literature

The third paper investigates the associations between different firms' innovation-related characteristics and web-based indicators. By doing this, the paper assesses how these associations compare to what would be expected from the literature if the signal corresponds to reality. Seven main propositions related to specific dimensions of a firm's innovation profile were tested, including real innovation evidence level, time to market, level of knowledge intensity and sectorial attributes. These propositions were tested using a large sample of 1594 respondents, primarily small technological enterprises from Finland and Canada that are seeking to improve their innovation capabilities by being served with innovation intermediary services.

In preparation for this paper, two new sets of web-based indicators were developed for one-way ANOVA analysis and regression analysis. The first set corresponds to a ratio based on standardised frequency, which allows for comparison of different signals among various characteristics in the ANOVA analysis. The second set uses the count of occurrences of innovation-related terms in sentences as dependent variables, corresponding to the real distribution without any transformation. A ratio based on the number of sentences standardised by firm size and sector was added as control variables to account for variations in websites and exposure to the Wayback Machine crawl. Due to the necessity of complying with publication length restrictions, the ANOVA analysis was not included directly in the paper, but was made available as an online Appendix. These analyses are available in Appendix E, along with additional supplementary materials, including summaries of the results of both ANOVA and regressions in Tables E-36 to E-41. Non-parametric tests like Kruskal-Wallis, Mann-Whitney U, and Wilcoxon rank were used due to non-normal distribution and small sample sizes. For instance, while the low number of subjects limits the validity of this result, the ANOVA analysis suggests that science-based firms (33 subjects) when compared to scale-based firms (10 subjects) exhibit stronger overall innovation, collaboration, and R&D signals which was anticipated (Castellacci, 2008). Other results from the ANOVA were very similar in terms of innovation evidence, knowledge intensity, age and country, but different for time to market, sectors and sectorial categories and the size. For instance, the ANOVA analysis suggested that a longer time to market, larger firms and life science and Health firms were generally associated with a stronger signal. It is appearing that the regression analysis that accounts for the other variables' impact on the signal significantly reduced the strength of

these relationships. Furthermore, the ANOVA through the Wilcoxon-ranked test allowed to observe the relative impact of each dimension among each other by category. Bottom line, it seems the distribution of the different dimensions are fairly evenly distributed among all the categories with R&D the dimension that is consistently the most signalled, followed by innovation culture and collaboration.

With the regression results, the paper demonstrates that the overall innovation signal strength, built from all the related innovation keywords, has a strong correspondence with what was anticipated from the literature. Out of seven propositions, four were validated and two partially validated. It was reported that firms with a stronger overall innovation signal seem to have some level of innovation results, a higher level of knowledge intensity, and come from sectors characterised by either a technological regime or a sectorial category that would predict higher innovation rates. Only the time to market and age do not have any influence on the overall signal strength.

Furthermore, one of the most important contributions of this paper is to bring valuable nuance to the findings from the second paper when it comes to the level of innovation evidence. Firms with strong evidence of innovation have a stronger signal than firms with no evidence of innovation at all, but they are not the ones related to the strongest signal. It is the firms with some evidence of innovation that have the strongest signal. Besides the conclusion from Paper 2, which indicates that firms with the highest signalling levels are associated with a greater need for external resources, the findings from Paper 3 suggest that these firms exhibiting the most signalling are also significantly linked to some innovation evidence. This implies that moderately innovative firms have the strongest overall signal.

The paper also revealed some tendencies related to sectoral differences. It was observed that firms with a stronger signal were more associated with firms in digital media and ICT or in the KIBS category. In contrast, it was also observed that manufacturing firms, even with a higher level of knowledge, and specialised-suppliers were having a weaker signal than anticipated. It was suggested that this finding may possibly capture attitudes towards digital technologies.

Another important contribution of Paper 3 is that it allows for the assessment of the web-based indicators themselves based on the results. For instance, we have seen that the overall innovation signal, the R&D signal, and to a slightly lesser extent, the IP signal, really correspond well with what was anticipated. The case of the IP signal being related to some unique patterns, such as being

associated with a longer time to market or with older firms, suggests a strong association with real IP activities, which was also reported in Paper 2. Innovation culture led by the term "innov" is also associated with at least some level of innovation evidence, which suggests that beyond the buzzword, there is something tangible behind the use of the "innovation" lexical field. This may even potentially help differentiate innovative firms from non-innovative firms (OECD & Eurostat, 2018).

However, collaboration, innovation culture, and especially OI proved to be much less corresponding to expectations. This points to the fact that all of these dimensions are more elusive and therefore harder to measure accurately with established keywords related to a lexical field that is shared by both academia and industry. This reveals the limitation of the keyword analysis method and suggests that more sophisticated methods should be considered to capture signals that correspond to these dimensions.

7.4 A note about trustworthiness

As shown in the literature section, many examples of corporations sending misleading and dishonest signals were given, which may call into question the validity of the corporate website as a source of data for research. However, we saw these examples in sectors that are particularly operating under environmental and socially sensitive industries, where dishonesty is typically supported by a cost-benefit analysis that estimates the punishment of being caught to be just as costly as telling the truth or nothing at all. However, our innovation signalling cost function does not share this dynamic, as firms promote their skills and attributes to seek external resources, such as to facilitate investment, as shown in the second paper. The dynamic is therefore closer to the evolutionary biology signalling perspective, in which firms act as males and investors as females, and firms compete among each other for investments. Firms that attempt to signal fake innovative attributes are less likely to receive investment, as investors perform due diligence. Therefore, there are little incentives to send innovation misleading signals, as it would more likely result in a waste of time and financial resources. This analysis suggests that it is reasonable for researchers to assume overall trustworthiness when it comes to corporate signalling of innovation. Finally, the validation of propositions 1.2, 2.2, and most propositions related to the overall innovation, R&D and IP signals strength in Paper 3 suggests that there is indeed a real fire behind the smoke.

CHAPTER 8 CONCLUSION AND RECOMMENDATIONS

The research articles presented explore the relationship between the language used on corporate websites and innovation, yielding meaningful insights for various stakeholders, such as scholars, businesses, investors, and policymakers. This thesis offers a comprehensive understanding of the use and implications of innovation-related language present on corporate websites and assesses the epistemic value of these indicators for research purposes. Additionally, it demonstrates how indicators built through keyword frequency methods can capture innovation-related signals and proposes potential applications while unveiling their limitations. The thesis enhances our understanding of web-based innovation signalling through the examination of 11 main propositions. To that end, we presented three research papers that complement each other by offering unique perspectives and collectively provide a comprehensive view of what can be expected from both corporate websites and web-based indicators built with a keyword frequency method. Moreover, to evaluate our propositions, we used two distinct datasets and employed five distinct web-based indicator constructs, all based on keyword frequency methods.

In addition to the innovation management literature supporting the main themes of the research, we also provide an analysis of corporate websites as an object of study based on academic publications in the field of semiotics, signalling theories, corporate management strategy, corporate social responsibility, and branding. This analysis provides examples and arguments that expose the epistemic limitations of corporate websites as a data source, but also highlights opportunities. Subjectivity and selection bias are inherent to this data source, making it unfit to replace a valid epistemic tool such as a questionnaire-based survey. However, viewing a corporate website as a signalling platform in which the signals may be decoded to assess the most important needs of a firm aimed toward external stakeholders is a valuable perspective, as we have demonstrated. As a signal includes both elements of corporate culture and strategy, there is an opportunity to obtain rich complementary information that would not be possible to obtain through a classic method. Furthermore, the signalling perspective teaches us that signalling occurs because there is a need to fulfil. Indeed, a firm would not waste space on its website and the time of its visitors on something that would not serve its objectives.

Furthermore, the research presented was made possible by the use and development of many different advanced methodologies that allowed for the empirical assessment of our propositions.

The MTMM matrix followed by the CFA methodology demonstrated the construct validity of the web-based and questionnaire-based indices, suggesting a tangible correspondence between reality and the signal. Although it is not clear what was measured, it is clear what it is not measuring, namely objective, precise measures related to specific activities. Therefore, it seems to provide some level of evidence for the importance of an innovation factor to the firm, which is a corporate culture artefact.

The signalling perspective is typically employed in game theory models rather than statistical inference models. However, the insight to use the investment received not only as a direct measure but also as a proxy for innovation activities within a simultaneous equations model allowed us to determine the best correspondence with the innovation signal. The results were not close, as the signal clearly has a strong correspondence with the investment received and none with the evidence of innovation. This demonstrates that from an innovation process perspective, signalling is more related to a stakeholder's input than a firm's output. It implies that these indicators can help identify firms that are actively seeking external support when it comes to their innovation process, and that signalling their innovation prowess is instrumental in receiving that support.

Finally, the last paper brought more nuances to what can be expected from these indicators. The signal has a stronger association with moderately innovative firms operating in sectors that are expected to have more innovation intensity. Moreover, the results of this paper seem to highlight the biggest limitations of the methodology. First, our results suggest that the signal is more likely to be stronger for firms within a sector that seems to have a higher level of comfort with digital technologies. Furthermore, the results of what was expected by factor showed good correspondence with our overall innovation and R&D indicators, but also revealed the limitation of keyword-based indicators when the lexical field does not seem to correspond well between academic and industrial language.

The results of our thesis have important policy-making implications. Innovation agencies could potentially use web-based indicators to assess firms' innovation-related needs and monitor them over time. By analysing a population of firms' corporate websites, agencies could gather up-to-date data on these needs and support innovation at a regional or national level. This process may enable timely responses and appropriate mechanisms to support innovation at a regional or national

level. In light of this analysis, we recommend that the Canadian government consider this approach to improve the national innovation capabilities.

8.1 Research limitations

The primary limitation of the first paper is the small sample size, which could compromise the robustness of the results. Furthermore, 10 websites were not captured by Nutch, the web scraper used, due to novel technologies implemented on those websites. This limitation motivated us to build our own web scraper for Papers 2 and 3. Firms operating in nanotechnology and advanced materials are known to develop general-purpose technologies. These technologies require a high innovation rate and significant external financing, creating favourable conditions for this study. As a result, this specific context may also affect the generalisability of the study's results.

Both papers 2 and 3 rely on the database collected between 2010 and 2016 provided by TEN. It is possible the current trends in website elaboration are not the same as the ones were back then, thus it may lead to different results. The surveys were designed to assess the impact of innovation intermediaries, and innovation was assessed with this goal in mind. Furthermore, many manual operations were performed to retrieve websites, categorise organisational types of firms, and recategorise many sectors. Moreover, significant operations were performed in Excel to harmonise the results into one comprehensive database. Finally, although we followed the state-of-the-art guidelines (Särdnal et al., 1992), we used imputation to maximise the regressions sample size which may have affected the results.

The Wayback Machine has many limitations, notably inconsistent crawling within and across websites. This may give more or less exposure to firms with characteristics that fit criteria that raise the likelihood of being crawled. Unfortunately, despite extensive research and attempts to communicate with people from the archive.org organisation, we were not able to obtain more clarity on this process. An example of a potential bias of the Wayback Machine may be related to its favouring English content, as we retrieved significantly less content associated with Finnish firms. Speaking of which, we performed translation to obtain only English content. However, the automatic translation services we used for all three papers are not perfect, and they may have altered the meaning behind words, leading to false positives or false negatives.

All papers use a keyword frequency analysis method, which was chosen in part due to its established status in webometrics (Thelwall, 2009) and in existing web-based innovation studies

(S. K. Arora et al., 2013; Gök et al., 2015; Libaers et al., 2016; Youtie et al., 2012). It was also selected for its simplicity of use and interpretation of results, which allowed for easy comparisons across all papers and the construction of a comprehensive knowledge base regarding its strengths and limitations. When an established lexical field is used by both industry and academia, it appears to correspond well with its corresponding innovation factor. However, if the lexical field is not well established or not adopted by the industry, or if the industry does not consider the lexical field important for meeting its strategic goals with stakeholders, the resulting measure may not be a valid representation of the factor it is intended to represent. It is also possible that the industry uses different terms associated with innovation factors, such as OI. Additionally, words may have different meanings in different situations, which is not considered by the methodology. A rule-based variable analysis may enhance true positive rates (Gök et al., 2015), but it may also increase the likelihood of false negatives.

8.2 Future research

A more effective long-term data acquisition process would be to implement regular web scraping on selected firms, in addition to periodic questionnaire-based surveys, to collect panel data and enable longitudinal studies that can further validate our findings and enhance our understanding of web-based signals and their meanings.

A keyword-based methodology is a top-down approach that is conceptually grounded, but it may lack correspondence with the field, as we just discussed. Bottom-up approaches such as topic modelling with LDA, or training machine learning models, and the late emergence of Large Language Models (LLM)-based technologies such as embeddings may have the potential to improve significantly the recall, i.e., to reveal underlying instances of innovation signalling that we would miss using keywords alone. These approaches may be particularly promising in capturing signals that firms may unwittingly send about their culture that supports innovation, or even practices related to OI or collaboration. As we saw, these dimensions are more elusive and therefore more challenging to define and identify using keywords alone.

We focused on corporate websites as a source of innovation-related signals, but other communication channels, such as social media and corporate reports, can also be examined to gain a deeper understanding of a firm's motivations and needs. By understanding the role and characteristics of a social media platform along with the strategic objectives of firms in relation to

their stakeholders, one may identify innovation-related dimensions worth reporting. For instance, LinkedIn can be viewed from a recruitment perspective where firms are expected to share information that showcases their expertise, values, and innovation capabilities to attract potential employees. An analysis of a sample of social media posts could reveal how innovation-related signals stand out in these contexts.

Other techniques, such as sentiment analysis, can help assess the subjectivity behind signals and reveal whether signals sent are positive, negative, or neutral. Furthermore, we solely considered textual content retrieved, but other media, such as images, photos, videos, and web design elements, are extremely rich in information. Although converting textual information into structured data is already challenging, the conversion of multimedia content into objective, quantifiable signalling units will require considerable research. It will be interesting to assess the potential use of large multimodal models in this area of research.

Finally, other dimensions related to a firm's needs may be studied using a similar signalling approach. Corporate strategies and other firm priorities may be unveiled from a signalling perspective. We hope that the methodologies and results presented in this thesis will inspire further research in this area, which may lead to unparalleled insights into our business ecosystems and a better understanding of the factors that foster innovation at scale.

REFERENCES

Acs, Z. J., & Audretsch, D. B. (1990). *Innovation and small firms*. MIT press.

Acs, Z. J., & Preston, L. (1997). Small and Medium-Sized Enterprises, Technology, and Globalization: Introduction to a special issue on Small and Medium-Sized Enterprises in the Global Economy. *Small Business Economics*, *9*(1), 1–6. https://doi.org/10.1023/A:1007945327618

Adams, D. R., Flatten, T. C., Brinkmann, H., & Brettel, M. (2015). Consequences And Antecedents Of Absorptive Capacity In A Cross-Cultural Context. *International Journal of Innovation Management*, *20*(01), 1650003.

Adams, R., Bessant, J., & Phelps, R. (2006). Innovation management measurement: A review. *International Journal of Management Reviews*, *8*(1), 21–47. https://doi.org/10.1111/j.1468-2370.2006.00119.x

Afonso, P., Nunes, M., Paisana, A., & Braga, A. (2008). The influence of time-to-market and target costing in the new product development success. *International Journal of Production Economics*, *115*(2), 559–568. https://doi.org/10.1016/j.ijpe.2008.07.003

Afuah, A. (1998). *Innovation Management: Strategies, Implementation and Profits*. Oxford University Press.

Aghion, P., Bond, S., Klemm, A., & Marinescu, I. (2004). Technology and Financial Structure: Are Innovative Firms Different? *Journal of the European Economic Association*, *2*(2–3), 277–288. https://doi.org/10.1162/154247604323067989

Al Baghal, T., Wenz, A., Sloan, L., & Jessop, C. (2021). Linking Twitter and survey data: Asymmetry in quantity and its impact. *EPJ Data Science*, *10*(1), Article 1. https://doi.org/10.1140/epjds/s13688-021-00286-7

Aldas-Manzano, J., Küster, I., & Vila, N. (2005). Market orientation and innovation: An inter-relationship analysis. *European Journal of Innovation Management*, *8*(4), 437–452. https://doi.org/10.1108/14601060510627812

Al-Hassan, A. A., Alshameri, F., & Sibley, E. H. (2013). A research case study: Difficulties and recommendations when using a textual data mining tool. *Information & Management*, *50*(7), 540–552. https://doi.org/10.1016/j.im.2013.05.010

Allen, R. C. (1983). Collective invention. *Journal of Economic Behavior & Organization*, *4*(1), 1–24.

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the world wide web: Methodological approaches to 'webometrics'. *Journal of Documentation*, *53*(4), 404–426. https://doi.org/10.1108/EUM0000000007205

Amabile, T. M. (1996). *Creativity in context: Update to*.

Amabile, T. M. (1997). Motivating Creativity in Organizations: On Doing What You Love and Loving What You Do. *California Management Review*, *40*(1), 39–58. https://doi.org/10.2307/41165921

Amabile, T. M., Conti, R., Coon, H., Lazenby, J., & Herron, M. (1996). Assessing the Work Environment for Creativity. *Academy of Management Journal*, *39*(5), 1154–1184. https://doi.org/10.5465/256995

Amara, N., Landry, R., Becheikh, N., & Ouimet, M. (2008). Learning and novelty of innovation in established manufacturing SMEs. *Technovation*, *28*(7), 450–463. https://doi.org/10.1016/j.technovation.2008.02.001

Amara, N., Landry, R., Becheikh, N., & Ouimet, M. (2004, June 14). Radical innovations in traditional manufacturing industries. *Druid Summer Conference,, Industrial Dynamics, Innovation and Development*. DRUID Summer Conference 2004, Copenhagen, Denmark.

Anderson, N., De Dreu, C. K. W., & Nijstad, B. A. (2004). The routinization of innovation research: A constructively critical review of the state-of-the-science. *Journal of Organizational Behavior*, *25*(2), 147–173. https://doi.org/10.1002/job.236

Anderson, P., & Tushman, M. L. (1990). Technological discontinuities and dominant designs: A cyclical model of technological change. *Administrative Science Quarterly*, 604–633.

Anderson, T. W. (1963). Asymptotic Theory for Principal Component Analysis. *The Annals of Mathematical Statistics*, *34*(1), 122–148. https://doi.org/10.1214/aoms/1177704248

Andersson, M. (2019). Sexual Selection. In *Sexual Selection*. Princeton University Press. https://doi.org/10.1515/9780691207278

Antonelli, C. (1998). Localized technological change, new information technology and the knowledge-based economy: The European evidence. *Journal of Evolutionary Economics*, *8*(2), 177–198. https://doi.org/10.1007/s001910050061

Antonelli, G. A., Leone, M. I., & Ricci, R. (2022). Exploring the Open COVID Pledge in the fight against COVID-19: A semantic analysis of the Manifesto, the pledgors and the featured patents. *R&D Management*, *52*(2), 255–272. https://doi.org/10.1111/radm.12493

Antons, D., Grünwald, E., Cichy, P., & Salge, T. O. (2020). The application of text mining methods in innovation research: Current state, evolution patterns, and development priorities. *R&D Management*, *50*(3), 329–351. https://doi.org/10.1111/radm.12408

Archibugi, D. (1992). Patenting as an indicator of technological innovation: A review. *Science and Public Policy*, *19*(6), 357–368. https://doi.org/10.1093/spp/19.6.357

Archibugi, D., & Sirilli, G. (2001). The direct measurement of technological innovation in business. in "Innovation and enterprise creation: Statistics and indicators. *Proceedings of the Conference Held at Sophia Antipolis*.

Armellini, F., Beaudry, C., & Kaminski, P. C. (2017). Open within a box: An analysis of open innovation patterns within Canadian aerospace companies. *Sinergie Italian Journal of Management*, *0*(0). http://sinergiejournal.eu/index.php/sinergie/article/view/1347

Armellini, F., Kaminski, P. C., Beaudry, C., Armellini, F., Kaminski, P. C., & Beaudry, C. (2014). The Open Innovation Journey in Emerging Economies: An Analysis of the Brazilian Aerospace Industry. *Journal of Aerospace Technology and Management*, *6*(4), 462–474. https://doi.org/10.5028/jatm.v6i4.390

Arora, A. (1995). Licensing Tacit Knowledge: Intellectual Property Rights And The Market For Know-How. *Economics of Innovation and New Technology*, *4*(1), 41–60. https://doi.org/10.1080/10438599500000013

Arora, A., & Gambardella, A. (1994). The changing technology of technological change: General and abstract knowledge and the division of innovative labour. *Research Policy*, *23*(5), 523–532.

Arora, S. K., Li, Y., Youtie, J., & Shapira, P. (2016). Using the wayback machine to mine websites in the social sciences: A methodological resource. *Journal of the Association for Information Science and Technology*, *67*(8), 1904–1915. https://doi.org/10.1002/asi.23503

Arora, S. K., Li, Y., Youtie, J., & Shapira, P. (2020). Measuring dynamic capabilities in new ventures: Exploring strategic change in US green goods manufacturing using website data. *The Journal of Technology Transfer*, *45*(5), 1451–1480. https://doi.org/10.1007/s10961-019-09751-y

Arora, S. K., Youtie, J., Shapira, P., Gao, L., & Ma, T. (2013). Entry strategies in an emerging technology: A pilot web-based study of graphene firms. *Scientometrics*, *95*(3), 1189–1207. https://doi.org/10.1007/s11192-013-0950-7

Arrow, K. (1962). Economic welfare and the allocation of resources for invention. In *The rate and direction of inventive activity: Economic and social factors* (pp. 609–626). Princeton University Press. http://www.nber.org/chapters/c2144.pdf

Arvanitis, S. (2012). How do different motives for R&D cooperation affect firm performance? – An analysis based on Swiss micro data. *Journal of Evolutionary Economics*, *22*(5), 981–1007. https://doi.org/10.1007/s00191-012-0273-5

Atkin, A. (2022). Peirce's Theory of Signs. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/peirce-semiotics/

Augusto, M., & Coelho, F. (2009). Market orientation and new-to-the-world products: Exploring the moderating effects of innovativeness, competitive strength, and environmental forces. *Industrial Marketing Management*, *38*(1), 94–108. https://doi.org/10.1016/j.indmarman.2007.09.007

Axenbeck, J., & Breithaupt, P. (2021). Innovation indicators based on firm websites—Which website characteristics predict firm-level innovation activity? *PLOS ONE*, *16*(4), e0249583. https://doi.org/10.1371/journal.pone.0249583

Bae, S. C., & Kim, D. (2003). The Effect of R&D Investments on Market Value of Firms: Evidence from the U.S., Germany, and Japan. *Multinational Business Review*, *11*(3), 51–76. https://doi.org/10.1108/1525383X200300016

Bagozzi, R. P. (1994). *Principles of marketing research*. Blackwell Business.

Bagozzi, R. P. (2011). Measurement and meaning in information systems and organizational research: Methodological and philosophical foundations. *Mis Quarterly*, 261–292.

Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing Construct Validity in Organizational Research. *Administrative Science Quarterly*, *36*(3), 421–458. https://doi.org/10.2307/2393203

Baldwin, J. R. (1995). *Productivity Growth, Plant Turnover and Restructuring in the Canadian Manufacturing Sector* (SSRN Scholarly Paper No. 4255). https://doi.org/10.2139/ssrn.4255

Baldwin, J. R., & Rafiquzzaman, M. (1995). Restructuring in the Canadian Manufacturing Sector from 1970 to 1990: Industry and Regional Dimensions of Job Turnover. *Publications Review Committee, Analytical Studies Branch, Statistics Canada*.

Bar-Anan, Y., & Vianello, M. (2018). A multi-method multi-trait test of the dual-attitude perspective. *Journal of Experimental Psychology: General*, *147*(8), 1264–1272. https://doi.org/10.1037/xge0000383

Barney, J. B. (1986). Organizational culture: Can it be a source of sustained competitive advantage? *Academy of Management Review*, *11*(3), 656–665.

Batini, C., Rula, A., Scannapieco, M., & Viscusi, G. (2015). From data quality to big data quality. *Journal of Database Management (JDM)*, *26*(1), 60–82.

Batini, C., & Scannapieco, M. (2006). *Data quality. Data-centric Systems and Applications*. Springer.

Baumol, W. (2005). Small firms: Why market-driven innovation can't get along without them. *The Small Business Economy: A Report to the President*, 183–206.

Baysinger, B., & Hoskisson, R. E. (1989). Diversification Strategy and R&D Intensity in Multiproduct Firms. *Academy of Management Journal*, *32*(2), 310–332. https://doi.org/10.2307/256364

Bayus, B. L. (1997). Speed-to-market and new product performance trade-offs. *Journal of Product Innovation Management*, *14*(6), 485–497. https://doi.org/10.1016/S0737-6782(97)00062-3

Beaudry, C., & Solar-Pelletier, L. (2019). The Superclusters Initiative: An Opportunity to Reinforce Innovation Ecosystems. *IRPP Study 79. Montreal: Institute for Research on Public Policy.* https://doi.org/10.26070/jft2-9055

Becheikh, N., Landry, R., & Amara, N. (2006). Lessons from innovation empirical studies in the manufacturing sector: A systematic review of the literature from 1993–2003. *Technovation*, *26*(5), 644–664. https://doi.org/10.1016/j.technovation.2005.06.016

Becker, M. C., & Knudsen, T. (2002). Schumpeter 1911: Farsighted visions on economic development. *American Journal of Economics and Sociology*, *61*(2), 387–403.

Belderbos, R., Carree, M., & Lokshin, B. (2004). Cooperative R&D and firm performance. *Research Policy*, *33*(10), 1477–1492. https://doi.org/10.1016/j.respol.2004.07.003

Bensaude Vincent, B. (2014). The politics of buzzwords at the interface of technoscience, market and society: The case of 'public engagement in science'. *Public Understanding of Science*, *23*(3), 238–253. https://doi.org/10.1177/0963662513515371

Bergfeld, M.-M. H., & Weber, F.-M. (2011). Dynasties of innovation: Highly performing German family firms and the owners' role for innovation. *International Journal of Entrepreneurship and Innovation Management*, *13*(1), 80–94. https://doi.org/10.1504/IJEIM.2011.038449

Bergh, D. D., Connelly, B. L., Ketchen Jr, D. J., & Shannon, L. M. (2014). Signalling Theory and Equilibrium in Strategic Management Research: An Assessment and a Research Agenda. *Journal of Management Studies*, *51*(8), 1334–1360. https://doi.org/10.1111/joms.12097

Bergstrom, C. T., & Lachmann, M. (1998). Signaling among relatives. III. Talk is cheap. *Proceedings of the National Academy of Sciences*, *95*(9), 5100–5105. https://doi.org/10.1073/pnas.95.9.5100

Berthon, P., Hulbert, J. M., & Pitt, L. F. (1999). To Serve or Create? Strategic Orientations toward Customers and Innovation. *California Management Review*, *42*(1), 37–58. https://doi.org/10.2307/41166018

Berthon, P. R., Lord Ferguson, S. T., Pitt, L. F., & Wang, E. (2023). The virtuous brand: The perils and promises of brand virtue signaling. *Business Horizons*, *66*(1), 27–36. https://doi.org/10.1016/j.bushor.2021.10.006

Bertozzi, M. (2007). Chasseurs d'indices. Quelques réflexions sur les formes de rationalité et les ruses de l'intelligence. *THOUARD, Denis (Éd.). L'interprétation Des Indices: Enquête Sur Le Paradigme Indiciaire Avec Carlo Ginzburg. Villeneuve d'Ascq: Presses Universitaires Du Septentrion*, 25–36.

Beugelsdijk, S., & Cornet, M. (2002). 'A Far Friend is Worth More than a Good Neighbour': Proximity and Innovation in a Small Country. *Journal of Management and Governance*, *6*, 169–188.

Biernaskie, J. M., Grafen, A., & Perry, J. C. (2014). The evolution of index signals to avoid the cost of dishonesty. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1790), 20140876. https://doi.org/10.1098/rspb.2014.0876

Birkinshaw, J., Bouquet, C., & Barsoux, J.-L. (2011). The 5 myths of innovation. *MIT Sloan Management Review*, *52*(2), 43.

Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, *55*(14), 1216–1227. https://doi.org/10.1002/asi.20077

Blazquez, D., & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, *130*, 99–113. https://doi.org/10.1016/j.techfore.2017.07.027

Blazquez, D., Domenech, J., & Debón, A. (2018). Do corporate websites' changes reflect firms' survival? *Online Information Review*, *42*(6), 956–970. https://doi.org/10.1108/OIR-11-2016-0321

Bliege Bird, R., & Smith, E. A. (2005). Signaling Theory, Strategic Interaction, and Symbolic Capital. *Current Anthropology*, *46*(2), 221–248. https://doi.org/10.1086/427115

Bock B. B. (2012). Social innovation and sustainability; how to disentangle the buzzword and its application in the field of agriculture and rural development. *Studies in Agricultural Economics*, *114*(2), Article 2.

Bock, D. E., Wolter, J. S., & Ferrell, O. C. (2020). Artificial intelligence: Disrupting what we know about services. *Journal of Services Marketing*, *34*(3), 317–334. https://doi.org/10.1108/JSM-01-2019-0047

Bogers, M., Zobel, A.-K., Afuah, A., Almirall, E., Brunswicker, S., Dahlander, L., Frederiksen, L., Gawer, A., Gruber, M., Haefliger, S., & others. (2017). The open innovation research landscape: Established perspectives and emerging themes across different levels of analysis. *Industry and Innovation*, *24*(1), 8–40.

Boone, R. T., & Buck, R. (2003). Emotional Expressivity and Trustworthiness: The Role of Nonverbal Behavior in the Evolution of Cooperation. *Journal of Nonverbal Behavior*, *27*(3), 163–182. https://doi.org/10.1023/A:1025341931128

Botcheva, L., White, C. R., & Huffman, L. C. (2002). Learning Culture and Outcomes Measurement Practices in Community Agencies. *American Journal of Evaluation*, *23*(4), 421–434. https://doi.org/10.1177/109821400202300404

Bottai, C., Crosato, L., Domenech, J., Guerzoni, M., & Liberati, C. (2022). Unconventional data for policy: Using Big Data for detecting Italian innovative SMEs. *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, 338–344. https://doi.org/10.1145/3524458.3547246

Bozdogan, K., Deyst, J., Hoult, D., & Lucas, M. (1998). Architectural innovation in product development through early supplier integration. *R&D Management*, *28*(3), 163–173. https://doi.org/10.1111/1467-9310.00093

Breitzman, A., & Hicks, D. (2008). *An analysis of small business patents by industry and firm size*.

Brem, A. (2009). The Term Innovation and its Front End – Is There a Specific Asian View? In P. Ordóñez de Pablos & M. D. Lytras (Eds.), *The China Information Technology Handbook* (pp. 1–12). Springer US. https://doi.org/10.1007/978-0-387-77743-6_5

Brentani, U. de. (2001). Innovative versus incremental new business services: Different keys for achieving success. *Journal of Product Innovation Management*, *18*(3), 169–187. https://doi.org/10.1111/1540-5885.1830169

Breschi, S., & Malerba, F. (1997). Sectoral innovation systems: Technological regimes, Schumpeterian dynamics, and spatial boundaries. *Systems of Innovation: Technologies, Institutions and Organizations*, *1*, 130–156.

Breschi, S., Malerba, F., & Orsenigo, L. (2000). Technological regimes and Schumpeterian patterns of innovation. *The Economic Journal*, *110*(463), 388–410.

Brown, J. R., Fazzari, S. M., & Petersen, B. C. (2009). Financing Innovation and Growth: Cash Flow, External Equity, and the 1990s R&D Boom. *The Journal of Finance*, *64*(1), 151–185. https://doi.org/10.1111/j.1540-6261.2008.01431.x

Brunswicker, S., & Van de Vrande, V. (2014). Exploring open innovation in small and medium-sized enterprises. *New Frontiers in Open Innovation*, *1*, 135–156.

Burke, J. (1998). Invention. In *Encyclopedia Britannica*. https://www.britannica.com/technology/invention-technology

Burns, T., & Stalker, G. M. (1961). Mechanistic and organic systems. *Classics of Organizational Theory*, 209–214.

Burton, M. D., Sørensen, J. B., & Beckman, C. M. (2002). 7. Coming from good stock: Career histories and new venture formation. In *Social structure and organisations revisited* (Vol. 19, pp. 229–262). Emerald Group Publishing Limited.

Bush, V. (1945). Science: The endless frontier. *Transactions of the Kansas Academy of Science (1903)*, 231–264.

Callegaro, M., & Yang, Y. (2018). The Role of Surveys in the Era of "Big Data". In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave Handbook of Survey Research* (pp. 175–192). Springer International Publishing. https://doi.org/10.1007/978-3-319-54395-6_23

Cameron, K. S., Freeman, S. J., & Mishra, A. K. (1991). Best practices in white-collar downsizing: Managing contradictions. *Academy of Management Perspectives*, *5*(3), 57–73.

Campbell, C. M., Michel, J. O., Patel, S., & Gelashvili, M. (2019). College Teaching from Multiple Angles: A Multi-trait Multi-method Analysis of College Courses. *Research in Higher Education*, *60*(5), 711–735. https://doi.org/10.1007/s11162-018-9529-8

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Capon, N., Farley, J. U., Lehmann, D. R., & Hulbert, J. M. (1992). Profiles of Product Innovators Among Large U.S. Manufacturers. *Management Science*, *38*(2), 157–169. https://doi.org/10.1287/mnsc.38.2.157

Carboni, O. A. (2013). Spatial and industry proximity in collaborative research: Evidence from Italian manufacturing firms. *The Journal of Technology Transfer*, *38*(6), 896–910. https://doi.org/10.1007/s10961-012-9279-2

Carpenter, C. R. (1969). Approaches to studies of the naturalistic communicative behavior in nonhuman primates. *Approaches to Animal Communication*, 40–70.

Castellacci, F. (2008). Technological paradigms, regimes and trajectories: Manufacturing and service industries in a new taxonomy of sectoral patterns of innovation. *Research Policy*, *37*(6), 978–994. https://doi.org/10.1016/j.respol.2008.03.011

Cebon, P., Newton, P., & Noble, P. (1999). Innovation in firms: Towards a framework for indicator development. *Melbourne Business School, Working Paper*, *99–9*.

Cenfetelli, R. T., & Bassellier, G. (2009). Interpretation of Formative Measurement in Information Systems Research. *MIS Quarterly*, *33*(4), 689–707.

Chandler, A. D. (1990). The enduring logic of industrial success. *Harvard Business Review*, *68*(2), 130–140.

Chaudhry, S. J., & Wald, K. A. (2022). Overcoming listener skepticism: Costly signaling in communication increases perceived honesty. *Current Opinion in Psychology*, *48*, 101442. https://doi.org/10.1016/j.copsyc.2022.101442

Chellappandi, P., & Vijayakumar, C. S. (2018). Bibliometrics, Scientometrics, Webometrics/Cybermetrics, Informetrics and Altmetrics—An Emerging Field in Library and Information Science Research. *Shanlax International Journal of Education*, *7*(1), 5–8.

Chen, L., Yan, Y., & Smith, A. N. (2022). What drives digital engagement with sponsored videos? An investigation of video influencers' authenticity management strategies. *Journal of the Academy of Marketing Science*. https://doi.org/10.1007/s11747-022-00887-2

Chesbrough, H. W. (2003a). *Open Innovation: The New Imperative for Creating And Profiting from Technology*. Harvard Business School Press.

Chesbrough, H. W. (2003b). The era of open innovation. *MIT Sloan Management Review*, *44*(3), 35–41.

Chesbrough, H. W. (2006a). *Open business models: How to Thrive in the New Innovation Landscape* (Boston: Harvard Business School Press.).

Chesbrough, H. W. (2006b). Open innovation: A new paradigm for understanding industrial innovation. *Open Innovation: Researching a New Paradigm*, 1–12.

Chesbrough, H. W. (2012). GE's ecomagination Challenge: An Experiment In Open Innovation. *California Management Review*, *54*(3), 140–154.

Chesbrough, H. W., & Bogers, M. (2014). *Explicating Open Innovation: Clarifying an Emerging Paradigm for Understanding Innovation* (SSRN Scholarly Paper No. 2427233). https://papers.ssrn.com/abstract=2427233

Chesbrough, H. W., & Brunswicker, S. (2014). A Fad or a Phenomenon? The Adoption of Open Innovation Practices in Large Firms. *Research Technology Management*, *57*(2), 16–25.

Chesbrough, H. W., & Rosenbloom, R. S. (2002). The role of the business model in capturing value from innovation: Evidence from Xerox Corporation's technology spin-off companies. *Industrial and Corporate Change*, *11*(3), 529–555.

Chesbrough, H. W., & Schwartz, K. (2007). Innovating business models with co-development partnerships. *Research-Technology Management*, *50*(1), 55–59.

Cho, C. H., & Patten, D. M. (2007). The role of environmental disclosures as tools of legitimacy: A research note. *Accounting, Organizations and Society*, *32*(7), 639–647. https://doi.org/10.1016/j.aos.2006.09.009

Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, *88*(s1), 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x

Christensen, C. M. (1997). *The innovator's dilemma*. Boston: Harvard Business School Press. http://soloway.pbworks.com/w/file/fetch/46695705/The%20Innovators%20Dilemma.pdf

Christensen, C. M. (2013). *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Harvard Business Review Press.

Christensen, C. M., & Raynor, M. (2013). *The innovator's solution: Creating and sustaining successful growth*. Harvard Business Review Press.

Chun, R., Da Silva, R., Davies, G., & Roper, S. (2005). *Corporate reputation and competitiveness*. Routledge.

Churchill, G. A. (1979). A Paradigm for Developing Better Measures of Marketing Constructs. *Journal of Marketing Research*, *16*(1), 64–73. https://doi.org/10.2307/3150876

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates.

Cohen, M. A., Eliasberg, J., & Ho, T.-H. (1996). New Product Development: The Performance and Time-to-Market Tradeoff. *Management Science*, *42*(2), 173–186. https://doi.org/10.1287/mnsc.42.2.173

Cohen, W. M., & Levin, R. C. (1989). Chapter 18 Empirical studies of innovation and market structure. In *Handbook of Industrial Organization* (Vol. 2, pp. 1059–1107). Elsevier. https://doi.org/10.1016/S1573-448X(89)02006-6

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quarterly*, *35*(1), 128–152. https://doi.org/10.2307/2393553

Cohen, W. M., Nelson, R. R., & Walsh, J. P. (2000a). *Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not)* (Working Paper No. 7552). National Bureau of Economic Research. https://doi.org/10.3386/w7552

Cohen, W. M., Nelson, R. R., & Walsh, J. P. (2002). Links and impacts: The influence of public research on industrial R&D. *Management Science*, *48*(1), 1–23.

Cohen, W. M., Nelson, R., & Walsh, J. P. (2000b). *Protecting their intellectual assets: Appropriability conditions and why US manufacturing firms patent (or not)*. National Bureau of Economic Research Cambridge, Mass., USA.

Coldbeck, B., & Ozkan, A. (2018). Comparison of adjustment speeds in target research and development and capital investment: What did the financial crisis of 2007 change? *Journal of Business Research*, *84*, 1–10.

Colombo, M. G., Guerini, M., Hoisl, K., & Zeiner, N. M. (2023). The dark side of signals: Patents protecting radical inventions and venture capital investments. *Research Policy*, *52*(5), 104741. https://doi.org/10.1016/j.respol.2023.104741

Cooke, S., & Ryan, P. (2000). Brand alliances: From reputation endorsement to collaboration on core competencies. *Irish Marketing Review*, *13*(2), 36.

Coombs, R., Narandren, P., & Richards, A. (1996). A literature-based innovation output indicator. *Research Policy*, *25*(3), 403–413. https://doi.org/10.1016/0048-7333(95)00842-X

Cooper, R. G., & Kleinschmidt, E. J. (1995). Benchmarking the Firm's Critical Success Factors in New Product Development. *Journal of Product Innovation Management*, *12*(5), 374–391. https://doi.org/10.1111/1540-5885.1250374

Coriat, B., Orsi, F., & Weinstein, O. (2003). Does Biotech Reflect a New Science-based Innovation Regime? *Industry and Innovation*, *10*(3), 231–253. https://doi.org/10.1080/1366271032000141634

Cornaggia, J., Mao, Y., Tian, X., & Wolfe, B. (2015). Does banking competition affect innovation? *Journal of Financial Economics*, *115*(1), 189–209.

Cornelissen, J. P. (2020). *Corporate Communication: A Guide to Theory and Practice*. SAGE Publications Ltd. http://digital.casalini.it/9781529712681

Council of Canadian Academies. (2018). *Competing in a Global Innovation Economy: The Current State of R&D in Canada* (Expert Panel on the State of Science and Technology

and Industrial Research and Development in Canada, Council of Canadian Academies). http://new-report.scienceadvice.ca/

Cowie, R. H., Bouchet, P., & Fontaine, B. (2022). The Sixth Mass Extinction: Fact, fiction or speculation? *Biological Reviews, 97*(2), 640–663. https://doi.org/10.1111/brv.12816

Crawley, T. (2007). Commercialization of Nanotechnology–Key Challenges. *Report on the Workshop Organised by Nanoforum, Helsinki*.

Cronk, L. (2005). The application of animal signaling theory to human phenomena: Some thoughts and clarifications. *Social Science Information*, *44*(4), 603–620. https://doi.org/10.1177/0539018405058203

Cropper, S., Ebers, M., & Huxham, C. (2008). *The Oxford handbook of inter-organizational relations*. Oxford Handbooks.

Cummings, J. L., & Teng, B.-S. (2003). Transferring R&D knowledge: The key factors affecting knowledge transfer success. *Journal of Engineering and Technology Management*, *20*(1), 39–68. https://doi.org/10.1016/S0923-4748(03)00004-3

Dahlander, L., & Gann, D. M. (2010). How open is innovation? *Research Policy*, *39*(6), 699–709. https://doi.org/10.1016/j.respol.2010.01.013

Damanpour, F., & Aravind, D. (2006). Product and process innovations: A review of organizational and environmental determinants. *Innovation, Science, and Industrial Change: A Research Handbook*, 38–66.

Datar, S., Jordan, C., Kekre, S., Rajiv, S., & Srinivasan, K. (1997). New Product Development Structures and Time-to-Market. *Management Science*, *43*(4), 452–464. https://doi.org/10.1287/mnsc.43.4.452

Day, G. S., & Schoemaker, P. J. H. (2000). Avoiding the Pitfalls of Emerging Technologies. *California Management Review*, *42*(2), 8–33. https://doi.org/10.2307/41166030

Deeds, D. L. (2001). The role of R&D intensity, technical development and absorptive capacity in creating entrepreneurial wealth in high technology start-ups. *Journal of Engineering and Technology Management*, *18*(1), 29–47. https://doi.org/10.1016/S0923-4748(00)00032-1

Dégh, L., & Vázsonyi, A. (1983). Does the Word 'Dog' Bite? Ostensive Action: A Means of Legend-Telling. *Journal of Folklore Research*, *20*(1), 5–34.

Deng, X., & Cheng, X. (2019). Can ESG Indices Improve the Enterprises' Stock Market Performance?—An Empirical Study from China. *Sustainability*, *11*(17), Article 17. https://doi.org/10.3390/su11174765

Derrida, J. (2016). *Of Grammatology*. JHU Press.

Diamantopoulos, A. (1999). Viewpoint – Export performance measurement: Reflective versus formative indicators. *International Marketing Review*, *16*(6), 444–457. https://doi.org/10.1108/02651339910300422

Diamantopoulos, A., & Siguaw, J. A. (2006). Formative Versus Reflective Indicators in Organizational Measure Development: A Comparison and Empirical Illustration. *British Journal of Management*, *17*(4), 263–282. https://doi.org/10.1111/j.1467-8551.2006.00500.x

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index Construction with Formative Indicators: An Alternative to Scale Development. *Journal of Marketing Research*, *38*(2), 269–277. https://doi.org/10.1509/jmkr.38.2.269.18845

Diercks, G., Larsen, H., & Steward, F. (2019). Transformative innovation policy: Addressing variety in an emerging policy paradigm. *Research Policy*, *48*(4), 880–894. https://doi.org/10.1016/j.respol.2018.10.028

Dilevko, J. (2019). Virtue-signaling and inclusivity. *Journal of Information Ethics*, *28*(2), 11–15.

Dobni, C. B. (2008). Measuring innovation culture in organisations: The development of a generalized innovation culture construct using exploratory factor analysis. *European Journal of Innovation Management*, *11*(4), 539–559. https://doi.org/10.1108/14601060810911156

Dodgson, M., Gann, D., & Salter, A. (2006). The role of technology in the shift towards open innovation: The case of Procter & Gamble. *R&D Management*, *36*(3), 333–346. https://doi.org/10.1111/j.1467-9310.2006.00429.x

Doloreux, D., & Frigon, A. (2022). Innovation Superclusters Initiative in Canada: A new policy strategy? *Science and Public Policy*, *49*(1), 148–158. https://doi.org/10.1093/scipol/scab071

Dörr, J. O., Kinne, J., Lenz, D., Licht, G., & Winker, P. (2022). An integrated data framework for policy guidance during the coronavirus pandemic: Towards real-time decision support for economic policymakers. *PLOS ONE*, *17*(2), e0263898. https://doi.org/10.1371/journal.pone.0263898

Dosi, G. (1988). Sources, Procedures, and Microeconomic Effects of Innovation. *Journal of Economic Literature*, *26*(3), 1120–1171.

Drechsler, W., & Natter, M. (2012). Understanding a firm's openness decisions in innovation. *Journal of Business Research*, *65*(3), 438–445.

Driskell, J. E., Salas, E., & Driskell, T. (2018). Foundations of teamwork and collaboration. *American Psychologist*, *73*, 334–348. https://doi.org/10.1037/amp0000241

Dumas, M., Barker, J. L., & Power, E. A. (2021). When does reputation lie? Dynamic feedbacks between costly signals, social capital and social prominence. *Philosophical Transactions*

of the Royal Society B: Biological Sciences, *376*(1838), 20200298. https://doi.org/10.1098/rstb.2020.0298

Dunlap, R. E., & McCright, A. M. (2011). Organized climate change denial. *The Oxford Handbook of Climate Change and Society*, *1*, 144–160.

Dziallas, M., & Blind, K. (2019). Innovation indicators throughout the innovation process: An extensive literature analysis. *Technovation*, *80–81*, 3–29. https://doi.org/10.1016/j.technovation.2018.05.005

Ebersberger, B., & Herstad, S. J. (2012). The relationship between international innovation collaboration, intramural R&D and SMEs' innovation performance: A quantile regression approach. *Applied Economics Letters*, *20*(7), 626–630. https://doi.org/10.1080/13504851.2012.724158

Eck, A., Cazar, A. L. C., Callegaro, M., & Biemer, P. (2021). "Big Data Meets Survey Science". *Social Science Computer Review*, *39*(4), 484–488. https://doi.org/10.1177/0894439319883393

Edler, J., & Georghiou, L. (2007). Public procurement and innovation—Resurrecting the demand side. *Research Policy*, *36*(7), 949–963.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. https://doi.org/10.1037/1082-989X.5.2.155

Eggers, F., O'Dwyer, M., Kraus, S., Vallaster, C., & Güldenberg, S. (2013). The impact of brand authenticity on brand trust and SME growth: A CEO perspective. *Journal of World Business*, *48*(3), 340–348. https://doi.org/10.1016/j.jwb.2012.07.018

Enkel, E., Gassmann, O., & Chesbrough, H. W. (2009). Open R&D and open innovation: Exploring the phenomenon. *R & D Management*, *39*(4). http://search.proquest.com/abicomplete/docview/233088215/3B06F135B4064E7DPQ/16?accountid=40695

Ernst, H. (2001). Corporate culture and innovative performance of the firm. *Management of Engineering and Technology, 2001. PICMET'01. Portland International Conference On*, 532–535. http://ieeexplore.ieee.org/abstract/document/952226/

Esposito, E. (2004). Strategic alliances and internationalisation in the aircraft manufacturing industry. *Technological Forecasting and Social Change*, *71*(5), 443–468. https://doi.org/10.1016/S0040-1625(03)00002-7

Ettlie, J. E., Bridges, W. P., & O'Keefe, R. D. (1984). Organization Strategy and Structural Differences for Radical Versus Incremental Innovation. *Management Science*, *30*(6), 682–695. https://doi.org/10.1287/mnsc.30.6.682

Etzkowitz, H., & Leydesdorff, L. (2000). The dynamics of innovation: From National Systems and "Mode 2" to a Triple Helix of university–industry–government relations. *Research Policy*, *29*(2), 109–123. https://doi.org/10.1016/S0048-7333(99)00055-4

Faems, D. (2020). Moving forward quantitative research on innovation management: A call for an inductive turn on using and presenting quantitative research. *R&D Management*, *50*(3), 352–363. https://doi.org/10.1111/radm.12406

Faems, D., Van Looy, B., & Debackere, K. (2005). Interorganizational Collaboration and Innovation: Toward a Portfolio Approach*. *Journal of Product Innovation Management*, *22*(3), 238–250. https://doi.org/10.1111/j.0737-6782.2005.00120.x

Fagerberg, J. (2017). Innovation Policy: Rationales, Lessons and Challenges. *Journal of Economic Surveys*, *31*(2), 497–512. https://doi.org/10.1111/joes.12164

Fazzari, S., Hubbard, R. G., & Petersen, B. (1988). Investment, financing decisions, and tax policy. *The American Economic Review*, *78*(2), 200–205.

Feldman, M. P., & Florida, R. (1994). The geographic sources of innovation: Technological infrastructure and product innovation in the United States. *Annals of the Association of American Geographers*, *84*(2), 210–229.

Fey, C. F., & Birkinshaw, J. (2005). External Sources of Knowledge, Governance Mode, and R&D Performance. *Journal of Management*, *31*(4), 597–621. https://doi.org/10.1177/0149206304272346

Fiske, D. W., & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin*, *112*(3), 393.

Fitzpatrick, I., Bertscher, A., & Gilmore, A. B. (2022). Identifying misleading corporate narratives: The application of linguistic and qualitative methods to commercial determinants of health research. *PLOS Global Public Health*, *2*(11), e0000379. https://doi.org/10.1371/journal.pgph.0000379

Flor, M. L., & Oltra, M. J. (2004). Identification of innovating firms through technological innovation indicators: An application to the Spanish ceramic tile industry. *Research Policy*, *33*(2), 323–336. https://doi.org/10.1016/j.respol.2003.09.009

Folkestad, J., & Gonzalez, R. (2010). Teamwork for Innovation: A Content Analysis of the Highly Read and Highly Cited Literature on Innovation. *Advances in Developing Human Resources*, *12*(1), 115–136. https://doi.org/10.1177/1523422310365486

Foltz, J., Barham, B., & Kim, K. (2000). Universities and agricultural biotechnology patent production. *Agribusiness*, *16*(1), 82–95. https://doi.org/10.1002/(SICI)1520-6297(200024)16:1<82::AID-AGR7>3.0.CO;2-V

Foray, D. (2004). *Economics of knowledge*. MIT press.

Fosfuri, A. (2006). The licensing dilemma: Understanding the determinants of the rate of technology licensing. *Strategic Management Journal*, *27*(12), 1141–1158. https://doi.org/10.1002/smj.562

Frear, C. R., & Metcalf, L. E. (1995). Strategic alliances and technology networks: A study of a cast-products supplier in the aircraft industry. *Industrial Marketing Management*, *24*(5), 379–390. https://doi.org/10.1016/0019-8501(95)00029-A

Freel, M. S. (2003). Sectoral patterns of small firm innovation, networking and proximity. *Research Policy*, *32*(5), 751–770. https://doi.org/10.1016/S0048-7333(02)00084-7

Fritz, K., Schoenmueller, V., & Bruhn, M. (2017). Authenticity in branding – exploring antecedents and consequences of brand authenticity. *European Journal of Marketing*, *51*(2), 324–348. https://doi.org/10.1108/EJM-10-2014-0633

Gammoh, B. S., Voss, K. E., & Chakraborty, G. (2006). Consumer evaluation of brand alliance signals. *Psychology & Marketing*, *23*(6), 465–486. https://doi.org/10.1002/mar.20130

Garcia, R., & Calantone, R. (2002). A critical look at technological innovation typology and innovativeness terminology: A literature review. *Journal of Product Innovation Management*, *19*(2), 110–132. https://doi.org/10.1016/S0737-6782(01)00132-1

Gassmann, O., & Enkel, E. (2004). *Towards a theory of open innovation: Three core process archetypes*. https://www.alexandria.unisg.ch/publications/274

Geroski, P. A. (1990). Innovation, technological opportunity, and market structure. *Oxford Economic Papers*, *42*(3), 586–602.

Geroski, P. A., Machin, S., & Van Reenen, J. (1993). The Profitability of Innovating Firms. *The RAND Journal of Economics*, *24*(2), 198–211. https://doi.org/10.2307/2555757

Gharbi, S., Sahut, J.-M., & Teulon, F. (2014). R&D investments and high-tech firms' stock return volatility. *Technological Forecasting and Social Change*, *88*, 306–312. https://doi.org/10.1016/j.techfore.2013.10.006

Giardini, F., Balliet, D., Power, E. A., Számadó, S., & Takács, K. (2022). Four Puzzles of Reputation-Based Cooperation. *Human Nature*, *33*(1), 43–61. https://doi.org/10.1007/s12110-021-09419-3

Ginzburg, C. (1989). Clues: Roots Aof and Evidential Paradigm. *Clues, Myths and the Historical*.

Godfray, H. C. J. (1991). Signalling of need by offspring to their parents. *Nature*, *352*(6333), 328–330. https://doi.org/10.1038/352328a0

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*(1), 653–671. https://doi.org/10.1007/s11192-014-1434-0

Government of Canada, I. (2020, December 10). *Key Small Business Statistics—2020* [Reports; Related Links; Statistical Reports]. Innovation, Science and Economic Development

Canada. https://ised-isde.canada.ca/site/sme-research-statistics/en/key-small-business-statistics/key-small-business-statistics-2020

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, *144*(4), 517–546. https://doi.org/10.1016/S0022-5193(05)80088-8

Grassly, N. C., Von haeseler, A., & Krakauer, D. C. (2000). Error, Population Structure and the Origin of Diverse Sign Systems. *Journal of Theoretical Biology*, *206*(3), 369–378. https://doi.org/10.1006/jtbi.2000.2133

Gray, E. R., & Balmer, J. M. T. (1998). Managing Corporate Image and Corporate Reputation. *Long Range Planning*, *31*(5), 695–702. https://doi.org/10.1016/S0024-6301(98)00074-0

Gray, S. G., Sütterlin, B., Siegrist, M., & Árvai, J. (2020). The benefit of virtue signaling: Corporate sleight-of-hand positively influences consumers' judgments about "social license to operate". *Journal of Environmental Management*, *260*, 110047.

Grayson, K., & Martinec, R. (2004). Consumer Perceptions of Iconicity and Indexicality and Their Influence on Assessments of Authentic Market Offerings. *Journal of Consumer Research*, *31*(2), 296–312. https://doi.org/10.1086/422109

Greene, W. H. (2008). The econometric approach to efficiency analysis. *The Measurement of Productive Efficiency and Productivity Growth*, *1*(1), 92–250.

Greene, W. H. (2012). *Econometric Analysis*, (7th Edition). Upper Saddle River, NJ: Prentice Hall. https://www.stata.com/bookstore/econometric-analysis/

Greve, H. R. (2003). A Behavioral Theory of R&D Expenditures and Innovations: Evidence from Shipbuilding. *Academy of Management Journal*, *46*(6), 685–702. https://doi.org/10.2307/30040661

Griliches, Z. (1990). *Patent Statistics as Economic Indicators: A Survey* (Working Paper No. 3301). National Bureau of Economic Research. https://doi.org/10.3386/w3301

Griliches, Z. (1994). Productivity, R&D and the Data Constraint. Presidential address, American Economic Association, Boston, January 4, 1994. *American Economic Review*, *84*(1), 115–119.

Griliches, Z. (1998). *R&D and Productivity*. University of Chicago Press. https://ideas.repec.org/b/ucp/bknber/9780226308869.html

Grossman, G. M., & Helpman, E. (1993). *Innovation and Growth in the Global Economy*. MIT Press.

Guilford, T., & Dawkins, M. S. (1995). What are conventional signals? *Animal Behaviour*, *49*(6), 1689–1695. https://doi.org/10.1016/0003-3472(95)90090-X

Guix, M., Ollé, C., & Font, X. (2022). Trustworthy or misleading communication of voluntary carbon offsets in the aviation industry. *Tourism Management*, *88*, 104430. https://doi.org/10.1016/j.tourman.2021.104430

Gulek, C. (1999). *Using Multiple Means of Inquiry To Gain Insight into Classrooms: A Multi-Trait Multi-Method Approach*. https://eric.ed.gov/?id=ED431016

Guo, B., Aveyard, P., Fielding, A., & Sutton, S. (2008). Testing the convergent and discriminant validity of the Decisional Balance Scale of the Transtheoretical Model using the Multi-Trait Multi-Method approach. *Psychology of Addictive Behaviors*, *22*(2), 288–294. https://doi.org/10.1037/0893-164X.22.2.288

Hagedoorn, J., & Cloodt, M. (2003). Measuring innovative performance: Is there an advantage in using multiple indicators? *Research Policy*, *32*(8), 1365–1379. https://doi.org/10.1016/S0048-7333(02)00137-3

Hagedoorn, J., Link, A. N., & Vonortas, N. S. (2000). Research partnerships. *Research Policy*, *29*(4), 567–586. https://doi.org/10.1016/S0048-7333(99)00090-6

Hagel, J., & Brown, J. S. (2008). Creation Nets: Harnessing The Potential Of Open Innovation. *Journal of Service Science (JSS)*, *1*(2), Article 2. https://doi.org/10.19030/jss.v1i2.4293

Hair, J. F., Black, W. C., & Anderson, R. E. (2010). Multivariate Data Analysis (7th Ed.). *Upper Saddle River, N.J.: Pearson Prentice Hall.* https://digitalcommons.kennesaw.edu/facpubs/2925

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). *Multivariate data analysis* (Vol. 5). Prentice hall Upper Saddle River, NJ.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a Silver Bullet. *Journal of Marketing Theory and Practice*, *19*(2), 139–152. https://doi.org/10.2753/MTP1069-6679190202

Hall, B. H. (1990, January 1). The Impact of Corporate Restructuring on Industrial Research and Development. *Brookings*. https://www.brookings.edu/bpea-articles/the-impact-of-corporate-restructuring-on-industrial-research-and-development/

Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2000). *Market value and patent citations: A first look*. national bureau of economic research Cambridge, Mass., USA.

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market Value and Patent Citations. *The RAND Journal of Economics*, *36*(1), 16–38.

Hamilton, T., & Cavello, S. (2021). Ethical product havens in the global diamond trade: Using the Wayback Machine to evaluate ethical market outcomes. *Environment and Planning A: Economy and Space*, 0308518X211029661. https://doi.org/10.1177/0308518X211029661

Hanna, V., & Walsh, K. (2002). Small firm networks: A successful approach to innovation? *R&D Management*, *32*(3), 201–207. https://doi.org/10.1111/1467-9310.00253

Hannan, M. T., & Freeman, J. (1984). Structural inertia and organizational change. *American Sociological Review*, 149–164.

Harari, Y. N. (2014). *Sapiens: A Brief History of Humankind*. Random House.

Harhoff, D., & Körting, T. (1998). Lending relationships in Germany – Empirical evidence from survey data. *Journal of Banking & Finance*, *22*(10), 1317–1353. https://doi.org/10.1016/S0378-4266(98)00061-2

Hauschildt, J., & Salomo, S. (2007). *Innovationsmanagement (4., überarb., erg. Und aktualisierte Aufl.). Vahlens Handbücher der Wirtschafts-und Sozialwissenschaften*. München: Vahlen.

Hausman, J. A., Hall, B. H., & Griliches, Z. (1984). *Econometric Models for Count Data with an Application to the Patents-R&D Relationship* (Working Paper No. 17). National Bureau of Economic Research. https://doi.org/10.3386/t0017

Haziza, D., & Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, *75*(1), 25–43.

Heil, O., & Robertson, T. S. (1991). Toward a theory of competitive market signaling: A research agenda. *Strategic Management Journal*, *12*(6), 403–418. https://doi.org/10.1002/smj.4250120602

Henkel, J. (2006). Selective revealing in open innovation processes: The case of embedded Linux. *Research Policy*, *35*(7), 953–969.

Henningsen, A., & Hamann, J. D. (2007). **systemfit**: A Package for Estimating Systems of Simultaneous Equations in *R*. *Journal of Statistical Software*, *23*(4). https://doi.org/10.18637/jss.v023.i04

Herbig, P., & Milewicz, J. (1995). The relationship of reputation and credibility to brand success. *Journal of Consumer Marketing*, *12*(4), 5–11.

Héroux-Vaillancourt, M., & Beaudry, C. (2018). Are the most innovative Canadian nanotechnology-related firms also the most open? In *Geography, Open Innovation and Entrepreneurship* (pp. 140–189). Edward Elgar Publishing. https://www.elgaronline.com/display/edcoll/9781786439895/9781786439895.00010.xml

Héroux-Vaillancourt, M., Beaudry, C., & Rietsch, C. (2020). Using web content analysis to create innovation indicators – What do we really measure? *Quantitative Science Studies*, 1–37. https://doi.org/10.1162/qss_a_00086

Herrouz, A., Khentout, C., & Djoudi, M. (2013). Overview of Web Content Mining Tools. *ArXiv:1307.1024 [Cs]*. http://arxiv.org/abs/1307.1024

Herzog, P. (2011). Innovation culture. In *Open and Closed Innovation* (pp. 59–82). Springer. http://link.springer.com/chapter/10.1007/978-3-8349-6165-5_3

Hicks, D. (1995). Published Papers, Tacit Competencies and Corporate Management of the Public/Private Character of Knowledge. *Industrial and Corporate Change*, *4*(2), 401–424. https://doi.org/10.1093/icc/4.2.401

Hicks, D., & Hegde, D. (2005). Highly innovative small firms in the markets for technology. *Research Policy*, *34*(5), 703–716.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hitt, M. A., Hoskisson, R. E., & Kim, H. (1997). International Diversification: Effects on Innovation and Firm Performance in Product-Diversified Firms. *Academy of Management Journal*, *40*(4), 767–798. https://doi.org/10.2307/256948

Holt, D. B. (2002). Why Do Brands Cause Trouble? A Dialectical Theory of Consumer Culture and Branding. *Journal of Consumer Research*, *29*(1), 70–90. https://doi.org/10.1086/339922

Horibe, F. (2016). *Creating the innovation culture: Leveraging visionaries, dissenters, and other useful troublemakers in your organization*. VisionArts Inc. https://books.google.ca/books?hl=en&lr=&id=vdHuCwAAQBAJ&oi=fnd&pg=PT6&dq=Groupthink+innovation+culture&ots=h0il33KxuQ&sig=W7kSid6l_ssqYFLoo-miBrniUAw

Hsu, P.-H., Tian, X., & Xu, Y. (2014). Financial development and innovation: Cross-country evidence. *Journal of Financial Economics*, *112*(1), 116–135. https://doi.org/10.1016/j.jfineco.2013.12.002

Hurley, R. F., & Hult, G. T. M. (1998). Innovation, Market Orientation, and Organizational Learning: An Integration and Empirical Examination. *Journal of Marketing*, *62*(3), 42–54. https://doi.org/10.1177/002224299806200303

Hwang, D. (2010). *Ranking the nations on nanotech | Solid State Technology*. http://electroiq.com/blog/2010/08/ranking-the-nations/

Hyun Kim, J. (2012). A Hyperlink and Semantic Network Analysis of the Triple Helix (University-Government-Industry): The Interorganizational Communication Structure of Nanotechnology. *Journal of Computer-Mediated Communication*, *17*(2), 152–170. https://doi.org/10.1111/j.1083-6101.2011.01564.x

Ivanov, C.-I., & Avasilcăi, S. (2014). Performance Measurement Models: An Analysis for Measuring Innovation Processes Performance. *Procedia - Social and Behavioral Sciences*, *124*, 397–404. https://doi.org/10.1016/j.sbspro.2014.02.501

Jackson, J. E. (2005). *A User's Guide to Principal Components*. John Wiley & Sons.

Janis, I. L. (1971). Groupthink. *Psychology Today*, *5*(6), 43–46.

Jaworski, B. J., & Kohli, A. K. (1993). Market Orientation: Antecedents and Consequences. *Journal of Marketing*, *57*(3), 53–70. https://doi.org/10.1177/002224299305700304

Jenifer, M., Kassahun, A., & Baharu, A. (2016). *Web Mining Overview, Taxonomy, Literature Survey: A Review*. *14*(11), 10.

Johnson, W. H. A., & Filippini, R. (2009). Internal vs. External Collaboration: What Works. *Research-Technology Management*, *52*(3), 15–17. https://doi.org/10.1080/08956308.2009.11657564

Jones, M. D., Hutcheson, S., & Camba, J. D. (2021). Past, present, and future barriers to digital transformation in manufacturing: A review. *Journal of Manufacturing Systems*, *60*, 936–948. https://doi.org/10.1016/j.jmsy.2021.03.006

Jordan, J., & Lowe, J. (2004). Protecting Strategic Knowledge: Insights from Collaborative Agreements in the Aerospace Sector. *Technology Analysis & Strategic Management*, *16*(2), 241–259. https://doi.org/10.1080/09537320410001682900

Jung, J. C., & Sharon, E. (2019). The Volkswagen emissions scandal and its aftermath. *Global Business and Organizational Excellence*, *38*(4), 6–15. https://doi.org/10.1002/joe.21930

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36. https://doi.org/10.1007/BF02291575

Kalil, T. A. (2005). Nanotechnology and the Valley of Death. *Nanotechnology Law & Business*, *2*, 265.

Kaminski, P. C., de Oliveira, A. C., & Lopes, T. M. (2008). Knowledge transfer in product development processes: A case study in small and medium enterprises (SMEs) of the metal-mechanic sector from São Paulo, Brazil. *Technovation*, *28*(1), 29–36. https://doi.org/10.1016/j.technovation.2007.07.001

Katz, J. S. (2006). Indicators for complex innovation systems. *Research Policy*, *35*(7), 893–909. https://doi.org/10.1016/j.respol.2006.03.007

Katz, J. S., & Cothey, V. (2006). Web indicators for complex innovation systems. *Research Evaluation*, *15*(2), 85–95. https://doi.org/10.3152/147154406781775922

Kayser, V., & Blind, K. (2017). Extending the knowledge base of foresight: The contribution of text mining. *Technological Forecasting and Social Change*, *116*, 208–215. https://doi.org/10.1016/j.techfore.2016.10.017

Keizer, J. A., Dijkstra, L., & Halman, J. I. M. (2002). Explaining innovative efforts of SMEs.: An exploratory survey among SMEs in the mechanical and electrical engineering sector in The Netherlands. *Technovation*, *22*(1), 1–13. https://doi.org/10.1016/S0166-4972(00)00091-2

Kerr, W. R., & Nanda, R. (2015). Financing Innovation. *Annual Review of Financial Economics*, *7*(1), 445–462. https://doi.org/10.1146/annurev-financial-111914-041825

Kim, J., Lee, S., & Marschke, G. (2014). Impact of university scientists on innovations in nanotechnology. In S. Ahn, B. Hall, & K. Lee, *Intellectual Property for Economic*

Development (pp. 141–158). Edward Elgar Publishing. http://www.elgaronline.com/view/9781782548041.00012.xml

Kinne, J., & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics*, *125*(3), 2011–2041. https://doi.org/10.1007/s11192-020-03726-9

Kinne, J., & Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. *PLOS ONE*, *16*(4), e0249071. https://doi.org/10.1371/journal.pone.0249071

Klaassen, C. A. J., Mokveld, P. J., & van Es, B. (2000). Squared skewness minus kurtosis bounded by 186/125 for unimodal distributions. *Statistics & Probability Letters*, *50*(2), 131–135. https://doi.org/10.1016/S0167-7152(00)00090-0

Klein, G., Hernandez, F., Nguyen, V., & Senellart, J. (2020). The OpenNMT Neural Machine Translation Toolkit: 2020 Edition. *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, 102–109. https://aclanthology.org/2020.amta-research.9

Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., & Rush, A. (2018). OpenNMT: Neural Machine Translation Toolkit. *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, 177–184. https://aclanthology.org/W18-1817

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *Proceedings of ACL 2017, System Demonstrations*, 67–72. https://aclanthology.org/P17-4012

Kleinknecht, A., Montfort, K. V., & Brouwer, E. (2002). The Non-Trivial Choice between Innovation Indicators. *Economics of Innovation and New Technology*, *11*(2), 109–121. https://doi.org/10.1080/10438590210899

Kleinschmidt, E. J., Brentani, U. D., & Salomo, S. (2007). Performance of Global New Product Development Programs: A Resource-Based View. *Journal of Product Innovation Management*, *24*(5), 419–441. https://doi.org/10.1111/j.1540-5885.2007.00261.x

Klette, T. J., Møen, J., & Griliches, Z. (2000). Do subsidies to commercial R&D reduce market failures? Microeconometric evaluation studies1We have benefited from comments by Tore Nilssen, John van Reenen and participants at the NBER productivity meeting in December 1998. This project has received partial financial support from the Research Council of Norway.1. *Research Policy*, *29*(4), 471–495. https://doi.org/10.1016/S0048-7333(99)00086-4

Kock, N. (2014). Advanced Mediating Effects Tests, Multi-Group Analyses, and Measurement Model Assessments in PLS-Based SEM: *International Journal of e-Collaboration*, *10*(1), 1–13. https://doi.org/10.4018/ijec.2014010101

Kock, N. (2015). PLS-based SEM Algorithms: The Good Neighbor Assumption, Collinearity, and Nonlinearity. *Information Management and Business Review*, *7*(2), 113–130. https://doi.org/10.22610/imbr.v7i2.1146

Kock, N. (2018). SHOULD BOOTSTRAPPING BE USED IN PLS-SEM? TOWARD STABLE P-VALUE CALCULATION METHODS. *Journal of Applied Structural Equation Modeling*, *2*(1), 1–12. https://doi.org/10.47263/JASEM.2(1)02

Kock, N. (2022). *WarpPLS User Manual: Version 8.0*. 149.

Koen, P. A., Bertels, H. M. J., & Kleinschmidt, E. J. (2014). Managing the Front End of Innovation—Part II: Results from a Three-Year Study. *Research-Technology Management*, *57*(3), 25–35. https://doi.org/10.5437/08956308X5703199

Koestler, A. (1964). *The act of creation* (Penguin Books).

Kor, Y. Y., & Mahoney, J. T. (2005). How dynamics, management, and governance of resource deployments influence firm-level performance. *Strategic Management Journal*, *26*(5), 489–496. https://doi.org/10.1002/smj.459

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Sage Publications.

Kuratko, D. F., Covin, J. G., & Hornsby, J. S. (2014). Why implementing corporate innovation is so difficult. *Business Horizons*, *57*(5), 647–655. https://doi.org/10.1016/j.bushor.2014.05.007

Kuznets, S. (1962). Inventive activity: Problems of definition and measurement. In *The rate and direction of inventive activity: Economic and social factors* (pp. 19–52). Princeton University Press. http://www.nber.org/chapters/c2112.pdf

Lachmann, M., Számadó, S., & Bergstrom, C. T. (2001). Cost and conflict in animal signals and human language. *Proceedings of the National Academy of Sciences*, *98*(23), 13189–13194. https://doi.org/10.1073/pnas.231216498

Lakhani, K. R. (2008). InnoCentive. Com (A). *Harvard Business School Case*, *608–170*. http://netgear.rohidekar.com:8025/SCU-MBA/MGMT716/HBSP%20(Harvard%20Business%20School%20cases)/Innocentive%20-%2022324790.pdf

Lane, P. J., & Lubatkin, M. (1998). Relative absorptive capacity and interorganizational learning. *Strategic Management Journal*, *19*(5), 461–477.

Lane, P. J., Salk, J. E., & Lyles, M. A. (2001). Absorptive capacity, learning, and performance in international joint ventures. *Strategic Management Journal*, *22*(12), 1139–1161.

Laursen, K., & Salter, A. (2004). Searching high and low: What types of firms use universities as a source of innovation? *Research Policy*, *33*(8), 1201–1215. https://doi.org/10.1016/j.respol.2004.07.004

Laursen, K., & Salter, A. (2006a). My precious technology: The role of legal appropriability strategy in shaping innovative performance. *Tanaka Business School, Imperial College London, Working Paper*.

Laursen, K., & Salter, A. (2006b). Open for innovation: The role of openness in explaining innovation performance among U.K. manufacturing firms. *Strategic Management Journal, 27*(2), 131–150. https://doi.org/10.1002/smj.507

Laursen, K., & Salter, A. J. (2014). The paradox of openness: Appropriability, external search and collaboration. *Research Policy*, *43*(5), 867–878. https://doi.org/10.1016/j.respol.2013.10.004

Lee, C.-J., Lee, S., Jhon, M. S., & Shin, J. (2013). Factors influencing nanotechnology commercialisation: An empirical analysis of nanotechnology firms in South Korea. *Journal of Nanoparticle Research*, *15*(2), 1444. https://doi.org/10.1007/s11051-013-1444-5

Lee, S., Park, G., Yoon, B., & Park, J. (2010). Open innovation in SMEs—An intermediated network model. *Research Policy*, *39*(2), 290–300. https://doi.org/10.1016/j.respol.2009.12.009

Levidow, L., & Papaioannou, T. (2018). Which inclusive innovation? Competing normative assumptions around social justice. *Innovation and Development*, *8*(2), 209–226. https://doi.org/10.1080/2157930X.2017.1351605

Levin, R. C., Klevorick, A. K., Nelson, R. R., Winter, S. G., Gilbert, R., & Griliches, Z. (1987). Appropriating the returns from industrial research and development. *Brookings Papers on Economic Activity*, *1987*(3), 783–831.

Li, Y., Arora, S. K., Youtie, J., & Shapira, P. (2018). Using web mining to explore Triple Helix influences on growth in small and mid-size firms. *Technovation*, *76–77*, 3–14. https://doi.org/10.1016/j.technovation.2016.01.002

Li, Y., Youtie, J., & Shapira, P. (2015). Why do technology firms publish scientific papers? The strategic use of science by small and midsize enterprises in nanotechnology. *The Journal of Technology Transfer*, *40*(6), 1016–1033. https://doi.org/10.1007/s10961-014-9391-6

Libaers, D., Hicks, D., & Porter, A. L. (2016). A taxonomy of small firm technology commercialisation. *Industrial and Corporate Change*, *25*(3), 371–405. https://doi.org/10.1093/icc/dtq039

Lieberman, M. B., & Montgomery, D. B. (1988). First-mover advantages. *Strategic Management Journal*, *9*(S1), 41–58. https://doi.org/10.1002/smj.4250090706

Little, R. J. A. (1986). Survey Nonresponse Adjustments for Estimates of Means. *International Statistical Review / Revue Internationale de Statistique*, *54*(2), 139–157. https://doi.org/10.2307/1403140

Liyanage, S., Greenfield, P. F., & Don, R. (1999). Towards a fourth generation R&D management model-research networks in knowledge management. *International Journal of Technology Management*, *18*(3–4), 372–393.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables* (Vol. 7). SAGE Publications, Inc. https://us.sagepub.com/en-us/nam/regression-models-for-categorical-and-limited-dependent-variables/book6071

Long, J. S., & Freese, J. (2001). FITSTAT: Stata module to compute fit statistics for single equation regression models. *Statistical Software Components*. https://ideas.repec.org//c/boc/bocode/s407201.html

Lugtig, P. (2017). The Relative Size of Measurement Error and Attrition Error in a Panel Survey. Comparing Them With New Multi-Trait Multi-Method Model. *Survey Research Methods*, *11*(4), Article 4. https://doi.org/10.18148/srm/2017.v11i4.7170

Lukas, B. A., & Ferrell, O. C. (2000). The effect of market orientation on product innovation. *Journal of the Academy of Marketing Science*, *28*(2), 239–247.

Lyon, T. P., & Maxwell, J. W. (2011). Greenwash: Corporate Environmental Disclosure under Threat of Audit. *Journal of Economics & Management Strategy*, *20*(1), 3–41. https://doi.org/10.1111/j.1530-9134.2010.00282.x

Maas, C. J., Lensvelt-Mulders, G. J., & Hox, J. J. (2009). A multilevel multitrait-multimethod analysis. *Methodology*, *5*(3), 72–77.

Maccoby, M. (2006). Creating collaboration. *Research-Technology Management*, *49*(6), 60–62.

Makkonen, T., & van der Have, R. P. (2013). Benchmarking regional innovative performance: Composite measures and direct innovation counts. *Scientometrics*, *94*(1), 247–262. https://doi.org/10.1007/s11192-012-0753-2

Markham, V. W. (2015). *Planning the corporate reputation*. Routledge.

Matthias Inauen & Andrea Schenker-Wicki. (2011). The impact of outside-in open innovation on innovation performance. *European Journal of Innovation Management*, *14*(4), 496–520. https://doi.org/10.1108/14601061111174934

Mavlanova, T., Benbunan-Fich, R., & Koufaris, M. (2012). Signaling theory and information asymmetry in online commerce. *Information & Management*, *49*(5), 240–247. https://doi.org/10.1016/j.im.2012.05.004

Mazzoleni, R., & Nelson, R. R. (1998). The benefits and costs of strong patent protection: A contribution to the current debate. *Research Policy*, *27*(3), 273–284. https://doi.org/10.1016/S0048-7333(98)00048-1

Mazzucato, M. (2011). The entrepreneurial state. *Soundings*, *49*(49), 131–142.

Mazzucato, M. (2021). Financing the Green New Deal. *Nature Sustainability*, 1–2. https://doi.org/10.1038/s41893-021-00828-x

McAndrew, F. T. (2018). Costly Signaling Theory. In T. K. Shackelford & V. A. Weekes-Shackelford (Eds.), *Encyclopedia of Evolutionary Psychological Science* (pp. 1–8). Springer International Publishing. https://doi.org/10.1007/978-3-319-16999-6_3483-1

McNeil, R. D., Lowe, J., Mastroianni, T., Cronin, J., & Ferk, D. (2007). *Barriers to nanotechnology commercialisation* (pp. 1–57). College of Business and Management, The University of Illinois at Springfield 62703-5407. http://www.wimb.fink.rs/docs/Report-BarriersNanotechnologyCommercialization.pdf

Merges, R. P. (1999). Institutions for intellectual property transactions: The case of patent pools. *University of California at Berkeley Working Paper*, 1–74.

Meuleman, M., & De Maeseneire, W. (2012). Do R&D subsidies affect SMEs' access to external financing? *Research Policy*, *41*(3), 580–591. https://doi.org/10.1016/j.respol.2012.01.001

Meyer, M. (2000). Does science push technology? Patents citing scientific literature. *Research Policy*, *29*(3), 409–434. https://doi.org/10.1016/S0048-7333(99)00040-2

Michie, J. (1998). Introduction. The Internationalisation of the Innovation Process. *International Journal of the Economics of Business*, *5*(3), 261–277. https://doi.org/10.1080/13571519884387

Migol, E., Tretyak, O., & Holm, A. B. (2018). Business Model Design Themes, Value Propositions and Firm Performance. *Journal of Business Models*, *6*(2), Article 2. https://doi.org/10.5278/ojs.jbm.v6i2.2464

Miller, K. D., & Bromiley, P. (1990). Strategic risk and corporate performance: An analysis of alternative risk measures. *Academy of Management Journal*, *33*(4), 756–779.

Miller, T., & Del Carmen Triana, M. (2009). Demographic Diversity in the Boardroom: Mediators of the Board Diversity–Firm Performance Relationship. *Journal of Management Studies*, *46*(5), 755–786. https://doi.org/10.1111/j.1467-6486.2009.00839.x

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. New York: Academic Press.

Minguillo, D., & Thelwall, M. (2012). Mapping the network structure of science parks: An exploratory study of cross-sectoral interactions reflected on the web. *Aslib Proceedings*, *64*(4), 332–357. https://doi.org/10.1108/00012531211244716

Mirończuk, M. M., & Protasiewicz, J. (2016). A Diversified Classification Committee for Recognition of Innovative Internet Domains. In S. Kozielski, D. Mrozek, P. Kasprowski, B. Małysiak-Mrozek, & D. Kostrzewa (Eds.), *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery* (Vol. 613,

pp. 368–383). Springer International Publishing. https://doi.org/10.1007/978-3-319-34099-9_29

Mirończuk, M. M., & Protasiewicz, J. (2020). Recognising innovative companies by using a diversified stacked generalisation method for website classification. *Applied Intelligence*, *50*(1), 42–60. https://doi.org/10.1007/s10489-019-01509-1

Mowery, D., & Rosenberg, N. (1979). The influence of market demand upon innovation: A critical review of some recent empirical studies. *Research Policy*, *8*(2), 102–153. https://doi.org/10.1016/0048-7333(79)90019-2

Muller, P., & Pénin, J. (2006). Why do firms disclose knowledge and how does it matter? *Journal of Evolutionary Economics*, *16*(1), 85–108. https://doi.org/10.1007/s00191-005-0009-x

Murphy, J., Hashim, N. H., & O'Connor, P. (2007). Take Me Back: Validating the Wayback Machine. *Journal of Computer-Mediated Communication*, *13*(1), 60–75. https://doi.org/10.1111/j.1083-6101.2007.00386.x

Nagaoka, S., Motohashi, K., & Goto, A. (2010). Chapter 25—Patent Statistics as an Innovation Indicator. In B. H. Hall & N. Rosenberg (Eds.), *Handbook of the Economics of Innovation* (Vol. 2, pp. 1083–1127). North-Holland. https://doi.org/10.1016/S0169-7218(10)02009-5

Nathan, M., & Rosso, A. (2017). *Innovative Events* (SSRN Scholarly Paper No. 3085935). https://doi.org/10.2139/ssrn.3085935

National Nanotechnology Coordination Office. (2017). *Supplement to the President's 2018 Budget* (p. 86).

Nelissen, R. M. A., & Meijers, M. H. C. (2011). Social benefits of luxury brands as costly signals of wealth and status. *Evolution and Human Behavior*, *32*(5), 343–355. https://doi.org/10.1016/j.evolhumbehav.2010.12.002

Nelson, P. R. C., Taylor, P. A., & MacGregor, J. F. (1996). Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics and Intelligent Laboratory Systems*, *35*(1), 45–65. https://doi.org/10.1016/S0169-7439(96)00007-X

Nowak, M. A., & Krakauer, D. C. (1999). The evolution of language. *Proceedings of the National Academy of Sciences*, *96*(14), 8028–8033. https://doi.org/10.1073/pnas.96.14.8028

Nowak, M. A., Plotkin, J. B., & Krakauer, D. C. (1999). The Evolutionary Language Game. *Journal of Theoretical Biology*, *200*(2), 147–162. https://doi.org/10.1006/jtbi.1999.0981

Nuvolari, A. (2004). Collective invention during the British Industrial Revolution: The case of the Cornish pumping engine. *Cambridge Journal of Economics*, *28*(3), 347–363. https://doi.org/10.1093/cje/28.3.347

Nyagadza, B., Kabonga, I., Hlungwani, P. M., Chigora, F., & Rukasha, T. (2021). The fourth Industrial Revolution and the Imperatives of African States' Sovereignty over natural resources. *From# RhodesMustFall Movements To# HumansMustFall Movements: African*

Liberation Movements in the Age of the Transhumanist Geographies of Death'. Langaa Research & Publishing Common Initiative Group. ISBN: 97809956551941.

Nyagadza, B., Kadembo, E. M., & Makasi, A. (2021). When corporate brands tell stories: A signalling theory perspective. *Cogent Psychology*, *8*(1), 1897063. https://doi.org/10.1080/23311908.2021.1897063

Oakey, R. P. (2013). Open innovation and its relevance to industrial research and development: The case of high-technology small firms. *International Small Business Journal*, *31*(3), 319–336. https://doi.org/10.1177/0266242612458942

O'Cass, A., & Viet Ngo, L. (2007). Market orientation versus innovative culture: Two routes to superior brand performance. *European Journal of Marketing*, *41*(7/8), 868–887. https://doi.org/10.1108/03090560710752438

OCDE, & Communities, S. O. of the E. (2005). *Oslo Manual*. https://www.oecd-ilibrary.org/content/publication/9789264013100-en

OCDE, & Eurostat. (2019). *Oslo Manual 2018*. https://www.oecd-ilibrary.org/content/publication/9789264304604-en

O'Connor, G. C., & DeMartino, R. (2006). Organizing for radical innovation: An exploratory study of the structural aspects of RI management systems in large established firms. *Journal of Product Innovation Management*, *23*(6), 475–497.

OECD. (2015). *Frascati Manual 2015*. https://www.oecd-ilibrary.org/content/publication/9789264239012-en

OECD & Eurostat. (2005). *Oslo Manual*. OECD Publishing. http://www.oecd-ilibrary.org/science-and-technology/oslo-manual_9789264013100-en

OECD, & Eurostat. (2018). *Oslo Manual 2018*. https://www.oecd-ilibrary.org/content/publication/9789264304604-en

OECD, O. M. (1997). *The Measurement of scientific and technological activites. Proposed Guidelines for Collecting and Interpeting Technological Innovation Data*. OECD/Eurostat, Paris.

Oinas, P. (2005). Finland: A success story? *European Planning Studies*, *13*(8), 1227–1244. https://doi.org/10.1080/09654310500336618

Ortiz de Guinea, A., Titah, R., & Léger, P.-M. (2013). Measure for Measure: A two study multi-trait multi-method investigation of construct validity in IS research. *Computers in Human Behavior*, *29*(3), 833–844. https://doi.org/10.1016/j.chb.2012.12.009

Parker, H. (2000). Interfirm collaboration and the new product development process. *Industrial Management & Data Systems*, *100*(6), 255–260. https://doi.org/10.1108/02635570010301179

Parthasarthy, R., & Hammond, J. (2002). Product innovation input and outcome: Moderating effects of the innovation process. *Journal of Engineering and Technology Management*, *19*(1), 75–91. https://doi.org/10.1016/S0923-4748(01)00047-9

Pavitt, K. (1984). Sectoral patterns of technical change: Towards a taxonomy and a theory. *Research Policy*, *13*(6), 343–373. https://doi.org/10.1016/0048-7333(84)90018-0

Pavitt, K. (1985). Patent statistics as indicators of innovative activities: Possibilities and problems. *Scientometrics*, *7*(1–2), 77–99. https://doi.org/10.1007/BF02020142

Peter, J. P., & Churchill, G. A. (1986). Relationships among Research Design Choices and Psychometric Properties of Rating Scales: A Meta-Analysis. *Journal of Marketing Research*, *23*(1), 1–10. https://doi.org/10.2307/3151771

Petter, S., Straub, D., & Rai, A. (2007). Specifying Formative Constructs in Information Systems Research. *MIS Quarterly*, *31*(4), 623–656.

Piller, F., & West, J. (2014). Firms, users, and innovation. *New Frontiers in Open Innovation*, *29*(1), 29–49.

Pinker, S. (2003). *The language instinct: How the mind creates language*. Penguin UK.

Pisano, G. P. (1997). *The Development Factory: Unlocking the Potential of Process Innovation*. Harvard Business Press.

Podmetina, D., Volchek, D., Dąbrowska, J., & Fiegenbaum, I. (2013). Human resource practices and open innovation. *International Journal of Innovation Management*, *17*(06), 1340019. https://doi.org/10.1142/S1363919613400197

Price, L., & Thelwall, M. (2005). The clustering power of low frequency words in academic Webs. *Journal of the American Society for Information Science and Technology*, *56*(8), 883–888. https://doi.org/10.1002/asi.20177

Puaschunder, J. M. (2019). *Artificial Intelligence Market Disruption* (SSRN Scholarly Paper No. 3398470). https://doi.org/10.2139/ssrn.3398470

Ramdani, A. (2014). *Revue systématique de la littérature sur les mesures de la collaboration inter-organisationnelle dans un contexte d'innovation* [Masters, École Polytechnique de Montréal]. https://publications.polymtl.ca/1624/

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. *Proceedings of the First Instructional Conference on Machine Learning*, *242*(1), 29–48.

Reinig, B. A., Briggs, R. O., & Nunamaker, J. F. (2007). On the Measurement of Ideation Quality. *Journal of Management Information Systems*, *23*(4), 143–161. https://doi.org/10.2753/MIS0742-1222230407

Richardson, H. A., Simmering, M. J., & Sturman, M. C. (2009). A Tale of Three Perspectives: Examining Post Hoc Statistical Techniques for Detection and Correction of Common

Method Variance. *Organizational Research Methods*, *12*(4), 762–800. https://doi.org/10.1177/1094428109332834

Rietsch, C. (2015). *Étude exploratoire de la commercialisation des nanotechnologies au Canada: L'utilisation de la fouille de données de contenu Web* [Masters, École Polytechnique de Montréal]. https://publications.polymtl.ca/1952/

Ritala, P., Schneider, S., & Michailova, S. (2020). Innovation management research methods: Embracing rigor and diversity. *R&D Management*, *50*(3), 297–308. https://doi.org/10.1111/radm.12414

Rivette, K. G., & Kline, D. (2000). Rembrandts in the Attic: Unlocking the Hidden Value of Patents. *Boston: Harvard Business School Press.*

Rohatgi, V. K., & Székely, G. J. (1989). Sharp inequalities between skewness and kurtosis. *Statistics & Probability Letters*, *8*(4), 297–299. https://doi.org/10.1016/0167-7152(89)90035-7

Roin, B. N. (2013). The Case for Tailoring Patent Awards Based on Time-to-Market. *UCLA Law Review*, *61*, 672.

Roja, A. I., & Nastase, M. (2013). Leveraging Organizational Capabilities through Collaboration and Collaborative Competitive Advantage. *Revista de Management Comparat International*, *14*(3), 359–366.

Rosenberg, N. (1990). Why do firms do basic research (with their own money)? *Research Policy*, *19*(2), 165–174.

Rothwell, R. (1989). Small firms, innovation and industrial change. *Small Business Economics*, 51–64.

Roussel, P. A., Saad, K. N., & Erickson, T. J. (1991). *Third generation r&d*. Harvard Business School Press.

Santarelli, E. (1991). Asset Specificity, R&D Financing, And The Signalling Properties of The Firm's Financial Structure. *Economics of Innovation and New Technology*, *1*(4), 279–294. https://doi.org/10.1080/10438599100000008

Särdnal, C., Swensson, B., & Wretman, J. (1992). Model Assisted Survey Sampling. *Springer, New York*, 1–694. https://doi.org /10.1007/978-1-4612-4378-6

Särndal, C. E., Swensson, B., & Wretman, J. (1992). Model assisted survey sampling Springer. *New York*.

Schein, E. H. (2004). *Organizational Culture and Leadership*. Jossey Bass Publishers.

Scherer, F. M. (1986). Innovation and Growth: Schumpeterian Perspectives. *MIT Press Books*, *1*. https://ideas.repec.org//b/mtp/titles/0262691027.html

Schmoch, U. (2007). Double-boom cycles and the comeback of science-push and market-pull. *Research Policy*, *36*(7), 1000–1015. https://doi.org/10.1016/j.respol.2006.11.008

Schrage, M. (1990). *Shared minds The new technologies of collaboration*.

Schumpeter, J. A. (1942). *Capitalism, socialism and democracy*. New York. Harper & Row. In Gunderson. https://msuweb.montclair.edu/~lebelp/SchumpeterChapter7.pdf

Scotchmer, S. (1991). Standing on the shoulders of giants: Cumulative research and the patent law. *Journal of Economic Perspectives*, *5*(1), 29–41.

Sebeok, T. A. (2001). *Signs: An introduction to semiotics*. University of Toronto Press.

Shafique, M. (2013). Thinking inside the box? Intellectual structure of the knowledge base of innovation research (1988–2008). *Strategic Management Journal*, *34*(1), 62–93. https://doi.org/10.1002/smj.2002

Shapira, P., Gök, A., & Salehi, F. (2016). Graphene enterprise: Mapping innovation and business development in a strategic emerging technology. *Journal of Nanoparticle Research*, *18*(9), 269. https://doi.org/10.1007/s11051-016-3572-1

Siano, A., Vollero, A., Conte, F., & Amabile, S. (2017). "More than words": Expanding the taxonomy of greenwashing after the Volkswagen scandal. *Journal of Business Research*, *71*, 27–37. https://doi.org/10.1016/j.jbusres.2016.11.002

Siguaw, J. A., Simpson, P. M., & Enz, C. A. (2006). Conceptualizing Innovation Orientation: A Framework for Study and Integration of Innovation Research*. *Journal of Product Innovation Management*, *23*(6), 556–574. https://doi.org/10.1111/j.1540-5885.2006.00224.x

Siqueira, J. R., Peña-García, N., ter Horst, E., Molina, G., & Villamil, M. (2021). The Role of Brand Commitment in the Retail Sector: The Relation with Open Innovation. *Journal of Open Innovation: Technology, Market, and Complexity*, *7*(2), 154. https://doi.org/10.3390/joitmc7020154

Song Shin, H. (2003). Disclosures and Asset Returns. *Econometrica*, *71*(1), 105–133. https://doi.org/10.1111/1468-0262.00391

Sood, A., & Tellis, G. J. (2009). Do Innovations Really Pay Off? Total Stock Market Returns to Innovation. *Marketing Science*, *28*(3), 442–456. https://doi.org/10.1287/mksc.1080.0407

Sørensen, J. B., & Stuart, T. E. (2000). Aging, Obsolescence, and Organizational Innovation. *Administrative Science Quarterly*, *45*(1), 81–112. https://doi.org/10.2307/2666980

Spence, M. (1973). Job Market Signaling. *The Quarterly Journal of Economics*, *87*(3), 355–374. https://doi.org/10.2307/1882010

Spence, M. (1974). *Market signaling: Informational transfer in hiring and related screening processes*. cambridge: harvard university press.

Squicciarini, M., Dernis, H., & Criscuolo, C. (2013). *Measuring Patent Quality*. https://www.oecd-ilibrary.org/content/paper/5k4522wkw1r8-en

StataCorp. (2019). *Stata 16 Base Reference Manual.* College Station, TX: Stata Press.

Straub, D., & Burton-Jones, A. (2007). Veni, Vidi, Vici: Breaking the TAM Logjam. *Journal of the Association for Information Systems; Atlanta*, *8*(4), 223–229.

Straub, D., Limayem, M., & Karahanna-Evaristo, E. (1995). Measuring System Usage: Implications for IS Theory Testing. *Management Science*, *41*(8), 1328–1342. https://doi.org/10.1287/mnsc.41.8.1328

Stuart, D., & Thelwall, M. (2006). Investigating triple helix relationships using URL citations: A case study of the UK West Midlands automobile industry. *Research Evaluation*, *15*(2), 97–106. https://doi.org/10.3152/147154406781775968

Stubbs, W. (2019). Strategies, practices, and tensions in managing business model innovation for sustainability: The case of an Australian BCorp. *Corporate Social Responsibility and Environmental Management*, *26*(5), 1063–1072. https://doi.org/10.1002/csr.1786

Summers, K., Speed, M. P., Blount, J. D., & Stuckert, A. M. M. (2015). Are aposematic signals honest? A review. *Journal of Evolutionary Biology*, *28*(9), 1583–1599. https://doi.org/10.1111/jeb.12676

Sundgren, M., Dimenäs, E., Gustafsson, J.-E., & Selart, M. (2005). Drivers of organizational creativity: A path model of creative climate in pharmaceutical R&D. *R&D Management*, *35*(4), 359–374. https://doi.org/10.1111/j.1467-9310.2005.00395.x

Számadó, S. (2011). The cost of honesty and the fallacy of the handicap principle. *Animal Behaviour*, *81*(1), 3–10. https://doi.org/10.1016/j.anbehav.2010.08.022

Számadó, S., Balliet, D., Giardini, F., Power, E. A., & Takács, K. (2021). The language of cooperation: Reputation and honest signalling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1838), 20200286. https://doi.org/10.1098/rstb.2020.0286

Szambelan, S., Jiang, Y., & Mauer, R. (2020). Breaking through innovation barriers: Linking effectuation orientation to innovation performance. *European Management Journal*, *38*(3), 425–434. https://doi.org/10.1016/j.emj.2019.11.001

Szulanski, G. (1996). Exploring internal stickiness: Impediments to the transfer of best practice within the firm. *Strategic Management Journal*, *17*(S2), 27–43.

Tchernichovski, O., Parra, L. C., Fimiarz, D., Lotem, A., & Conley, D. (2019). Crowd wisdom enhanced by costly signaling in a virtual rating system. *Proceedings of the National Academy of Sciences*, *116*(15), 7256–7265. https://doi.org/10.1073/pnas.1817392116

Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy*, *15*(6), 285–305. https://doi.org/10.1016/0048-7333(86)90027-2

Thelwall, M. (2009). Introduction to Webometrics: Quantitative Web Research for the Social Sciences. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *1*(1), 1–116. https://doi.org/10.2200/S00176ED1V01Y200903ICR004

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, *62*(2), 406–418. https://doi.org/10.1002/asi.21462

Thomsen, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. *Statistisk Tidskrift*, *4*, 278–283.

Tian Xie, F., & Johnston, W. J. (2004). Strategic alliances: Incorporating the impact of e-business technological innovations. *Journal of Business & Industrial Marketing*, *19*(3), 208–222. https://doi.org/10.1108/08858620410531342

Tibbetts, E. A., & Izzo, A. (2010). Social Punishment of Dishonest Signalers Caused by Mismatch between Signal and Behavior. *Current Biology*, *20*(18), 1637–1640. https://doi.org/10.1016/j.cub.2010.07.042

Toivonen, T. (2016). What is the Social Innovation Community? Conceptualizing an Emergent Collaborative Organization. *Journal of Social Entrepreneurship*, *7*(1), 49–73. https://doi.org/10.1080/19420676.2014.997779

Torkkeli, M., Kotonen, T., & Ahonen, P. (2007). Regional open innovation system as a platform for SMEs: A survey. *International Journal of Foresight and Innovation Policy*, *3*(4), 336–350.

Torregrosa-Hetland, S., Pelkonen, A., Oksanen, J., & Kander, A. (2019). The prevalence of publicly stimulated innovations –A comparison of Finland and Sweden, 1970–2013. *Research Policy*, *48*(6), 1373–1384. https://doi.org/10.1016/j.respol.2019.02.001

Tosi, J., & Warmke, B. (2016). Moral Grandstanding. *Philosophy & Public Affairs*, *44*(3), 197–217. https://doi.org/10.1111/papa.12075

Trott, P., & Hartmann, D. (2009). Why 'open innovation' is old wine in new bottles. *International Journal of Innovation Management*, *13*(04), 715–736. https://doi.org/10.1142/S1363919609002509

Tsai, W. (2001). Knowledge transfer in intraorganizational networks: Effects of network position and absorptive capacity on business unit innovation and performance. *Academy of Management Journal*, *44*(5), 996–1004.

Utterback, J. M., & Abernathy, W. J. (1975). A dynamic model of process and product innovation. *Omega*, *3*(6), 639–656.

Van de Lei, T. E., & Cunningham, S. W. (2006). Use of the internet for future-oriented technology analysis. *2nd International Seville Seminar on 'Future-Oriented Technology Analysis: Impact of FTA Approaches on Policy and Decision-Making'. Seville, Spain*, 28–29.

Van de Ven, A. H., & Chu, Y. (1989). A psychometric assessment of the Minnesota innovation survey. *Research on the Management of Innovation: The Minnesota Studies*, 55–103.

Van Dijk, B., Den Hertog, R., Menkveld, B., & Thurik, R. (1997). Some new evidence on the determinants of large-and small-firm innovation. *Small Business Economics*, *9*, 335–343.

van Dijk, J., Antonides, G., & Schillewaert, N. (2014). Effects of co-creation claim on consumer brand perceptions and behavioural intentions. *International Journal of Consumer Studies*, *38*(1), 110–118. https://doi.org/10.1111/ijcs.12071

Vaughan, L. (2004). Exploring website features for business information. *Scientometrics*. https://akademiai.com/doi/abs/10.1023/B%3ASCIE.0000045122.93018.2a

Veblen, T. (2017). *The Theory of the Leisure Class*. Routledge. https://doi.org/10.4324/9781315135373

Veltri, G. A. (2013). Microblogging and nanotweets: Nanotechnology on Twitter. *Public Understanding of Science*, *22*(7), 832–849. https://doi.org/10.1177/0963662512463510

Verrecchia, R. E. (1983). Discretionary disclosure. *Journal of Accounting and Economics*, *5*, 179–194. https://doi.org/10.1016/0165-4101(83)90011-3

Vossen, R. W. (1998). Relative strengths and weaknesses of small firms in innovation. *International Small Business Journal*, *16*(3), 88–94.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 307–333.

Weare, C., & Lin, W.-Y. (2000). Content Analysis of the World Wide Web: Opportunities and Challenges. *Social Science Computer Review*, *18*(3), 272–292. https://doi.org/10.1177/089443930001800304

Wellalage, N. H., & Fernandez, V. (2019). Innovation and SME finance: Evidence from developing countries. *International Review of Financial Analysis*, *66*, 101370. https://doi.org/10.1016/j.irfa.2019.06.009

Wells, Valacich, & Hess. (2011). What Signal Are You Sending? How Website Quality Influences Perceptions of Product Quality and Purchase Intentions. *MIS Quarterly*, *35*(2), 373. https://doi.org/10.2307/23044048

West, J., & Bogers, M. (2014). Leveraging External Sources of Innovation: A Review of Research on Open Innovation. *Journal of Product Innovation Management*, *31*(4), 814–831. https://doi.org/10.1111/jpim.12125

West, J., & Lakhani, K. R. (2008). Getting Clear About Communities in Open Innovation. *Industry and Innovation*, *15*(2), 223–231. https://doi.org/10.1080/13662710802033734

West, J., & Wood, D. (2014). Evolving an open ecosystem: The rise and fall of the Symbian platform. In *Collaboration and competition in business ecosystems*. Emerald Group Publishing Limited.

Westra, E. (2021). Virtue Signaling and Moral Progress. *Philosophy & Public Affairs*, *49*(2), 156–178. https://doi.org/10.1111/papa.12187

Wilson, P. (2015). The misuse of the Vuong test for non-nested models to test for zero-inflation. *Economics Letters*, *127*, 51–53. https://doi.org/10.1016/j.econlet.2014.12.029

Winter, S. G. (1997). Knowledge and Competence as Strategic Assets. In *The Strategic Management of Intellectual Capital*. Routledge.

Wold, H. (1975). 11 - Path Models with Latent Variables: The NIPALS Approach**NIPALS = Nonlinear Iterative PArtial Least Squares. In H. M. Blalock, A. Aganbegian, F. M. Borodkin, R. Boudon, & V. Capecchi (Eds.), *Quantitative Sociology* (pp. 307–357). Academic Press. https://doi.org/10.1016/B978-0-12-103950-9.50017-4

Wold, H. (1980). Model Construction and Evaluation When Theoretical Knowledge Is Scarce: Theory and Application of Partial Least Squares. In J. Kmenta & J. B. Ramsey (Eds.), *Evaluation of Econometric Models* (pp. 47–74). Academic Press. https://doi.org/10.1016/B978-0-12-416550-2.50007-8

Wood, D. J., & Gray, B. (1991). Toward a comprehensive theory of collaboration. *The Journal of Applied Behavioral Science*, *27*(2), 139–162.

Worren, N., Moore, K., & Cardona, P. (2002). Modularity, strategic flexibility, and firm performance: A study of the home appliance industry. *Strategic Management Journal*, *23*(12), 1123–1140. https://doi.org/10.1002/smj.276

Xu, G., Zhang, Y., & Li, L. (2011). Web Content Mining. In G. Xu, Y. Zhang, & L. Li (Eds.), *Web Mining and Social Networking: Techniques and Applications* (pp. 71–87). Springer US. https://doi.org/10.1007/978-1-4419-7735-9_4

Yakin, H. S. Mohd., & Totu, A. (2014). The Semiotic Perspectives of Peirce and Saussure: A Brief Comparative Study. *Procedia - Social and Behavioral Sciences*, *155*, 4–8. https://doi.org/10.1016/j.sbspro.2014.10.247

Youtie, J., Hicks, D., Shapira, P., & Horsely, T. (2012). Pathways from Discovery to Commercialization: Using Web Sources to Track Small and Medium-sized Enterprise Strategies in Emerging Nanotechnologies. *Technology Analysis & Strategic Management*, *24*(10), 981–995. https://doi.org/10.1080/09537325.2012. 724163.

Youtie, J., Hicks, D., Shapira, P., & Horsley, T. (2012). Pathways from discovery to commercialisation: Using web sources to track small and medium-sized enterprise

strategies in emerging nanotechnologies. *Technology Analysis & Strategic Management*, *24*(10), 981–995. https://doi.org/10.1080/09537325.2012.724163

Zahavi, A. (1975). Mate selection—A selection for a handicap. *Journal of Theoretical Biology*, *53*(1), 205–214. https://doi.org/10.1016/0022-5193(75)90111-3

Zahavi, A. (1977). The cost of honesty. *Journal of Theoretical Biology*, *67*(3), 603–605. https://doi.org/10.1016/0022-5193(77)90061-3

Zahra, S. A., & George, G. (2002). Absorptive capacity: A review, reconceptualization, and extension. *Academy of Management Review*, *27*(2), 185–203.

Zaltman, G., Duncan, R., & Holbek, J. (1973). *Innovations and organisations*. John Wiley & Sons.

APPENDIX A – SUPPLEMENTARY MATERIALS ARTICLE 1

Relevant questions from the nano questionnaire-based survey

R&D

1. How many nanotechnology-related and/or advanced material products in development do you actually have in each of the following phases?
    - Applied Research;
    - Product Scoping and Business Case Building;
    - Development, Testing and Validation;
    - Commercialisation.
2. How important to your plant's innovation activities are each of the following sources of knowledge and innovation? (1-Not important, 2-Very low, 3-Low, 5-High, 6-Very high, 7-Essential).
    - Internal R&D in your firm;
    - Commercial laboratories/R&D firms/Technical Consultants.
3. Please indicate the level of importance of each of the following innovation activities to your plant during the period 2010 to 2014 (1-Not important, 2-Very low, 3-Low, 5-High, 6-Very high, 7-Essential).
    - Contracting of external R&D service providers;
    - Providing R&D services to third parties.
4. How long did it take to develop your most significant and recent (MSR) nanotechnology-related product innovation?
5. How important were each of the following organisations as collaborators in the development and commercialisation of your MSR product innovation? (1-Not important, 2-Very low, 3-Low, 5-High, 6-Very high, 7-Essential).
    - Private research laboratories/Research and Development firms.

6. How important were the following reasons in deciding to collaborate for the development and the commercialisation of your MSR product innovation? (1-Not important, 2-Very Low, 3-Low, 5-High, 6-Very high, 7-Essential)
    - Accessing research and development.
7. What proportion of Canadian employees from your firm are assigned primarily in R&D (%)?

Collaboration

8. Did your firm participate in alliances or collaborative agreements with other organisations to develop or commercialise your MSR product innovation? Y/N
9. How important were each of the following organisations as collaborators in the development and commercialisation of your MSR product innovation? (1-Not important, 2-Very low, 3-Low, 5-High, 6-Very high, 7-Essential).
    - Universities or higher education institutions, College centres for technology transfer (CCTT) and CEGEPs, university technology transfer offices.

External financing

10. Please indicate the proportion (%) of the total amount of financing provided by each of the following sources for the development and commercialisation of your MSR product innovation.

	Development of innovation	Commercialisation of innovation
Internal funds of your firm or establishment		
Government subsidies/tax credits/academics grants		
Debt capital (such as bank loans)		
Venture capital (public/private)		
Collaboration agreements		

Programs from organisations such as nanoQuebec (now PRIMA-Quebec), nanoOntario, nanoAlberta, etc.		
Other		

Intellectual property

11. Which of the following mechanisms are used by your firm to protect the intellectual property rights (IPR) for your MSR product innovation?

- Patents
- Trademarks
- Confidentiality agreements
- Trade secrets
- First mover advantage
- Other

12. How many patents does your firm own? Please note that the same patent filed in different countries is considered as only one patent.

	All patents	Nanotechnology-related and advanced materials patents
Patent applications		
Existing patents		
Patents assigned/(sold) to others		

Table A-1. Correlations table

		1	2	3	4	5	6	7	8	9
WEB_R&D	**1**	1.00								
WEB_IP	**2**	-0.18	1.00							
WEB_Collab	**3**	0.35 **	-0.30 **	1.00						
WEB _ExtFin	**4**	0.23 *	-0.37 **	0.42 **	1.00					
ln(nb_R&D_proj)	**5**	0.00	-0.13	-0.00	0.04	1.00				
dInt_R&D_source	**6**	0.17	0.01	0.08	0.05	0.09	1.00			
Ext_R&D_source	**7**	0.15	0.20	0.09	0.07	0.19	-0.07	1.00		
Contract_In_R&D	**8**	0.28 **	-0.21 *	0.28 **	0.15	0.14	-0.06	0.40 **	1.00	
Contract_Out_R&D	**9**	0.31 **	-0.07	0.08	0.05	-0.05	0.10	-0.11	0.06	1.00
ln(time_R&D)	**10**	0.20 *	-0.21 *	0.24 *	0.17	0.10	-0.09	0.16	0.25 *	0.07
propEmp_R&D	**11**	0.31 **	-0.46 **	0.12	0.39 **	-0.07	-0.17	0.01	0.21 *	0.17
dCollab_R&D	**12**	-0.02	-0.19	0.12	0.14	0.10	-1.00 **	0.21	0.01	-0.24
dCollab_R&D_Reason	**13**	-0.27	-0.02	0.09	-0.15	0.11	-0.23	0.37 *	0.22	-0.18
nbIP	**14**	0.01	-0.37 **	0.23 *	0.38 **	0.24 *	-0.12	0.17	0.43 **	-0.07
ln(nbPatent)	**15**	0.04	-0.40 **	0.09	0.01	0.09	0.33 *	-0.12	-0.08	-0.13
dCollab	**16**	0.17	-0.25 *	0.22 *	0.11	0.19	0.06	-0.03	0.22 *	0.07
propExtFin_R&D	**17**	0.36 **	-0.18	0.31 **	0.15	0.16	-0.07	0.07	0.23 *	0.10
propExtFin_Comm	**18**	0.37 **	-0.30 **	0.28 **	0.22 *	0.11	0.01	0.03	0.26 *	0.24 *
propTotExtFin	**19**	0.40 **	-0.27 *	0.33 **	0.19	0.15	-0.04	0.07	0.26 *	0.17

Table A-1. Correlations table (continued)

		10	11	12	13	14	15	16	17	18	19
ln(time_R&D)	**10**	1.00									
propEmp_R&D	**11**	0.31 **	1.00								
dCollab_R&D	**12**	0.12	0.21	1.00							
dCollab_R&D_Reason	**13**	0.57 **	-0.13	0.23	1.00						
nbIP	**14**	0.26 *	0.28 **	0.20	0.27	1.00					
ln(nbPatent)	**15**	0.18	0.02	-0.24	0.16	0.14	1.00				
dCollab	**16**	0.14	0.11	.a	.a	0.22 *	0.13	1.00			
propExtFin_R&D	**17**	0.18	0.24 *	0.19	-0.28	0.14	-0.23	0.36 **	1.00		
propExtFin_Comm	**18**	0.11	0.16	-0.17	-0.06	0.20 *	-0.09	0.29 **	0.68 **	1.00	
propTotExtFin	**19**	0.16	0.21 *	-0.01	-0.18	0.18	-0.16	0.38 **	0.91 **	0.92 **	1.00

a. Cannot be computed because at least one of the variance is constant. *p<0.05; **p<0.01

Table A-2. Descriptive statistics

	N	Mean	Std. Dev	Min	Median	Max
WEB_R&D	78	2.430	3.322	0.000	1.304	17.942
WEB_IP	78	0.577	1.694	0.000	0.046	13.619
WEB_Collab	78	1.900	2.820	0.000	0.702	16.802
WEB _ExtFin	78	0.272	0.590	0.000	0.000	2.858
nb_R&D_proj	76	10.671	15.356	0	5.500	90
dInt_R&D_source	84	0.952	0.214	0	1	1
Ext_R&D_source	88	3.898	1.781	1	4	7
Contract_In_R&D	86	3.465	1.813	1	4	7
Contract_Out_R&D	86	3.767	2.135	1	4	7
time_R&D	80	38.063	28.611	2	30	144
propEmp_R&D	79	33.848	30.695	0	20	100
dCollab_R&D	37	0.027	0.164	0	0	1
dCollab_R&D_Reason	31	0.419	0.502	0	0	1
nbIP	80	3.225	1.283	0	3	6
nbPatent	44	17.705	33.272	0	4	141
dCollab	76	0.500	0.503	0	0.50	1
propExtFin_R&D	80	44.088	33.693	0	40	100
propExtFin_Comm	77	26.312	35.023	0	5	100
propTotExtFin	77	34.630	31.393	0	25	100
RD_INDEX	80	0.004	0.167	-0.310	-0.008	0.495
IP_INDEX	78	-0.014	0.492	-0.912	-0.169	1.873

Table A-3. Correlations between individual questionnaire-based and web-based indicators

Variables	1	II	III	IV
	WEB_R&D	WEB _IP	WEB_ Collab	WEB _ExtFin
ln(nb_R&D_proj)	0.002			
dInt_R&D_source	0.171			
Ext_R&D_source	0.152			
Contract_In_R&D	0.284**			
Contract_Out_R&D	0.306***			
ln(time_R&D)	0.197*			
propEmp_R&D	0.306***			
dCollab_R&D	-0.017			
dCollab_R&D_Reason	-0.270			
nbIP		-0.368***		
ln(nbPatent)		-0.396**		
dCollab			0.222*	
propExtFin_R&D				0.150
propExtFin_Comm				0.222*
propTotExtFin				0.192

Notes: Group I, II, III performed with Pearson correlations and IV performed with Spearman correlations. Two-tailed test. *p<0.10; **p<0.05; ***p<0.01.

Table A-4. PLS regression results

PLS Model B Basic Linear Stable 3 for INDEX_R&D

Items	**Weights**	**SE**	**p-value**	**VIF**	**WLS**	**ES**
Z_propEmp_R&D	0.323	0.101	0.001	1.229	1	0.166
Z_nb_R&D_proj	0.323	0.101	0.001	1.087	1	0.129
Z_time_R&D	0.323	0.101	0.001	1.255	1	0.206
Z_Ext_R&D_source	0.323	0.101	0.001	1.346	1	0.175
Z_Contract_Out_R&D	0.323	0.101	0.001	1.071	1	0.122
Z_Contract_In_R&D	0.323	0.101	0.001	1.352	1	0.201

PLS Model B Basic Warp3 Basic Stable 3 for INDEX_R&D

Items	**Weights**	**SE**	**p-value**	**VIF**	**WLS**	**ES**
Z_propEmp_R&D	0.323	0.101	0.001	1.229	1	0.166
Z_nb_R&D_proj	0.323	0.101	0.001	1.087	1	0.129
Z_time_R&D	0.323	0.101	0.001	1.255	1	0.206
Z_Ext_R&D_source	0.323	0.101	0.001	1.346	1	0.175
Z_Contract_Out_R&D	0.323	0.101	0.001	1.071	1	0.122
Z_Contract_In_R&D	0.323	0.101	0.001	1.352	1	0.201

PLS Model B Basic Warp3 Basic Stable 3 for INDEX_IP

Items	**Weights**	**SE**	**p-value**	**VIF**	**WLS**	**ES**
Z_nbIP	0.636	0.093	<0.001	1.059	1	0.5
Z_nbPatent	0.636	0.093	<0.001	1.059	1	0.5

PLS Model B Basic Linear Basic Stable 3 for INDEX_IP

Items	Weights	SE	p-value	VIF	WLS	ES
Z_nbIP	0.636	0.093	<0.001	1.059	1	0.5
Z_nbPatent	0.636	0.093	<0.001	1.059	1	0.5

APPENDIX B – EXAMPLES OF TEN'S QUESTIONS

Innovation number

(Program ID:14)

How many new-to-the-world products or services has your company introduced over the past three years?

- 5 or more
- 4
- 3
- 2
- 1
- None

Innovation proportion

(Program ID: 3)

What percentage of your company's revenues can be attributed to new or substantially improved products, processes, or services introduced since your engagement with Program ID: 3 or Program ID: 3 -supported researchers?

- 90% or more
- Between 60% and 90%
- Between 30% and 60%
- Less than 10%
- We have not introduced a new or substantially improved product, process, or service since our engagement with Program ID: 3

Degree of innovation

(Program ID: 4)

How innovative do you consider your company to be?

- International leader, we rank first in the world
- International advanced, there are few companies in the world with our capabilities
- National leader, we rank first in Canada
- National advanced, there are few companies in Canada with our capabilities
- We have not yet developed an innovative new or improved product, process or service

(Program ID: 2)

How innovative do you consider your company to be?

- International leader, we rank first in the world
- International advanced, there are few companies in the world with our capabilities
- National leader, we rank first in Canada
- National advanced, there are few companies in Canada with our capabilities
- Regional leader, we rank first in Canada
- Regional advanced, there are few companies in Canada with our capabilities
- We have not yet developed an innovative new or improved product, process or service

Time to market

(Program ID: 3)

How quickly does your company typically introduce new products, processes, or services to the market? Please specify the time between the product, process, or service conceptualisation and market launch.

- Less than 1 year

- Between 1 year and 2 years
- Between 2 years and 5 years
- More than 5 years
- We have not yet offered products, processes, or services to the market

Investment

(Program ID: 4)

What is the value of investments (equity, loans, grants, etc.) received by your company in the past year?

- Greater than $10 million
- Between $1 million and $10 million
- Between $500K and $1 million
- Between $100K and $500K
- Less than $100K
- We have not received investments in the past year

(Program ID: 14)

How much equity financing has your company raised in the past three years?

- More than $10 million
- Between $5 million and $9.99 million
- Between $1 million and $4.99 million
- Between $500K and $999K
- Between $100K and $499K
- Less than $100K
- Not at all

- Our company has not tried to raise private equity

Sectors

(Program ID: 4)

In what industrial sector does your company participate?

- Advanced Health Technologies
- Digital Media and ICT
- Green Energy Technologies
- Environmental – Clean Technologies
- Environmental – Water Technologies
- Agriculture
- Forestry
- Materials and Advanced Manufacturing
- Financial Services
- Other (please specify)

(Program ID: 5)

In which industrial sector does your company belong?

- Energy
- Environment
- Information and Communications Technologies
- Life sciences, Health or Medical
- Materials or Manufacturing
- Natural Resources
- Other (please specify)

APPENDIX C – KEYWORDS AND STATISTICS ARTICLE 2 AND 3

Table C-1. The web keyword analysis construct

Concepts	Keywords considered	Removed	Processed stemmed keywords
Innovation (Herzog, 2011; OECD & Eurostat, 2018)	Innovation, Innovative, innovation culture, innovator, innovate, creativity, create, creation, new ideas, risk taking, new product, new process, new method, new procedure, new service, market introduction	risk taking, new product, new process, new method, new procedure, new service, market introduction	innov, creativ, create, creation, idea
Collaboration (Arvanitis, 2012; Belderbos et al., 2004; Carboni, 2013; Ramdani, 2014)	affiliation, collaboration, cooperation, partners, partnership, innovation partnership collaborate, collaborator, collaborative network, collaborative innovation, collaborative platform, collaborative research		affiliat, collab, cooperat, partner
R&D (Gök et al., 2015; Héroux-Vaillancourt et al., 2020; Meyer, 2000; OECD, 2015; Schmoch, 2007)	research & development, research and development, r&d, researcher, product development, technology development, technical development, development phase, development program, development process, development project, development centre, development facility, technological development, development effort, development cycle, development research, development activity, fundamental research, basic research, science, scientist, scientific, inventive, inventor, invention, experimentation, experiment, experience, laboratory, test	experience, test	research, develop, r&d, tech, scien, invent, experiment, lab

Table C-1. The web keyword analysis construct (continued)

Concepts	Keywords considered	Removed	Processed stemmed keywords
Open innovation (Bogers et al., 2017; Chesbrough, 2003a, 2006b; Liyanage et al., 1999)	consorti, international consorti, global consorti, open innovation, business ecosystem, innovation ecosystem, co-development, community, innovation contest, technolo scout, technolo intelligence, technolo watch, technolo scan, technolo survey, competitive technical intelligen monitor technolo, lead users, early adopters, external innovators, crowdsourcing, crowdfunding, community sourcing, spin off, creative commons licenses, public copyright licenses, public license, open copyright licenses, technology transfer, design thinking, open source, open data, open content, strategic alliance, inside-out, outside-in, startup	innovation contest, technolo scout, technolo intelligence, technolo watch, technolo scan, technolo survey, competitive technical intelligen monitor technolo, lead users, external innovators, community sourcing, creative commons licenses, public copyright licenses, open copyright licenses, open content, strategic alliance, inside-out, outside-in	consorti, open innov, business ecosystem, innovation ecosystem, co-dev, community, early adopter, crowdsourcing, crowdfunding, spin off, public licence, tech transfert, design thinking, open source, open data, startup
Intellectual property (Rietsch, 2015; Teece, 1986)	patent, intellectual property, trade secret, industrial design	intellectual property	patent, secret industrial design

Table C-2. Document keywords frequencies by unique website-questionnaire year pair

	Total	Min	Max	Mean	Std. dev.	Freq.	Cumul. freq.
Sentences	511,815	1	4,785	220.80	333.19		
Document keywords							
tech	27,360	0	401	11.80	27.51	21.1%	21%
develop	23,998	0	452	10.35	25.23	18.5%	40%
lab	17,188	0	531	7.42	18.47	13.3%	53%
innov	8,741	0	1,220	3.77	27.70	6.7%	60%
partner	8,369	0	152	3.61	8.91	6.5%	66%
research	8,218	0	246	3.55	13.71	6.3%	72%
create	7,622	0	347	3.29	10.28	5.9%	78%
idea	4,790	0	657	2.07	14.56	3.7%	82%
scien	4,747	0	180	2.05	8.75	3.7%	86%
community	4,028	0	161	1.74	7.63	3.1%	89%
collab	3,201	0	152	1.38	6.79	2.5%	91%
creativ	1,857	0	370	0.80	8.30	1.4%	93%
invent	1,636	0	119	0.71	4.25	1.3%	94%
creation	1,476	0	115	0.64	3.83	1.1%	95%
patent	1,470	0	107	0.63	3.31	1.1%	96%
affiliat	1,056	0	69	0.46	2.34	0.8%	97%
r&d	845	0	44	0.36	2.07	0.7%	98%
experiment	585	0	48	0.25	1.62	0.5%	98%
cooperat	575	0	18	0.25	1.03	0.4%	99%
secret	565	0	41	0.24	1.33	0.4%	99%
startup	528	0	28	0.23	1.30	0.4%	99%
open source	240	0	28	0.10	1.21	0.2%	100%
consorti	154	0	14	0.07	0.56	0.1%	100%
open innov	78	0	50	0.03	1.07	0.1%	100%
technology transfert	41	0	6	0.02	0.20	0.0%	100%
industrial design	38	0	5	0.02	0.20	0.0%	100%
early adopter	35	0	7	0.02	0.23	0.0%	100%
design thinking	34	0	13	0.01	0.30	0.0%	100%

Table C-2. Document keywords frequencies by unique website-questionnaire year pair (continued)

	Total	Min	Max	Mean	Std. dev.	Freq.	Cumul. freq.
crowdsourcing	33	0	12	0.01	0.32	0.0%	100%
open data	32	0	8	0.01	0.25	0.0%	100%
codev	23	0	3	0.01	0.13	0.0%	100%
innovation ecosystem	21	0	13	0.01	0.28	0.0%	100%
crowdfunding	20	0	16	0.01	0.34	0.0%	100%
spin off	16	0	15	0.01	0.31	0.0%	100%
public licence	9	0	7	0.00	0.15	0.0%	100%
business ecosystem	2	0	2	0.00	0.04	0.0%	100%
Total	129,631	2	27,360	3,600.86	6,570.07		

Table C-3. Standardised document keyword frequencies by unique website-questionnaire year pair

	Total	min	max	Mean	Std. dev.	Freq	Cumul. freq.
Document keywords							
tech	138.069	0.000	1.000	0.060	0.096	22.5%	22.5%
develop	123.568	0.000	1.000	0.053	0.089	20.1%	42.6%
lab	76.190	0.000	0.800	0.033	0.055	12.4%	55.0%
innov	44.254	0.000	1.000	0.019	0.063	7.2%	62.2%
partner	41.855	0.000	1.000	0.018	0.045	6.8%	69.0%
research	37.826	0.000	1.000	0.016	0.050	6.2%	75.1%
create	33.687	0.000	0.500	0.015	0.032	5.5%	80.6%
scien	22.469	0.000	1.000	0.010	0.043	3.7%	84.3%
idea	20.547	0.000	0.500	0.009	0.026	3.3%	87.6%
community	14.102	0.000	0.361	0.006	0.022	2.3%	89.9%
collab	12.631	0.000	0.400	0.005	0.020	2.1%	92.0%
patent	8.295	0.000	0.333	0.004	0.018	1.3%	93.3%
creativ	7.531	0.000	0.500	0.003	0.017	1.2%	94.5%
invent	6.903	0.000	0.272	0.003	0.014	1.1%	95.6%
creation	6.145	0.000	0.333	0.003	0.014	1.0%	96.6%

Table C-3. Standardised document keyword frequencies by unique website-questionnaire year pair (continued)

	Total	min	max	Mean	Std. dev.	Freq	Cumul. freq.
r&d	4.057	0.000	0.211	0.002	0.009	0.7%	97.3%
cooperat	3.725	0.000	0.500	0.002	0.014	0.6%	97.9%
affiliat	3.634	0.000	0.333	0.002	0.009	0.6%	98.5%
experiment	2.741	0.000	0.250	0.001	0.008	0.4%	99.0%
startup	2.189	0.000	0.138	0.001	0.006	0.4%	99.3%
secret	1.616	0.000	0.069	0.001	0.003	0.3%	99.6%
open source	0.746	0.000	0.111	0.000	0.003	0.1%	99.7%
consorti	0.523	0.000	0.048	0.000	0.002	0.1%	99.8%
industrial design	0.297	0.000	0.063	0.000	0.002	0.0%	99.8%
innovation ecosystem	0.187	0.000	0.146	0.000	0.003	0.0%	99.9%
design thinking	0.162	0.000	0.038	0.000	0.001	0.0%	99.9%
technology transfert	0.150	0.000	0.019	0.000	0.001	0.0%	99.9%
codev	0.129	0.000	0.026	0.000	0.001	0.0%	99.9%
crowdsourcing	0.124	0.000	0.063	0.000	0.002	0.0%	99.9%
early adopter	0.096	0.000	0.021	0.000	0.001	0.0%	100.0%
open innov	0.093	0.000	0.035	0.000	0.001	0.0%	100.0%
open data	0.084	0.000	0.023	0.000	0.001	0.0%	100.0%
public licence	0.027	0.000	0.014	0.000	0.000	0.0%	100.0%
crowdfunding	0.020	0.000	0.012	0.000	0.000	0.0%	100.0%
spin off	0.004	0.000	0.004	0.000	0.000	0.0%	100.0%
business ecosystem	0.002	0.000	0.002	0.000	0.000	0.0%	100.0%
Total	614.679	0.002	138.069	17.074	32.721	1.000	1.000

Table C-4. Standardised document keywords frequencies by concept by unique website-questionnaire year pair

	Total	Mean	Std. dev.	min	max	Freq	Cumul. freq.
Innovation							
innov	44.254	0.019	0.063	0.000	1.000	39.45%	39.45%
create	33.687	0.015	0.032	0.000	0.500	30.03%	69.49%
idea	20.547	0.009	0.026	0.000	0.500	18.32%	87.81%
creativ	7.531	0.003	0.017	0.000	0.500	6.71%	94.52%
creation	6.145	0.003	0.014	0.000	0.333	5.48%	100.00%
Total	112.164						
Collaboration							
partner	41.855	0.018	0.045	0.000	1.000	67.68%	67.68%
collab	12.631	0.005	0.020	0.000	0.400	20.42%	88.10%
cooperat	3.725	0.002	0.014	0.000	0.500	6.02%	94.12%
affiliat	3.634	0.002	0.009	0.000	0.333	5.88%	100.00%
Total	61.845						
Open innovation							
community	14.102	0.006	0.022	0.000	0.361	75.66%	75.66%
startup	2.189	0.001	0.006	0.000	0.138	11.74%	87.40%
open source	0.746	0.000	0.003	0.000	0.111	4.00%	91.40%
consorti	0.523	0.000	0.002	0.000	0.048	2.81%	94.21%
innovation ecosystem	0.187	0.000	0.278	0.000	0.146	1.00%	95.21%
design thinking	0.162	0.000	0.001	0.000	0.038	0.87%	96.08%
technology transfert	0.150	0.000	0.001	0.000	0.019	0.81%	96.89%
codev	0.129	0.000	0.001	0.000	0.026	0.69%	97.58%
crowdsourcing	0.124	0.000	0.002	0.000	0.063	0.67%	98.25%
early adopter	0.096	0.000	0.001	0.000	0.021	0.52%	98.76%
open innov	0.093	0.000	0.001	0.000	0.035	0.50%	99.26%

Table C-4. Standardised document keywords frequencies by concept by unique website-questionnaire year pair (continued)

	Total	Mean	Std. dev.	min	max	Freq	Cumul. freq.
open data	0.084	0.000	0.001	0.000	0.023	0.45%	99.71%
public licence	0.027	0.000	0.151	0.000	0.014	0.15%	99.86%
crowdfunding	0.020	0.000	0.336	0.000	0.012	0.11%	99.96%
spin off	0.004	0.000	0.312	0.000	0.004	0.02%	99.99%
business ecosystem	0.002	0.000	0.042	0.000	0.002	0.01%	100.00%
Total	18.639						
R&D							
tech	138.069	0.060	0.096	0.000	1.000	33.53%	33.53%
develop	123.568	0.053	0.089	0.000	1.000	30.01%	63.53%
lab	76.190	0.033	0.055	0.000	0.800	18.50%	82.03%
research	37.826	0.016	0.050	0.000	1.000	9.18%	91.22%
scien	22.469	0.010	0.043	0.000	1.000	5.46%	96.67%
invent	6.903	0.003	0.014	0.000	0.272	1.68%	98.35%
r&d	4.057	0.002	0.009	0.000	0.211	0.99%	99.33%
experiment	2.741	0.001	0.008	0.000	0.250	0.67%	100.00%
Total	411.823						
Intellectual property							
patent	8.295	0.004	0.018	0.000	0.333	81.26%	81.26%
secret	1.616	0.001	0.003	0.000	0.069	15.83%	97.09%
industrial design	0.297	0.000	0.002	0.000	0.063	2.91%	100.00%
Total	10.207						

Table C-5. Proportion of observations without keyword occurrences

Keywords	% of observations without keyword occurrences
business ecosystem	99.96%
public license	99.92%
spin off	99.92%
crowdfunding	99.83%
innovation ecosystem	99.66%
crowdsourcing	99.58%
open innov	99.50%
design thinking	99.45%
open data	99.45%
codev	99.33%
early adopter	99.20%
industrial design	99.03%
technology transfer	98.74%
open source	97.35%
consorti	96.85%
startup	92.52%
experiment	90.67%
r&d	90.00%
secret	89.58%
cooperat	88.73%
affiliat	86.38%
patent	83.66%
invent	83.27%
creativ	81.80%
creation	81.72%
collab	74.99%
community	73.14%
scien	72.89%
research	60.91%

Table C-5. Proportion of observations without keyword occurrences (continued)

Keywords	**% of observations without keyword occurrences**
idea	58.97%
innov	54.98%
create	48.47%
partner	46.66%
lab	32.96%
tech	31.36%
develop	29.17%

APPENDIX D – SUPPLEMENTARY MATERIALS ARTICLE 2

Table D-1. Sampling summary

# program	Year	Invitations	Entries	Websites found	Firms	Websites retrieved	Final
1	2010	53	33	26	25	18	0
2	2010	72	31	25	23	21	12
3	2011	754	282	194	184	151	66
4	2011	224	71	40	40	37	31
5	2012	142	68	62	61	34	13
6	2012	110	54	48	44	31	24
7	2012	812	382	264	226	194	129
8	2012	207	94	73	71	60	27
9	2012	305	163	45	41	34	13
10	2012	73	49	22	21	15	11
11	2012	29	21	15	15	14	0
12	2012	59	27	18	18	13	0
13	2013	605	260	138	130	116	70
14	2013	894	313	100	98	80	55
15	2013	151	109	75	75	68	0
16	2013	200	140	75	74	60	0
17	2013	34	29	19	17	15	6
18	2014	34	29	27	27	23	19
19	2014	108	87	68	53	44	37
20	2014	23	23	17	15	14	0
21	2014	3,952	1,307	929	905	797	600
22	2014	82	48	44	43	38	31
23	2015	545	222	174	168	153	0
24	2015	750	230	196	166	136	64
25	2015	323	150	104	99	87	52
26	2015	38	23	21	19	19	8
27	2016	38	30	21	19	16	14
28	2016	1,529	425	74	72	61	7

Table D-1. Sampling summary (continued)

# program	Year	Invitations	Entries	Websites found	Firms	Websites retrieved	Final
29	2016	56	40	35	31	30	0
n		12,202	4,740	2,949	2,780	2,379	1,289
%			39%	62%	94%	86%	54%

Table D-2. Questionnaire-based survey variables

Variable	Type	Definition	Transformation
InnovEvidence	Continuous	Measure of the intensity of innovation evidence, ranging from 0 corresponding to no innovation evidence, to 1 with the strongest innovation evidence	-
TimetoMarket	Discrete	Number of months required by a company to develop and introduce new or improved products, processes, or services to the market	ln(TimetoMarket/100 + 1)
Investment	Continuous	Amount of funding or investments received	ln(Investment/50,000 + 1)
TechFirm	Dummy	Dummy variable that takes the value 1 when a firm uses or develops technologies that require the involvement of STEM disciplines, and 0 otherwise	-
Revenue	Continuous	Company's annual revenues	ln(Revenue/1,000 + 1)
d2011, d2013, d2015	Dummy	Dummy variable that takes the value 1 for the year of the investigation, and 0 otherwise	-
dDigMediaICT, dLifeSciHealth, dMaterialMnfg	Dummy	Dummy variable that takes the value 1 for the sector in which the firm operates and 0 otherwise.	-
dFinland	Dummy	Dummy variable that takes the value 1 if the company is located in Finland	-
dYoung	Dummy	Dummy variable that takes the value 1 if the firm is considered young (4 to 8 years old) in the year of the investigation year	-

Notes : Continuous variables were transformed using the natural logarithm to normalize them.

Table D-3. Web-based variables construct

Concepts	Indicators	Variable	Transformed
Web-based innovation signalling	The sum of all standardised keyword frequencies on a firm's website corresponding to concepts related to innovation	webIndex	lndwebIndex ln(10x+1)
	The weighted average from PLS weights	webIndexPLS	lncwebIndexPLS ln(100x+1)
Web-based innovation culture signalling	The sum of all standardised keyword frequencies referring to innovation on a firm's website	webInnov	lncwebInnov ln(100x+1)
	The weighted average from PLS weights	webInnovPLS	lniicwebInnovPLS ln(200x+1)
Web-based collaboration signalling	The sum of all standardised keyword frequencies referring to collaboration on a firm's website	webCollab	lncwebCollab ln(100x+1)
	The weighted average from PLS weights	webCollabPLS	LnvcwebCollabPLS ln(500x+1)
Web-based OI signalling	The sum of all standardised keyword frequencies referring to OI on a firm's website	webOpen	lndmwebOpen ln(10,000x + 1)
	The weighted average from PLS weights	webOpenPLS	lncmwebOpenPLS ln(100,000x + 1)
Web-based R&D signalling	The sum of all standardised keyword frequencies referring to R&D on a firm's website	webRD	lndwebRD ln(10x + 1)
	The weighted average from PLS weights	webRDPLS	lnvdwebRDPLS ln(50x + 1)

Table D-3. Web-based variables construct (continued)

Concepts	Indicators	Variable	Transformed
Web-based IP signalling	The sum of all standardised keyword frequencies referring to IP on a firm's website	webIP	lndMwebIP ln(10,000,000x +1)
	The weighted average from PLS weights	webIPPLS	lnMMwebIPPLS ln(1,000,000,000x+1)

Table D-4 Descriptive statistics

Variable	mean	std. dev.	min	median	max
InnovEvidence	0.383	0.331	0.000	0.400	1.000
Investment	899,563	1,981,832	0.000	300,000	12,000,000
webIndex	0.272	0.271	0.000	0.200	3.000
webInnov	0.053	0.087	0.000	0.029	1.000
webCollab	0.028	0.054	0.000	0.009	0.800
webOpen	0.006	0.023	0.000	0.000	0.361
webRD	0.182	0.207	0.000	0.121	2.000
webIP	0.005	0.018	0.000	0.000	0.275
webIndexPLS	0.015	0.016	0.000	0.011	0.184
webInnovPLS	0.017	0.028	0.000	0.009	0.304
webCollabPLS	0.011	0.024	0.000	0.003	0.366
webOpenPLS	0.002	0.007	0.000	0.000	0.106
webRDPLS	0.027	0.033	0.000	0.017	0.384
webIPPLS	0.005	0.018	0.000	0.000	0.275
TimetoMarket	20.713	16.490	6.000	18.000	108.000
TechFirm	0.618	0.486	0.000	1.000	1.000
dYoung	0.401	0.490	0.000	0.000	1.000
d2011	0.075	0.264	0.000	0.000	1.000
d2013	0.102	0.302	0.000	0.000	1.000
d2015	0.096	0.295	0.000	0.000	1.000
Revenue	2,894,559	7,402,423	0.000	500,000	50,000,000
dDigMediaICT	0.293	0.455	0.000	0.000	1.000
dLifeSciHealth	0.123	0.328	0.000	0.000	1.000
dMaterialMnfg	0.157	0.364	0.000	0.000	1.000
dFinland	0.465	0.499	0.000	0.000	1.000

Table D-5 Pearson's correlation table

		1	2	3	4	5	6	7	8	9	10	11
InnovEvidence	**1**	1.000										
Investment	**2**	-0.018	1.000									
webIndex	**3**	0.068 *	0.160 *	1.000								
webInnov	**4**	0.126 *	0.078 *	0.594 *	1.000							
webCollab	**5**	0.020	0.144 *	0.485 *	0.232 *	1.000						
webOpen	**6**	0.122 *	0.021	0.219 *	0.228 *	0.202 *	1.000					
webRD	**7**	0.033	0.180 *	0.909 *	0.365 *	0.331 *	0.110 *	1.000				
webIP	**8**	0.041	0.174 *	0.242 *	0.165 *	0.114 *	0.110 *	0.235 *	1.000			
webIndexPLS	**9**	0.063 *	0.155 *	0.974 *	0.640 *	0.551 *	0.214 *	0.827 *	0.222 *	1.000		
webInnovPLS	**10**	0.124 *	0.071 *	0.577 *	0.975 *	0.224 *	0.220 *	0.331 *	0.134 *	0.645 *	1.000	
webCollabPLS	**11**	0.014	0.141 *	0.473 *	0.235 *	0.963 *	0.220 *	0.329 *	0.123 *	0.548 *	0.223 *	1.000
webOpenPLS	**12**	0.141 *	0.052	0.265 *	0.275 *	0.229 *	0.907 *	0.156 *	0.163 *	0.259 *	0.266 *	0.243 *
webRDPLS	**13**	0.028	0.190 *	0.900 *	0.347 *	0.323 *	0.100 *	0.992 *	0.234 *	0.832 *	0.328 *	0.321 *
webIPPLS	**14**	0.042	0.175 *	0.238 *	0.165 *	0.115 *	0.115 *	0.232 *	1.000 *	0.218 *	0.135 *	0.125 *
TimetoMarket	**15**	-0.106 *	0.093 *	0.091 *	-0.016	0.035	-0.033	0.142 *	0.089 *	0.076 *	-0.035	0.028
TechFirm	**16**	0.013	0.218 *	0.240 *	0.121 *	0.095 *	-0.022	0.267 *	0.154 *	0.214 *	0.101 *	0.102 *
dYoung	**17**	-0.039	0.079 *	-0.006	0.001	0.033	-0.043	0.013	-0.037	0.004	0.003	0.045
d2011	**18**	0.269 *	-0.086 *	0.044	0.066 *	0.008	0.040	0.020	-0.035	0.047	0.074 *	0.003
d2013	**19**	0.133 *	-0.142 *	-0.070 *	-0.021	-0.050	0.117 *	-0.091 *	-0.021	-0.074 *	-0.020	-0.057 *
d2015	**20**	0.018	0.038	0.047	0.047	-0.003	0.111 *	0.050	0.050	0.044	0.044	-0.001

Table D-5 Pearson's correlation table (continued)

		1	2	3	4	5	6	7	8	9	10	11
Revenue	**21**	-0.047	0.020	-0.022	-0.004	0.093 *	0.074 *	-0.025	-0.042	-0.021	-0.009	0.118 *
dDigMediaICT	**22**	0.221 *	-0.029	0.058 *	0.162 *	0.051	0.159 *	-0.004	-0.024	0.062 *	0.180 *	0.050
dLifeSciHealth	**23**	0.086 *	0.235 *	0.199 *	0.052	0.084 *	-0.012	0.238 *	0.159 *	0.190 *	0.038	0.080 *
dMaterialMnfg	**24**	-0.115 *	0.012	-0.009	-0.048	-0.009	-0.138 *	0.014	-0.008	-0.018	-0.062 *	-0.005
dFinland	**25**	-0.385 *	0.084 *	-0.190 *	-0.172 *	-0.056 *	-0.316 *	-0.157 *	-0.129 *	-0.181 *	-0.164 *	-0.041

		12	13	14	15	16	17	18	19	20	21	22
webOpenPLS	**12**	1.000										
webRDPLS	**13**	0.146 *	1.000									
webIPPLS	**14**	0.168 *	0.230 *	1.000								
TimetoMarket	**15**	-0.011	0.153 *	0.086 *	1.000							
TechFirm	**16**	0.025	0.267 *	0.153 *	0.135 *	1.000						
dYoung	**17**	-0.046	0.020	-0.038	-0.015	0.051	1.000					
d2011	**18**	0.027	0.023	-0.034	-0.110 *	-0.078 *	0.103 *	1.000				
d2013	**19**	0.093 *	-0.101 *	-0.020	-0.102 *	-0.116 *	-0.081 *	-0.096 *	1.000			
d2015	**20**	0.107 *	0.053	0.050	0.274 *	0.094 *	-0.079 *	-0.093 *	-0.110 *	1.000		
Revenue	**21**	0.033	-0.030	-0.039	0.050	0.038	0.095 *	-0.019	-0.060 *	0.090 *	1.000	
dDigMediaICT	**22**	0.191 *	-0.019	-0.021	-0.220 *	0.041	-0.058 *	0.152 *	0.082 *	-0.054	-0.094 *	1.000
dLifeSciHealth	**23**	0.006	0.253 *	0.158 *	0.144 *	0.153 *	0.042	-0.062 *	-0.040	0.079 *	-0.124 *	-0.241 *
dMaterialMnfg	**24**	-0.149 *	0.011	-0.010	0.116 *	0.178 *	0.068 *	-0.083 *	-0.110 *	-0.026	0.111 *	-0.279 *
dFinland	**25**	-0.315 *	-0.156 *	-0.131 *	-0.078 *	0.008	0.125 *	-0.266 *	-0.314 *	-0.304 *	0.032	-0.116 *

Notes: Two-tailed test. *p<0.05. Performed with natural logarithm transformed variables when relevant.

Table D-5 Pearson's correlation table (continued)

		23	24	25
dLifeSciHealth	**23**	1.000		
dMaterialMnfg	**24**	-0.162 *	1.000	
dFinland	**25**	-0.230 *	0.233 *	1.000

Notes: Two-tailed test. *p<0.05. Performed with natural logarithm transformed variables when relevant.

Robustness checks and post-estimation tests

Because the reg3 procedure does not allow the clustering of repeated observations, the effect of repeated measures (n=69 observations) were tested by estimating a two-stage least-squares (2SLS) model using the ivreg procedure in Stata accounting for such non independence of repeated measures for the same firm. The 2SLS regression results with and without accounting for the clustering of repeated measures yield no differences in the coefficients and their significance level. The impact of such a small number of repeated measures is therefore considered to be negligible.

The 3SLS and 2SLS regression results were then compared, and show very close values for the coefficients and their significance level. Breusch-Pagan tests using the sureg procedure to assess heteroskedasticity were all non-significant, suggesting that the residuals are independently and identically distributed. Potential endogeneity was examined using the endogenous post-estimation tests, Durbin (score) and Wu-Hausman metrics, after the 2SLS regressions (StataCorp., 2019). Although these tests assume a valid set of instruments (which we do not have), all the results were non-significant implying that exogeneity cannot be rejected. If there were proper instruments, this would suggest that the models could be treated using OLS regressions without biases. For further robustness purposes, OLS regressions were estimated treating each equation of the model separately. Not surprisingly, the results were again almost identical to those obtained with the other methods. Table D-6 shows the results.

Table D-6. OLS regression results

Variables	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1. InnovEvidence							
Investment	0.006	0.004	0.005	0.006	0.005	0.030	0.047
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.041)	(0.049)
TimetoMarket	-0.209***	-0.211***	-0.212***	-0.204***	-0.213***	-0.216***	-0.199***
	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.073)	(0.073)
TechFirm							
dYoung	-0.010	-0.011	-0.010	-0.010	-0.010	-0.010	-0.010
	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
d2011	0.201***	0.201***	0.201***	0.200***	0.200***	0.201***	0.199***
	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)
d2013	0.031	0.036	0.033	0.034	0.029	0.034	0.032
	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)
d2015	-0.044*	-0.042*	-0.043*	-0.042*	-0.045*	-0.042*	-0.045*
	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)
ln(*Revenue*)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
dDigMediaICT	0.124***	0.118***	0.123***	0.123***	0.124***	0.122***	0.120***
	(0.020)	(0.021)	(0.020)	(0.020)	(0.020)	(0.020)	(0.021)
dLifeSciHealth	0.084***	0.079***	0.081***	0.080***	0.086***	0.080	0.086***
	(0.029)	(0.029)	(0.029)	(0.029)	(0.030)	(0.029)	(0.030)
dMaterialMnfg	0.044**	0.041**	0.043**	0.042**	0.045**	0.042**	0.043**
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
dFinland	-0.218***	-0.211***	-0.216***	-0.217***	-0.219***	-0.215***	-0.217***
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.023)
ln(webIndex)	-0.011						
	(0.014)						
ln(webInnov)		0.010					0.016**
		(0.008)					(0.009)

Table D-6. OLS regression results (continued)

Variables	**1**	**2**	**3**	**4**	**5**	**6**	**7**
ln(webCollab)			-0.005				-0.003
			(0.009)				(0.010)
ln(webOpen)				-0.018			-0.002
				(0.015)			(0.004)
ln(webRD)					-0.037*		-0.026**
					(0.024)		(0.016)
ln(webIP)						0.000	0.000
						(0.002)	(0.002)
Constant	0.464***	0.440***	0.456***	0.456***	0.437***	0.453***	0.456***
	(0.034)	(0.033)	(0.032)	(0.032)	(0.045)	(0.032)	(0.034)
Nb obs.	1289	1289	1289	1289	1289	1289	1289
R^2	0.219	0.220	0.219	0.219	0.220	0.219	0.222
F	29.869***	29.967***	29.829***	29.821***	29.959***	29.803***	22.682***
Variables	**1´**	**2´**	**3´**	**4´**	**5´**	**6´**	**7´**
2. Investment							
Investment							
TimetoMarket							
TechFirm	0.485***	0.485***	0.481***	0.507***	0.430***	0.448***	0.402***
	(0.088)	(0.088)	(0.088)	(0.088)	(0.090)	(0.088)	(0.090)
dYoung	0.153***	0.148**	0.142**	0.148**	0.148**	0.163**	0.151**
	(0.084)	(0.084)	(0.084)	(0.084)	(0.084)	(0.084)	(0.083)
d2011	-0.264**	-0.273*	-0.257*	-0.224*	-0.263*	-0.224*	-0.191
	(0.171)	(0.171)	(0.170)	(0.171)	(0.170)	(0.170)	(0.169)
d2013	-0.342*	-0.380***	-0.368***	-0.387***	-0.329**	-0.362***	-0.293**
	(0.152)	(0.151)	(0.151)	(0.151)	(0.151)	(0.150)	(0.150)
d2015	0.180**	0.154	0.175	0.151	0.178	0.170	0.200*
	(0.153)	(0.153)	(0.152)	(0.152)	(0.152)	(0.151)	(0.151)

Table D-6. OLS regression results (continued)

Variables	**1´**	**2´**	**3´**	**4´**	**5´**	**6´**	**7´**
ln(Revenue)	0.019	0.019	0.011	0.016	0.020	0.021*	0.016
	(0.015)	(0.015)	(0.016)	(0.015)	(0.015)	(0.015)	(0.015)
dDigMediaICT	0.161	0.153*	0.158*	0.143*	0.182**	0.192**	0.145*
	(0.10)	(0.101)	(0.100)	(0.100)	(0.099)	(0.099)	(0.100)
dLifeSciHealth	1.068*	1.127***	1.092***	1.156***	1.039***	1.075***	1.015***
	(0.142)	(0.141)	(0.141)	(0.141)	(0.142)	(0.140)	(0.142)
dMaterialMnfg	-0.042***	-0.028	-0.030	-0.003	-0.041	-0.031	-0.028
	(0.124)	(0.123)	(0.123)	(0.123)	(0.123)	(0.123)	(0.122)
dFinland	0.408	0.373***	0.362***	0.434***	0.405***	0.405***	0.497***
	(0.107)	(0.106)	(0.105)	(0.109)	(0.106)	(0.105)	(0.108)
ln(webIndex)	0.242***						
	(0.068)						
ln(webInnov)		0.089**					0.008
		(0.040)					(0.043)
ln(webCollab)			0.183***				0.125***
			(0.045)				(0.048)
ln(webOpen)				0.049***			0.027*
				(0.020)			(0.020)
ln(webRD)					0.266***		0.141**
					(0.072)		(0.079)
ln(webIP)						0.042***	0.036***
						(0.009)	(0.009)
Constant	0.590***	0.735***	0.741***	0.770***	0.639***	0.717***	0.509***
	(0.158)	(0.150)	(0.144)	(0.145)	(0.152)	(0.143)	(0.156)
Nb obs.	1289	1289	1289	1289	1289	1289	1289
R^2	0.129	0.123	0.131	0.124	0.129	0.137	0.138
F	17.157***	16.343***	17.527***	16.475***	17.252***	18.396***	14.534***

Notes: Standard errors in parentheses. Unilateral level of significance test. ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table D-7. OLS regression results – *post-hoc*

Investment	
oInnovEvidence	0.049
	(0.061)
TechFirm	0.438***
	(0.089)
dYoung	0.152*
	(0.085)
d2011	-0.278
	(0.195)
d2013	-0.346**
	(0.141)
d2015	0.164
	(0.181)
ln(Revenue)	0.019
	(0.018)
dICT	0.149
	(0.102)
dLifeSciHealth	1.068***
	(0.177)
dMatManuf	-0.043
	(0.119)
dFinland	0.420***
	(0.111)
ln(webIndex)	0.242***
	(0.070)
Constant	0.494***
	(0.199)
Nb obs.	1,289
R^2	0.129
F	(0.06)***

Note: Standard errors in parentheses. Unilateral level of significance test. ***, **, * show significance at the 1%, 5% and 10% level respectively.

PLS methodology

To assess the validity of the web-based construct, classic partial least squares (PLS) regressions were estimated[78]. A PLS model consists of three parts: a structural component, which shows the connection between latent variables (the inner model); a measurement component (the outer model), which reveals how the latent variables and their indicators are associated; and a weight relation component, which is used to calculate case values for the latent variable (Hair et al., 2011). PLS has shown to be a robust method when performing complex structural equation models (SEM) when the conditions for confirmatory factor analysis (CFA) were not met, such as a small sample size, non-normality or having nonlinear relationships (Hair et al., 2011; Kock, 2022). PLS is also a better method to build latent formative construct (Hair et al., 2011).

Before treating the keywords into variables, some pretreatment was required. First, to be coherent with the other keywords that were built with their stemmed version, we combined the frequencies of “create”, “creativ” and “creation” into the item “creat”. Next, we noticed the presence of 15 items, from which 14 are items representing the open innovation factor, are constituted with more than 95% of 0 observation (see Table C-5). These items are “business ecosystem”, “public license”, “spin off”, “crowdfunding”, “innovation ecosystem”, “crowdsourcing”, “open innov”, “design thinking”, “open data”, “codev”, “early adopter”, “industrial design”, “technology transfer”, “open source” and “consorti”. We combined these items to increase the variance needed to perform PLS regressions. All the items that constitute the open innovation factor were added together in a group call GroupOI, which resulted in 91.26% of 0 observation. We then combined “industrial design” with “secret” into OtherIP which has 88.69% of 0 observation.

A first-order PLS was generated by building a formative latent variable by allocating each standardised keyword to its corresponding factor. As a result, the only item that did not load positively and significantly was “affiliat” from the collaboration construct. This keyword was removed from the web-based collaboration construct to generate the final results. Then, a second-order model where all five indicators explained their corresponding innovation concept was built. Following the formative construct guidelines from Hair et al. (2011), the estimation generated positive and significant loadings and weights, providing empirical support for the construct. Furthermore, all the relevant tests to ensure the construct’s validity is adequate. All Variance

[78] We used the WarpPLS 8.0 software with Warp3 inner model analysis algorithm and stable3 option.

Inflation Factors (VIF) are well below 2.5 which shows the absence of multicollinearity. All indicators effect sizes are equal or superior to 0.02 (Kock, 2014). The resulting distributions for each formative variable are unimodal with the Rohatgi-Székely test (Rohatgi & Székely, 1989) and the Klaassen-Mokveld-van Es test (Klaassen et al., 2000). The average variance extracted for the first five latent variables in the first-order can be seen in Table D-12 and the average variance extracted from the second-order model is 0.292.

Adding each item multiplied by the indicator weight obtained from the first-order PLS regressions and dividing the results by the sum of all the weights for each concept, six more variables were then generated: *webInnovPLS*, *webCollabPLS*, *webOpenPLS*, *webRDPLS* and *webIPPLS*. Finally, the index *webIndexPLS* was built by adding each of these five variables multiplied by their corresponding indicator weight obtained from the second-order PLS regression and dividing the result by the sum of all the weights. These indicators therefore take into account each factor's and keyword's individual contribution to explain a higher concept, which we deem are more precise signals. Results from the PLS regressions are provided in Table D-8 to Table D-15. A summary of the web-based variable construct is presented in Table D-3 along with descriptive and correlation results in Table D-4 and Table D-5.

Finally, results of the 3SLS regression with the PLS construct can be seen in Table D-16. No significant differences between the results obtained with our two types (simple and PLS) web-based constructs are observed. Whether each individual contribution to a concept is optimised or not, the measures captured remains consistent, hinting that the model is robust.

Table D-8. First order - Combined loadings and cross-loadings - first iteration

	Innov	**Collab**	**Open**	**RD**	**IP**	**SE**	**P value**
innov	0.54	-0.04	-0.11	0.40	-0.08	0.03	<0.001
idea	0.55	0.07	0.02	-0.14	0.08	0.03	<0.001
creat	0.69	-0.03	0.07	-0.21	0.00	0.03	<0.001
collab	0.00	0.77	0.01	0.12	-0.02	0.03	<0.001
affiliat	0.07	-0.06	-0.04	0.02	0.00	0.03	0.02
cooperat	-0.12	0.14	-0.05	0.04	-0.05	0.03	<0.001
partner	0.03	0.79	0.00	-0.12	0.03	0.03	<0.001
GroupOI	-0.03	-0.03	0.74	0.09	0.01	0.03	<0.001
community	0.02	0.04	0.53	-0.13	-0.06	0.03	<0.001
startup	0.01	-0.01	0.55	0.01	0.04	0.03	<0.001
tech	0.02	0.14	0.09	0.56	0.24	0.03	<0.001
research	-0.01	-0.13	-0.07	0.66	-0.10	0.03	<0.001
R&D	-0.11	0.01	0.02	0.35	0.03	0.03	<0.001
develop	0.17	0.11	0.06	0.50	0.08	0.03	<0.001
invent	0.03	-0.07	0.08	0.25	0.05	0.03	<0.001
scien	-0.05	-0.19	-0.08	0.66	-0.15	0.03	<0.001
lab	0.01	0.28	-0.05	0.39	-0.03	0.03	<0.001
experiment	-0.10	-0.03	0.02	0.31	-0.06	0.03	<0.001
patent	-0.05	-0.07	-0.01	0.16	0.71	0.03	<0.001
OtherIP	0.05	0.07	0.01	-0.16	0.71	0.03	<0.001

Notes: Loadings are unrotated and cross-loadings are oblique-rotated. SEs and P-values are for loadings.

Table D-9. First Order - indicator weights - first iteration

	Innov	Collab	Open	RD	IP	SE	P-value	VIF	WLS	ES
innov	0.51	0.00	0.00	0.00	0.00	0.027	<0.001	1.00	1.00	0.27
idea	0.51	0.00	0.00	0.00	0.00	0.027	<0.001	1.00	1.00	0.28
creat	0.65	0.00	0.00	0.00	0.00	0.027	<0.001	1.00	1.00	0.45
collab	0.00	0.62	0.00	0.00	0.00	0.027	<0.001	1.06	1.00	0.47
affiliat	0.00	-0.04	0.00	0.00	0.00	0.028	0.055	1.00	1.00	0.00
cooperat	0.00	0.12	0.00	0.00	0.00	0.028	<0.001	1.01	1.00	0.02
partner	0.00	0.64	0.00	0.00	0.00	0.027	<0.001	1.07	1.00	0.51
GroupOI	0.00	0.00	0.66	0.00	0.00	0.027	<0.001	1.01	1.00	0.48
community	0.00	0.00	0.47	0.00	0.00	0.027	<0.001	1.01	1.00	0.25
startup	0.00	0.00	0.49	0.00	0.00	0.027	<0.001	1.01	1.00	0.27
tech	0.00	0.00	0.00	0.30	0.00	0.027	<0.001	1.17	1.00	0.17
research	0.00	0.00	0.00	0.35	0.00	0.027	<0.001	1.29	1.00	0.24
R&D	0.00	0.00	0.00	0.19	0.00	0.027	<0.001	1.05	1.00	0.07
develop	0.00	0.00	0.00	0.27	0.00	0.027	<0.001	1.13	1.00	0.13
invent	0.00	0.00	0.00	0.14	0.00	0.028	<0.001	1.02	1.00	0.03
scien	0.00	0.00	0.00	0.35	0.00	0.027	<0.001	1.32	1.00	0.23
lab	0.00	0.00	0.00	0.21	0.00	0.027	<0.001	1.07	1.00	0.08
experiment	0.00	0.00	0.00	0.17	0.00	0.028	<0.001	1.03	1.00	0.05
patent	0.00	0.00	0.00	0.00	0.71	0.026	<0.001	1.00	1.00	0.50
OtherIP	0.00	0.00	0.00	0.00	0.71	0.026	<0.001	1.00	1.00	0.50

Notes: P values < 0.05 and VIFs < 2.5 are desirable for formative indicators; VIF = indicator variance inflation factor; WLS = indicator weight-loading sign (-1 = Simpson's paradox in l.v.); ES = indicator effect size.

Table D-10. First order - Combined loadings and cross-loadings – Final results

	Innov	**Collab**	**Open**	**RD**	**IP**	**SE**	**P value**
innov	0.54	-0.04	-0.11	0.40	-0.08	0.03	<0.001
idea	0.55	0.07	0.02	-0.13	0.08	0.03	<0.001
creat	0.69	-0.03	0.07	-0.21	0.00	0.03	<0.001
collab	0.00	0.77	0.01	0.12	-0.02	0.03	<0.001
affiliat	-0.12	0.14	-0.05	0.04	-0.05	0.03	<0.001
cooperat	0.02	0.79	0.00	-0.12	0.03	0.03	<0.001
partner	-0.03	-0.03	0.74	0.09	0.01	0.03	<0.001
GroupOI	0.02	0.04	0.53	-0.13	-0.06	0.03	<0.001
community	0.01	-0.01	0.55	0.01	0.04	0.03	<0.001
startup	0.01	0.14	0.09	0.56	0.24	0.03	<0.001
tech	0.00	-0.14	-0.07	0.66	-0.10	0.03	<0.001
research	-0.11	0.01	0.02	0.35	0.03	0.03	<0.001
R&D	0.17	0.10	0.06	0.50	0.08	0.03	<0.001
develop	0.03	-0.07	0.08	0.25	0.05	0.03	<0.001
invent	-0.05	-0.20	-0.08	0.66	-0.15	0.03	<0.001
scien	0.01	0.29	-0.05	0.39	-0.03	0.03	<0.001
lab	-0.09	-0.03	0.02	0.31	-0.06	0.03	<0.001
experiment	-0.05	-0.07	-0.01	0.16	0.71	0.03	<0.001
patent	0.05	0.07	0.01	-0.16	0.71	0.03	<0.001

Notes: Loadings are unrotated and cross-loadings are oblique-rotated. SEs and P-values are for loadings.

Table D-11. First Order - indicator weights - final results

	Innov	Collab	Open	RD	IP	SE	P value	VIF	WLS	ES
innov	0.51	0.00	0.00	0.00	0.00	0.03	<0.001	1.00	1.00	0.27
idea	0.51	0.00	0.00	0.00	0.00	0.03	<0.001	1.00	1.00	0.28
creat	0.65	0.00	0.00	0.00	0.00	0.03	<0.001	1.00	1.00	0.45
collab	0.00	0.62	0.00	0.00	0.00	0.03	<0.001	1.06	1.00	0.48
affiliat	0.00	0.12	0.00	0.00	0.00	0.03	<0.001	1.01	1.00	0.02
cooperat	0.00	0.64	0.00	0.00	0.00	0.03	<0.001	1.07	1.00	0.51
partner	0.00	0.00	0.66	0.00	0.00	0.03	<0.001	1.01	1.00	0.48
GroupOI	0.00	0.00	0.47	0.00	0.00	0.03	<0.001	1.01	1.00	0.25
community	0.00	0.00	0.49	0.00	0.00	0.03	<0.001	1.01	1.00	0.27
startup	0.00	0.00	0.00	0.30	0.00	0.03	<0.001	1.17	1.00	0.17
tech	0.00	0.00	0.00	0.35	0.00	0.03	<0.001	1.29	1.00	0.24
research	0.00	0.00	0.00	0.19	0.00	0.03	<0.001	1.05	1.00	0.07
R&D	0.00	0.00	0.00	0.27	0.00	0.03	<0.001	1.13	1.00	0.13
develop	0.00	0.00	0.00	0.14	0.00	0.03	<0.001	1.02	1.00	0.03
invent	0.00	0.00	0.00	0.35	0.00	0.03	<0.001	1.32	1.00	0.23
scien	0.00	0.00	0.00	0.21	0.00	0.03	<0.001	1.07	1.00	0.08
lab	0.00	0.00	0.00	0.17	0.00	0.03	<0.001	1.03	1.00	0.05
experiment	0.00	0.00	0.00	0.00	0.71	0.03	<0.001	1.00	1.00	0.50
patent	0.00	0.00	0.00	0.00	0.71	0.03	<0.001	1.00	1.00	0.50

Notes: P values < 0.05 and VIFs < 2.5 are desirable for formative indicators; VIF = indicator variance inflation factor; WLS = indicator weight-loading sign (-1 = Simpson’s paradox in l.v.); ES = indicator effect size.

Table D-12. Average variances extracted

Innov	Collab	Open	RD	IP
0.356	0.414	0.374	0.234	0.502

Table D-13. Full Collinearity VIFs

Innov	Collab	Open	RD	IP
1.068	1.070	1.025	1.128	1.014

Table D-14. Second order - Combined loadings and cross-loadings

	INDEX	SE	P-value
Innov	0.607	0.027	<0.001
Collab	0.592	0.027	<0.001
Open	0.417	0.027	<0.001
RD	0.732	0.026	<0.001
IP	0.182	0.027	<0.001

Notes: Loadings are unrotated and cross-loadings are oblique-rotated. SEs and P values are for loadings.

Table D-15. Second order - First order - Indicator weights

	INDEX	SE	P value	VIF	WLS	ES
Innov	0.415	0.027	<0.001	1.068	1	0.252
Collab	0.405	0.027	<0.001	1.07	1	0.239
Open	0.285	0.027	<0.001	1.025	1	0.119
RD	0.501	0.027	<0.001	1.128	1	0.367
IP	0.125	0.028	<0.001	1.014	1	0.023

Notes: P values < 0.05 and VIFs < 2.5 are desirable for formative indicators; VIF = indicator variance inflation factor; WLS = indicator weight-loading sign (-1 = Simpson's paradox in l.v.); ES = indicator effect size

Table D-16. 3SLS regression results - PLS web-based construct

Variables	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1. InnovEvidence							
Investment	0.006	0.004	0.005	0.006	0.005	0.030	0.047
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.041)	(0.049)
TimetoMarket	-0.209***	-0.211***	-0.212***	-0.204***	-0.213***	-0.216***	-0.199***
	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.073)	(0.073)
TechFirm							
dYoung	-0.010	-0.011	-0.010	-0.010	-0.010	-0.010	-0.010
	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
d2011	0.201***	0.201***	0.201***	0.200***	0.200***	0.201***	0.199***
	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)
d2013	0.031	0.036	0.033	0.034	0.029	0.034	0.032
	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)
d2015	-0.044*	-0.042*	-0.043*	-0.042*	-0.045*	-0.042*	-0.045*
	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)
ln(*Revenue*)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
dDigMediaICT	0.124***	0.118***	0.123***	0.123***	0.124***	0.122***	0.120***
	(0.020)	(0.021)	(0.020)	(0.020)	(0.020)	(0.020)	(0.021)
dLifeSciHealth	0.046	0.053	0.051	0.051	0.042	0.050	0.043
	(0.057)	(0.056)	(0.055)	(0.056)	(0.059)	(0.057)	(0.061)
dMaterialMnfg	0.041*	0.039*	0.040*	0.039*	0.041*	0.040*	0.041*
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
dFinland	-0.231***	-0.219***	-0.224***	-0.223***	-0.236***	-0.225***	-0.235***
	(0.028)	(0.026)	(0.026)	(0.028)	(0.029)	(0.028)	(0.034)
ln(webIndex)	-0.027						
	(0.023)						
ln(webInnov)		0.008					0.015**
		(0.011)					(0.010)

Table D-16. 3SLS regression results - PLS web-based construct (continued)

Variables	**1**	**2**	**3**	**4**	**5**	**6**	**7**
ln(webCollab)			-0.007				-0.006
			(0.009)				(0.009)
ln(webOpen)				0.000			0.000
				(0.004)			(0.004)
ln(webRD)					-0.037*		-0.040**
					(0.024)		(0.022)
ln(webIP)						-0.001	-0.001
						(0.002)	(0.002)
Constant	0.442***	0.424***	0.433***	0.430***	0.437***	0.431***	0.429***
	(0.043)	(0.044)	(0.045)	(0.044)	(0.045)	(0.045)	(0.042)
Nb obs.	1,289	1,289	1,289	1,289	1,289	1,289	1,289
R^2	0.200	0.212	0.208	0.209	0.191	0.207	0.189
F	352.52***	358.02***	355.19***	355.84***	349.35***	354.65***	351.52***
Variables	**1´**	**2´**	**3´**	**4´**	**5´**	**6´**	**7´**
2. Investment							
Investment							
TimetoMarket							
TechFirm	0.485***	0.485***	0.481***	0.507***	0.430***	0.448***	0.402***
	(0.088)	(0.088)	(0.088)	(0.088)	(0.090)	(0.088)	(0.090)
dYoung	0.153***	0.148**	0.142**	0.148**	0.148**	0.163**	0.151**
	(0.084)	(0.084)	(0.084)	(0.084)	(0.084)	(0.084)	(0.083)
d2011	-0.264**	-0.273*	-0.257*	-0.224*	-0.263*	-0.224*	-0.191
	(0.171)	(0.171)	(0.170)	(0.171)	(0.170)	(0.170)	(0.169)
d2013	-0.342*	-0.380***	-0.368***	-0.387***	-0.329**	-0.362***	-0.293**
	(0.152)	(0.151)	(0.151)	(0.151)	(0.151)	(0.150)	(0.150)
d2015	0.180**	0.154	0.175	0.151	0.178	0.170	0.200*
	(0.153)	(0.153)	(0.152)	(0.152)	(0.152)	(0.151)	(0.151)

Table D-16. 3SLS regression results - PLS web-based construct (continued)

Variables	**1´**	**2´**	**3´**	**4´**	**5´**	**6´**	**7´**
ln(Revenue)	0.019	0.019	0.011	0.016	0.020	0.021*	0.016
	(0.015)	(0.015)	(0.016)	(0.015)	(0.015)	(0.015)	(0.015)
dDigMediaICT	0.161	0.153*	0.158*	0.143*	0.182**	0.192**	0.145*
	(0.10)	(0.101)	(0.100)	(0.100)	(0.099)	(0.099)	(0.100)
dLifeSciHealth	1.071***	1.127***	1.092***	1.156***	1.039***	1.075***	1.015***
	(0.142)	(0.141)	(0.141)	(0.141)	(0.142)	(0.140)	(0.142)
dMaterialMnfg	-0.038	-0.028	-0.030	-0.003	-0.041	-0.031	-0.028
	(0.123)	(0.123)	(0.123)	(0.123)	(0.123)	(0.123)	(0.122)
dFinland	0.404***	0.373***	0.362***	0.434***	0.405***	0.405***	0.497***
	(0.106)	(0.106)	(0.105)	(0.109)	(0.106)	(0.105)	(0.108)
ln(webIndex)	0.296***						
	(0.084)						
ln(webInnov)		0.010**					0.008
		(0.047)					(0.049)
ln(webCollab)			0.131***				0.075**
			(0.035)				(0.037)
ln(webOpen)				0.048***			0.026**
				(0.015)			(0.016)
ln(webRD)					0.322***		0.196**
					(0.080)		(0.087)
ln(webIP)						0.031***	0.025***
						(0.006)	(0.006)
Constant	0.621***	0.745***	0.759***	0.735***	0.637***	0.718***	0.499***
	(0.154)	(0.149)	(0.143)	(0.145)	(0.150)	(0.143)	(0.154)
Nb obs.	1,289	1,289	1,289	1,289	1,289	1,289	1,289
R^2	0.129	0.123	0.130	0.127	0.131	0.137	0.150
F	190.17***	181.16***	192.18***	187.34***	194.22***	204.51***	227.74***

Notes: Standard errors in parentheses. Unilateral level of significance test. ***, **, * show significance at the 1%, 5% and 10% level respectively.

APPENDIX E – SUPPLEMENTARY MATERIALS ARTICLE 3

Table E-1. Sampling summary

# program	Year	Invitations	Entries	Websites found	Firms	Websites retrieved	Final
1	2010	53	33	26	25	18	0
2	2010	72	31	25	23	21	13
3	2011	754	282	194	184	151	67
4	2011	224	71	40	40	37	31
5	2012	142	68	62	61	34	19
6	2012	110	54	48	44	31	24
7	2012	812	382	264	226	194	132
8	2012	207	94	73	71	60	28
9	2012	305	163	45	41	34	13
10	2012	73	49	22	21	15	11
11	2012	29	21	15	15	14	0
12	2012	59	27	18	18	13	13
13	2013	605	260	138	130	116	87
14	2013	894	313	100	98	80	50
15	2013	151	109	75	75	68	0
16	2013	200	140	75	74	60	0
17	2013	34	29	19	17	15	6
18	2014	34	29	27	27	23	20
19	2014	108	87	68	53	44	37
20	2014	23	23	17	15	14	0
21	2014	3952	1307	929	905	797	739
22	2014	82	48	44	43	38	30
23	2015	545	222	174	168	153	89
24	2015	750	230	196	166	136	65
25	2015	323	150	104	99	87	51
26	2015	38	23	21	19	19	8

Table E-1. Sampling summary (continued)

# program	Year	Invitations	Entries	Websites found	Firms	Websites retrieved	Final
27	2016	38	30	21	19	16	13
28	2016	1529	425	74	72	61	28
29	2016	56	40	35	31	30	20
n		12202	4740	2949	2780	2379	1594
%			39%	62%	94%	86%	67%

Table E-2. Average of sentences obtained by sector and country

	Materials & manufacturing	**Life science & health**	**Digital media & ICT**	**Natural resources, energy & environment**	**Other**	**Canada**	**Finland**
Number of sentences	123.20	228.57	263.47	159.69	190.64	246.71	188.40
Number of pages	14.83	16.46	15.83	15.75	16.17	17.37	14.72

Table E-3. Dependent variables

Concepts	Variable	Type	Description
Overall innovation signal	countAll	Discrete	Number of all sentences with innovation-related terms on a firm's website
Innovation culture signal	countInnov	Discrete	Number of all sentences with innovation culture-related terms on a firm's website
Collaboration signal	countCollab	Discrete	Number of all sentences with collaboration-related terms on a firm's website
OI signal	countOpen	Discrete	Number of all sentences with OI-related terms on a firm's website
R&D signal	countRD	Discrete	Number of all sentences with R&D-related terms on a firm's website
IP signal	countIP	Discrete	Number of all sentences with IP-related terms on a firm's website

Table E-4. Independent and control variables

Concepts	Variable	Type	Description
Innovation evidence	dNoInnovEvidence	Binary	Dummy variable that takes the value 1 if the firm exhibits no innovation evidence, and 0 otherwise.
	dSomeInnovEvidence	Binary	Dummy variable that takes the value 1 if the firm exhibits some innovation evidence, and 0 otherwise.
	dStrongInnovEvidence	Binary	Dummy variable that takes the value 1 if the firm exhibits strong innovation evidence, and 0 otherwise.
Time to market	TimetoMarket	Continuous	Number of months needed to introduce new or improved products, processes, or services to the market
	dLongTMkt	Binary	Dummy variable that takes the value 1 if the firm is in the group with a longer time to market (*TimetoMarket*), and 0 otherwise.

Table E-4. Independent and control variables (continued)

Concepts	Variable	Type	Description
	dAverageTMkt	Binary	Dummy variable that takes the value 1 if the firm is neither in the group with a longer or a shorter time to market (*TimetoMarket*), and 0 otherwise.
	dShortTMkt	Binary	Dummy variable that takes the value 1 if the firm is in the group with a shorter time to market (*TimetoMarket*), and 0 otherwise.
Knowledge intensity (KI)	dLowKI	Binary	Dummy variable that takes the value 1 if the firm is in low-tech industries and other services, and 0 otherwise.
	dHighKI	Binary	Dummy variable that takes the value 1 if the firm is in advanced knowledge services or a tech firm, and 0 otherwise.
Sectors	dMaterialMnfg	Binary	Dummy variable that takes the value 1 if the firm is in the Materials & Manufacturing sector, and 0 otherwise.
	dLifeSciHealth	Binary	Dummy variable that takes the value 1 if the firm is in the Life sciences & health sector, and 0 otherwise.
	dDigMediaICT	Binary	Dummy variablesthat takes the value 1 if the firm is in the Digital media & ICT sector, and 0 otherwise.
	dEnerEnvironNat	Binary	Dummy variable that takes the value 1 if the firm is in the Natural resources, energy & environment sector, and 0 otherwise.
	dOther	Binary	Dummy variable that takes the value 1 if the firm is in any other sector, and 0 otherwise.
Sectorial category	dSupplierDominated	Binary	Dummy variable that takes the value 1 if the firm is in the Supplier-dominated sector, and 0 otherwise.
	dSpecializedSupplier	Binary	Dummy variable that takes the value 1 if the firm is in the Specialised supplier sector, and 0 otherwise.

Table E-4. Independent and control variables (continued)

Concepts	Variable	Type	Description
	dKIBS	Binary	Dummy variable that takes the value 1 if the firm is in the KIBS sector, and 0 otherwise.
	dMassProd	Binary	Dummy variable that takes the value 1 if the firm is in the Mass production sector, and 0 otherwise.
Standardised sentence ratio	SentenceRatio	Continuous	Total number of sentences divided by maximum sentences by size and sector
Size	nbEmployee	Continuous	Number of employees of the firm
	dVerySmall	Binary	Dummy variable that takes the value 1 if the firm has fewer than 10 employees, and 0 otherwise.
	dSmall	Binary	Dummy variable that takes the value 1 if the firm has 11-99 employees, and 0 otherwise.
	dMedium	Binary	Dummy variable that takes the value 1 if the firm has 100-499 employees, and 0 otherwise.
	dLarge	Binary	Dummy variable that takes the value 1 if the firm has 500 and more employees, and 0 otherwise.
Age	Age	Discrete	Age at the time of the questionnaire
	dEarlyStage	Binary	Dummy variable that takes the value 1 if the firm is 3 years old or younger at the time of the questionnaire, and 0 otherwise
	dYoung	Binary	Dummy variable that takes the value 1 if the firm is between 4 and 8 years old at the time of the questionnaire, and 0 otherwise
	dMature	Binary	Dummy variable that takes the value 1 if the firm is at least 9 years old at the time of the questionnaire, and 0 otherwise

Table E-4. Independent and control variables (continued)

Concepts	Variable	Type	Description
Country	*dCanada*	Binary	Dummy variable that takes the value 1 if the firm is located in Canada, and 0 otherwise.
	dFinland	Binary	Dummy variable that takes the value 1 if the firm is located in Finland, and 0 otherwise.
Period of investigation	*d2010_2011*	Binary	Dummy variable that takes the value 1 if the firm has responded to the survey in 2010 or 2011, and 0 otherwise.
	d2012_2013	Binary	Dummy variable that takes the value 1 if the firm has responded to the survey in 2012 or 2013, and 0 otherwise.
	d2014	Binary	Dummy variable that takes the value 1 if the firm has responded to the survey in 2014, and 0 otherwise.
	d2015_2016	Binary	Dummy variable that takes the value 1 if the firm has responded to the survey in 2015 or 2016, and 0 otherwise.

Table E-5. Descriptive statistics

Concepts	Variables	Mean	Std. Dev.	Min	Median	Max	Avg countAll
Signal	Overall innovation *(countAll)*	50.040	92.466	0 [9.16%]	19	1133	-
	Innovation culture (*countInnov*)	10.103	20.886	0 [27.85%]	3	275	-
	Collaboration (*countCollab*)	5.242	13.652	0 [41.59%]	1	237	-
	Open innovation (*countOpen*)	2.065	8.610	0 [71.52%]	0	161	-
	R&D (*countRD*)	22.181	48.519	0 [17.38%]	7	784	-
	IP (*countIP*)	0.788	3.590	0 [78.86%]	0	107	-
	Overall innovation *(ratioAll)*	0.089	0.093	0.000	0.063	1.000	-
	Innovation culture (ratio*Innov*)	0.044	0.080	0.000	0.023	1.000	-
	Collaboration (*ratioCollab*)	0.028	0.060	0.000	0.009	1.000	-
	Open innovation (*ratioOpen*)	0.016	0.060	0.000	0.000	1.000	-
	R&D (*ratioRD*)	0.089	0.103	0.000	0.056	1.000	-
	IP (*ratioIP*)	0.013	0.054	0.000	0.000	1.000	-
Innovation evidence	*dNoInnovEvidence*	0.231	0.422	0	0	1	40.01
	dSomeInnovEvidence	0.544	0.498	0	1	1	55.58
	dStrongInnovEvidence	0.225	0.417	0	0	1	46.97
Time market	*TimetoMarket*	20.594	16.505	6	18	108	-
	dLongTMkt	0.297	0.457	0	0	1	63.02
	dAverageTMkt	0.361	0.481	0	0	1	41.41
	dShortTMkt	0.341	0.474	0	0	1	47.87

Table E-5. Descriptive statistics (continued)

Concepts	Variables	Mean	Std. Dev.	Min	Median	Max	Avg countAll
Knowledge level	*dLowKI*	0.304	0.460	0	1	1	27.40
	dHighKI	0.696	0.460	0	1	1	59.94
Sector	*dMaterialMnfg*	0.161	0.367	0	0	1	30.56
	dLifeSciHealth	0.110	0.314	0	0	1	92.34
	dDigMediaICT	0.277	0.448	0	0	1	63.48
	dEnerEnvironNat	0.074	0.262	0	0	1	36.44
	dOther	0.378	0.485	0	0	1	38.76
Sectorial category	*dSupplierDominated*	0.306	0.461	0	0	1	25.15
	dSpecializedSupplier	0.385	0.487	0	0	1	47.47
	dKIBS	0.339	0.474	0	0	1	62.29
	dMassProd	0.027	0.162	0	0	1	161.81
Sentence ratio	*SentenceRatio*	0.112	0.165	0.000	0.053	1.000	-
Size	*nbEmployee*	361.908	5,299.790	1	10	181,000	-
	dVerySmall	0.602	0.490	0	1	1	39.88
	dSmall	0.329	0.470	0	0	1	53.14
	dMedium	0.036	0.187	0	0	1	82.34
	dLarge	0.032	0.176	0	0	1	172.71
Age	*Age*	7.989	9.289	0	5	114	-
	dEarlyStage	0.332	0.471	0	0	1	36.32
	dYoung	0.386	0.487	0	0	1	46.20
	dMature	0.282	0.450	0	0	1	71.49
Country	*dFinland*	0.519	0.500	0	1	1	31.94
	dCanada	0.481	0.500	0	0	1	69.61

Table E-5. Descriptive statistics (continued)

Concepts	Variables	Mean	Std. Dev.	Min	Median	Max	Avg countAll
Period	*d2010_2011*	0.070	0.255	0	0	1	48.02
	d2012_2013	0.240	0.427	0	0	1	59.52
	d2014	0.518	0.500	0	1	1	38.45
	d2015_2016	0.172	0.377	0	0	1	72.55

Table E-6. Pearson's correlation table

		1	2	3	4	5	6	7	8	9	10	11
countAll	**1**	1.000										
countInnov	**2**	0.836 *	1.000									
countCollab	**3**	0.675 *	0.663 *	1.000								
countOpen	**4**	0.445 *	0.417 *	0.360 *	1.000							
countRD	**5**	0.921 *	0.682 *	0.560 *	0.318 *	1.000						
countIP	**6**	0.355 *	0.218 *	0.154 *	0.047	0.288 *	1.000					
dSomeInnovEvidence	**7**	0.065 *	0.039	0.070 *	0.021	0.063 *	0.029	1.000				
dStrongInnovEvidence	**8**	-0.018	0.019	-0.027	0.005	-0.022	-0.003	-0.588 *	1.000			
dLonger	**9**	0.091 *	0.072 *	0.032	-0.008	0.076 *	0.081 *	0.136 *	-0.031	1.000		
dShorter	**10**	-0.017	0.040	0.023	0.044	-0.040	-0.047	-0.069 *	0.076 *	-0.468 *	1.000	
dHighKI	**11**	0.162 *	0.107 *	0.100 *	0.007	0.150 *	0.085 *	0.098 *	0.046	0.147 *	-0.099 *	1.000
dMaterialMnfg	**12**	-0.092 *	-0.104 *	-0.079 *	-0.080 *	-0.082 *	-0.034	-0.045	-0.027	0.003	-0.077 *	0.130 *
dLifeSciHealth	**13**	0.161 *	0.060 *	0.037	0.000	0.220 *	0.099 *	0.017	0.002	0.003	-0.064 *	0.085 *
dDigMediaICT	**14**	0.090 *	0.162 *	0.096 *	0.062 *	0.033	-0.021	0.086 *	0.097 *	0.219 *	0.116 *	0.129 *
dOther	**15**	-0.095 *	-0.081 *	-0.028	0.013	-0.096 *	-0.012	-0.087 *	-0.044	-0.195 *	0.034	-0.270 *
dSupplierDominated	**16**	-0.164 *	-0.111 *	-0.098 *	-0.004	-0.152 *	-0.088 *	-0.097 *	-0.048	-0.155 *	0.102 *	-0.984 *
dKIBS	**17**	0.095 *	0.174 *	0.135 *	0.094 *	0.041	0.010	0.063 *	0.125 *	0.174 *	0.104 *	0.203 *
dMassProd	**18**	0.201 *	0.133 *	0.044	0.055 *	0.189 *	0.069 *	0.083 *	-0.062 *	0.137 *	-0.071 *	0.102 *
dVerySmall	**19**	-0.135 *	-0.083 *	-0.111 *	-0.034	-0.128 *	-0.074 *	-0.060 *	0.057 *	-0.080 *	0.055 *	-0.173 *
dMedium	**20**	0.068 *	0.020	0.039	0.026	0.033	0.120 *	0.070 *	-0.016	0.057 *	-0.034	0.092 *

Table E-6 Pearson's correlation table (continued)

		1	2	3	4	5	6	7	8	9	10	11
dLarge	**21**	0.241 *	0.207 *	0.101 *	0.109 *	0.218 *	0.075 *	0.081 *	-0.055 *	0.155 *	-0.093 *	0.097 *
dYoung	**22**	-0.033	-0.034	0.008	-0.056 *	-0.019	-0.052 *	0.038	-0.016	-0.008	0.014	0.023
dMature	**23**	0.145 *	0.086 *	0.054 *	0.078 *	0.133 *	0.115 *	-0.026	0.031	0.130 *	-0.101 *	0.017
dCanada	**24**	0.204 *	0.178 *	0.119 *	0.149 *	0.166 *	0.110 *	0.036	0.171 *	0.108 *	0.023	0.101 *
SentenceRatio	**25**	0.493 *	0.424 *	0.331 *	0.233 *	0.423 *	0.212 *	0.058 *	0.007	0.058 *	-0.019	0.103 *
d2010_2011	**26**	-0.006	0.002	0.013	-0.012	-0.017	-0.020	0.013	0.095 *	-0.027	0.110 *	0.052 *
d2012_2013	**27**	0.058 *	0.110 *	0.043	0.094 *	0.005	0.003	-0.042	0.137 *	0.039	0.038	0.015
d2015_2016	**28**	0.111 *	0.040	0.044	0.085 *	0.114 *	0.099 *	0.000	0.050 *	0.093 *	-0.075 *	-0.057 *

		12	13	14	15	16	17	18	19	20	21	22
dMaterialMnfg	**12**	1.000										
dLifeSciHealth	**13**	-0.154 *	1.000									
dDigMediaICT	**14**	-0.271 *	-0.218 *	1.000								
dOther	**15**	-0.341 *	-0.274 *	-0.483 *	1.000							
dSupplierDominated	**16**	-0.131 *	-0.086 *	-0.135 *	0.269 *	1.000						
dKIBS	**17**	-0.296 *	-0.248 *	0.841 *	-0.329 *	-0.212 *	1.000					
dMassProd	**18**	0.012	0.077 *	-0.051 *	-0.034	-0.111 *	-0.119 *	1.000				
dVerySmall	**19**	-0.070 *	-0.086 *	0.117 *	0.044	0.167 *	0.087 *	-0.205 *	1.000			
dMedium	**20**	0.034	0.017	-0.053 *	-0.055 *	-0.071 *	-0.033	0.216 *	-0.239 *	1.000		
dLarge	**21**	-0.002	0.016	-0.025	-0.031	-0.098 *	-0.002	0.652 *	-0.224 *	-0.035	1.000	
dYoung	**22**	0.046	0.033	-0.025	-0.011	-0.018	-0.021	-0.084 *	-0.009	0.004	-0.108 *	1.000
dMature	**23**	0.060 *	0.029	-0.139 *	-0.005	-0.020	-0.107 *	0.214 *	-0.323 *	0.139 *	0.251 *	-0.496 *

Table E-6 Pearson's correlation table (continued)

		24	25	26	27	28
dCanada	**24**	1.000				
SentenceRatio	**25**	0.088 *	1.000			
d2010_2011	**26**	0.284 *	-0.031	1.000		
d2012_2013	**27**	0.585 *	0.012	-0.154 *	1.000	
d2015_2016	**28**	0.178 *	0.154 *	-0.125 *	-0.256 *	1.000

Notes: Two-tailed test. *p<0.05. Performed with natural logarithm transformed variables when relevant.

One way ANOVA results

One-way ANOVA[79] analysis is performed using the signalling ratios to explore the differences within each categorical variable in terms of different signalling dimensions. First, a Kruskal-Wallis test is performed whenever a categorical variable as more than two categories to assess the overall differences in mean rank. Because the variables are not normally distributed and some categories include a small number of firms, the Mann-Whitney *U* test is applied to assess one-on-one differences between categories on the different signalling dimensions. Finally, Wilcoxon rank tests are performed to assess the differences in different signalling dimension ranking by category. These non-parametric tests are appropriate to assess the differences in signalling between items that have a small number of observations (e.g., between scale-based and science-based firms.)

Table E-7 displays the differences in innovation evidence profiles among firms. As one may expect, firms with no innovation evidence ranked third on all metrics, with significant differences observed in all metrics except collaboration when compared to firms with strong innovation evidence. Firms with some innovation evidence displayed a stronger signal for overall innovation, collaboration, and R&D compared to firms with strong innovation evidence, while firms with strong innovation evidence exhibited a stronger innovation culture signal compared to those with some evidence. However, there were no significant differences between these two groups in terms of open innovation or IP signalling. When it comes to differences in terms of dimensions signalled, firms with some innovation evidence are displaying more R&D signal, followed innovation culture and collab. Firms with strong evidence are also doing more R&D signalling followed with OI and Collab.

Table E-8 illustrates the differences in signalling dimensions from firms among different time to market profiles. Our findings indicate that firms with a longer time to market consistently exhibit a stronger signal for innovation-related terms compared to those in the average group. Additionally, these firms displayed a significantly stronger signal in overall innovation, R&D, and IP compared to firms with a shorter time to market. Conversely, firms with a shorter time to market ranked have stronger overall innovation and open innovation signals compared to those in the average group. Among firms with a longer time to market, we can see a stronger R&D signal, followed with innovation culture and collab. For the firms with a shorter time to market, all the

[79] We used IBM SPSS Statistics Version 29.

signalling dimensions seem to be closer to each other, although they remain statistically different, with R&D first, followed again by innovation culture and collaboration.

Table E-9 shows the differences in signalling across different knowledge intensity levels. Our analysis indicates that, as expected, firms with high knowledge intensity exhibit a stronger signal across all innovation factors, with the exception of open innovation. Firms with higher knowledge intensity have stronger signals in R&D followed by innovation culture and collaboration.

Table E-10 shows the results between sectors. At first glance, firms in Life science and health and Digital media and ICT have a stronger signal followed by Energy and environment, Natural resources, Materials and manufacturing and Other. Firms in Life sciences and health have significantly stronger signal than firms in Materials and manufacturing and Other, and they also have a stronger signal across all innovation factors except for open innovation, compared to firms in Natural resources, Energy and environment sector. Furthermore, there are meaningful differences between firms in Life sciences and health and Digital media and ICT. Firms in Life sciences have a stronger signal in terms of overall innovation, as well as collaboration, R&D, and IP. Firms in Digital media and ICT have a stronger signal compared to Materials and manufacturing firms in terms of innovation overall, as well as innovation culture, collaboration, and open innovation. They also have a stronger signal than the Natural resources firms with innovation overall and innovation culture. Additionally, they have a stronger signal than firms in the Energy and environment sector in terms of innovation culture and collaboration, but less in terms of IP. Finally, firms in Digital media and ICT have a stronger signal for all the dimensions than firms in the sector labelled as Other. Finally, within each sector, R&D arrives first, innovation culture second, and collaboration third.

Table E-11 presents the results with the sectorial categories. Among all the firms, Science-based firms have the strongest innovation overall and R&D signals. Additionally, they have a stronger collaboration signal than scale-based industries. This is not surprising as science-based firms are characterised by higher levels of innovativeness, cumulativeness, appropriability through industrial design, process-product orientation, marketing innovation, and intramural R&D than scale-based industries (Castellacci, 2008). KIBS and specialised-suppliers does not have differences when it comes to innovation overall and collaboration. However, there have differences with innovation culture (KIBS is stronger), Open innovation (KIBS is stronger), R&D

(specialised-suppliers is stronger) and IP (specialised-suppliers is stronger). Finally, supplier dominated firms seem to have a weaker signal than most firms in for most dimensions. For most sectorial categories, R&D is the stronger signal followed by innovation culture except for scale-based which does not have a difference between R&D and innovation culture. The IP signal especially strong in science-based firms arriving third behind innovation culture. The IP signal within specialised-suppliers appears to be lower than anticipated by arriving fourth with only open innovation behind.

Differences in firms' sizes are shown in Table E-12. Generally, the larger the firm, the more there seems to be stronger signals. This seems to be especially true between small firms and either medium firms or larger firms. There is a significant difference between very small and small firms with the innovation culture, collaboration and R&D signals. Very small firms are almost all significantly different than medium and large firms, except for the innovation culture that does not seem to be different with medium-size firms. Small firms versus both medium and large firms have significant differences in all factors excepting R&D with medium firms and collaboration with large firms. Finally, there are differences between medium and large firms in terms of innovation culture, but most dimensions signalled are either not significant or only marginally significant (between 0.065 and 0.1) with open innovation and R&D. Finally, the dimensions retrieved within each firms' size group are consistently ranking R&D first, then innovation culture, then collaboration followed with OI and IP, although the differences are not always significant especially between OI and IP and also between OI and collaboration for large firms.

Table E-13 presents the differences across different age groups. Both the overall innovation and innovation culture signals are not significant across all age groups. Then the collaboration, OI, R&D and IP signals generally increase with age. Between early-stage and young firms, there were only two significant differences observed in collaboration and R&D. Comparing early-stage and mature firms, all remaining dimensions exhibited significant differences. Finally, mature firms compared to young firms have only two significant differences, which were related to the open innovation and IP signals. Each age group seems to signal the same dimensions which is R&D followed by innovation culture, followed by collaboration. The OI and IP signals do not appear to be different except for mature companies.

Table E-14 shows the differences between Canadian and Finish firms. There is a consistent difference suggesting a stronger signals is captured from Canadian firms. Both Finnish and Canadian firms signal R&D first, innovation culture second and collaboration third. In contrast to Finnish firms, Canadian firms have a stronger OI than IP signal.

Table E-7. One-way ANOVA - Innovation signalling and innovation evidence

	Innovation evidence									
	None (1)		**Some (2)**		**Strong (3)**					
	$\overline{x}$	δ	$\overline{x}$	δ	$\overline{x}$	δ	K-W		M-W	
	N_1=369		N_2=867		N_3=358			1 vs 2	1 vs 3	2 vs 3
									Sig.	
All	0.073	0.093	0.098	0.097	0.081	0.077	0.000	0.000	0.000	0.003
Innov	0.034	0.070	0.047	0.084	0.049	0.081	0.000	0.000	0.000	0.084
Collab	0.027	0.057	0.030	0.068	0.022	0.038	0.003	0.002	0.163	0.023
Open	0.011	0.049	0.016	0.052	0.023	0.083	0.000	0.000	0.000	0.399
RD	0.073	0.102	0.100	0.110	0.077	0.083	0.000	0.000	0.008	0.000
IP	0.008	0.029	0.016	0.066	0.012	0.041	0.032	0.014	0.021	0.444

Innovation evidence 1:None; 2: Some; 3: Strong. Performed with the signalling variable ratio. Kruskal-Wallis and Mann-Whitney One-tailed tests.

Table E-7. One-way ANOVA - Innovation signalling and innovation evidence (continued)

		Innovation evidence											
		None				**Some**				**Strong**			
		1	2	3	4	1	2	3	4	1	2	3	4
Innov	1												
Collab	2	0.016				0.000				0.000			
Open	3	0.000	0.000			0.000	0.000			0.000	0.000		
RD	4	0.000	0.000	0.000		0.000	0.000	0.000		0.000	0.000	0.000	
IP	5	0.000	0.000	0.985	0.000	0.000	0.000	0.154	0.000	0.000	0.000	0.058	0.000

Performed with the signalling variable ratio. Kruskal-Wallis and Mann-Whitney One-tailed tests. Wilcoxon rank two-tailed tests.

Table E-8. One-way ANOVA - Innovation signalling and time to market

	Time to Market									
	Longer (1)		**Average (2)**		**Shorter (3)**					
	$\overline{x}$	δ	$\overline{x}$	δ	$\overline{x}$	δ	K-W		M-W	
	N_1=474		N_2=576		N_3=544			1 vs 2	1 vs 3	2 vs 3
								Sig.		
All	0.103	0.105	0.086	0.097	0.079	0.074	0.000	0.000	0.000	0.758
Innov	0.050	0.095	0.039	0.076	0.045	0.070	0.001	0.000	0.595	0.002
Collab	0.027	0.041	0.029	0.073	0.028	0.057	0.018	0.005	0.146	0.165
Open	0.014	0.036	0.013	0.056	0.021	0.077	0.000	0.000	0.369	0.002
RD	0.107	0.117	0.088	0.109	0.073	0.079	0.000	0.000	0.000	0.569
IP	0.019	0.075	0.013	0.049	0.008	0.032	0.003	0.008	0.001	0.240

Performed with the signalling variable ratio. Kruskal-Wallis and Mann-Whitney two-tailed tests, except one tailed test for IPRatio

Table E-8 One-way ANOVA - Innovation signalling and time to market (continued)

		Time to Market											
		Longer				**Average**				**Shorter**			
		1	2	3	4	1	2	3	4	1	2	3	4
Innov	1												
Collab	2	0.000				0.000				0.000			
Open	3	0.000	0.000			0.000	0.000			0.000	0.000		
RD	4	0.000	0.000	0.000		0.000	0.000	0.000		0.000	0.000	0.000	
IP	5	0.000	0.000	0.847	0.000	0.000	0.000	0.550	0.000	0.000	0.000	0.000	0.000

Performed with the signalling variable ratio. Wilcoxon rank two-tailed tests.

Table E-9. One-way ANOVA - Innovation signalling and Knowledge intensity

	Knowledge intensity				
	Low (1)		**High (2)**		**M-W**
	$\overline{x}$	δ	$\overline{x}$	δ	1 vs 2
	N_1=485		N_2=1109		Sig.
All	0.050	0.061	0.106	0.099	0.000
Innov	0.029	0.049	0.051	0.090	0.000
Collab	0.020	0.056	0.031	0.061	0.000
Open	0.018	0.067	0.015	0.056	0.023
RD	0.044	0.060	0.108	0.112	0.000
IP	0.006	0.033	0.016	0.061	0.000

Performed with the signalling variable ratio. Kruskal-Wallis and Mann-Whitney two-tailed tests.

Table E-9. One-way ANOVA - Innovation signalling and Knowledge intensity (continued)

		Knowledge intensity							
		Low				**High**			
		1	2	3	4	1	2	3	4
Innov	1								
Collab	2	0.000				0.000			
Open	3	0.000	0.000			0.000	0.000		
RD	4	0.000	0.000	0.000		0.000	0.000	0.000	
IP	5	0.000	0.000	0.000	0.000	0.000	0.000	0.969	0.000

Performed with the signalling variable ratio. Wilcoxon rank two-tailed tests.

Table E-10. One-way ANOVA - Innovation signalling and sector

	Sector												
	Materials and manufacturing		**Life sciences and health**		**Natural resources**		**Digital media and ICT**		**Energy and environment**		**Other**		
	$\overline{x}$	δ	$\overline{x}$	δ	$\overline{x}$	δ	$\overline{x}$	δ	$\overline{x}$	δ	$\overline{x}$	δ	K-W
	N_1=256		N_2=176		N_3=58		N_4=442		N_5=60		N_6=602		Sig.
All	0.079	0.074	0.138	0.119	0.084	0.084	0.084	0.084	0.099	0.098	0.099	0.098	0.000
Innov	0.036	0.070	0.047	0.065	0.031	0.045	0.057	0.094	0.043	0.093	0.040	0.081	0.000
Collab	0.027	0.056	0.036	0.073	0.029	0.052	0.027	0.045	0.016	0.028	0.027	0.068	0.001
Open	0.006	0.033	0.013	0.033	0.014	0.027	0.021	0.057	0.014	0.044	0.018	0.077	0.000
RD	0.088	0.103	0.157	0.144	0.090	0.106	0.085	0.084	0.115	0.134	0.068	0.087	0.000
IP	0.017	0.068	0.021	0.043	0.005	0.019	0.011	0.052	0.026	0.074	0.010	0.051	0.000

Kruskal-Wallis one-tailed tests. Performed with the signalling variable ratio.

Table E-10. One-way ANOVA - Innovation signalling and sector (continued)

	Sector														
	M-W														
	1 vs 2	1 vs 3	1 vs 4	1 vs 5	1 vs 6	2 vs 3	2 vs 4	2 vs 5	2 vs 6	3 vs 4	3 vs 5	3 vs 6	4 vs 5	4 vs 6	5 vs 6
	Sig.														
All	0.000	0.472	0.003	0.197	0.003	0.000	0.000	0.003	0.000	0.059	0.295	0.064	0.295	0.000	0.014
Innov	0.003	0.412	0.000	0.419	0.442	0.038	0.018	0.050	0.001	0.001	0.494	0.398	0.001	0.000	0.394
Collab	0.005	0.437	0.065	0.064	0.325	0.078	0.056	0.001	0.001	0.244	0.081	0.296	0.007	0.008	0.093
Open	0.000	0.000	0.000	0.000	0.000	0.137	0.028	0.484	0.010	0.448	0.188	0.002	0.108	0.000	0.071
RD	0.000	0.499	0.210	0.128	0.000	0.000	0.000	0.006	0.000	0.280	0.199	0.021	0.214	0.000	0.002
IP	0.000	0.111	0.417	0.137	0.082	0.001	0.000	0.090	0.000	0.118	0.035	0.569	0.096	0.072	0.021

Sector 1: Materials and manufacturing; 2: Life sciences and health; 3: Natural resources; 4: Digital media and ICT; 5 Energy and environment; 6: Other. Mann-Whitney one-tailed tests.

Table E-10. One-way ANOVA - Innovation signalling and sector (continued)

		Sector											
		Materials and manufacturing				**Life sciences and health**				**Natural resources**			
		1	2	3	4	1	2	3	4	1	2	3	4
Innov	1												
Collab	2	0.003				0.001				0.151			
Open	3	0.000	0.000			0.000	0.000			0.001	0.093		
RD	4	0.000	0.000	0.000		0.000	0.000	0.000		0.000	0.000	0.000	
IP	5	0.000	0.000	0.001	0.000	0.000	0.002	0.010	0.000	0.000	0.000	0.048	0.000

Table E-10. One-way ANOVA - Innovation signalling and sector (continued)

		Sector											
		Digital media and ICT				**Energy and environment**				**Other**			
		1	2	3	4	1	2	3	4	1	2	3	4
Innov	1												
Collab	2	0.000				0.002				0.000			
Open	3	0.000	0.000			0.001	0.330			0.000	0.000		
RD	4	0.000	0.000	0.000		0.000	0.000	0.000		0.000	0.000	0.000	
IP	5	0.000	0.000	0.000	0.000	0.013	0.512	0.633	0.000	0.000	0.000	0.022	0.000

Wilcoxon rank two-tailed tests

Table E-11. One-way ANOVA - Innovation signalling and sectorial category

	Sectorial category										
	Supplier-dominated		**Specialised-suppliers**		**KIBS**		**Science-based**		**Scale-based**		
	$\overline{x}$	δ	$\overline{x}$	δ	$\overline{x}$	δ	$\overline{x}$	δ	$\overline{x}$	δ	K-W
	N_1=396		N_2=614		N_3=541		N_4=33		N_5=10		Sig.
All	0.046	0.063	0.105	0.104	0.097	0.080	0.180	0.170	0.068	0.048	0.000
Innov	0.024	0.043	0.045	0.084	0.055	0.075	0.098	0.235	0.066	0.126	0.000
Collab	0.021	0.058	0.028	0.051	0.033	0.070	0.029	0.046	0.009	0.014	0.000
Open	0.017	0.066	0.007	0.028	0.026	0.079	0.017	0.025	0.017	0.024	0.000
RD	0.041	0.060	0.113	0.121	0.090	0.087	0.193	0.162	0.055	0.044	0.000
IP	0.005	0.034	0.019	0.058	0.010	0.050	0.047	0.145	0.008	0.012	0.000

Kruskal-Wallis one-tailed tests. Performed with the signalling variable ratio.

Table E-11. One-way ANOVA - Innovation signalling and sectorial category (continued)

	Sectorial category									
	M-W									
	1 vs 2	1 vs 3	1 vs 4	1 vs 5	2 vs 3	2 vs 4	2 vs 5	3 vs 4	3 vs 5	4 vs 5
	Sig.									
All	0.000	0.000	0.000	0.014	0.446	0.000	0.158	0.000	0.096	0.001
Innov	0.000	0.000	0.000	0.017	0.000	0.061	0.336	0.494	0.262	0.202
Collab	0.000	0.000	0.001	0.449	0.111	0.110	0.090	0.232	0.053	0.023
Open	0.223	0.000	0.003	0.011	0.000	0.000	0.003	0.307	0.228	0.386
RD	0.000	0.000	0.000	0.056	0.015	0.000	0.064	0.000	0.101	0.001
IP	0.000	0.002	0.000	0.006	0.000	0.112	0.311	0.003	0.065	0.461

Sectorial category 1: Supplier-dominated; 2: Specialised-suppliers; 3: KIBS; 4: Science-based; 5 Scale-based. Mann-Whitney one-tailed tests.

Table E-11. One-way ANOVA - Innovation signalling and sectorial category (continued)

		Sectorial category																		
		Supplier-dominated				**Specialised supplier**				**KIBS**				**Science-based**				**Scale-based**			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Innov	1																				
Collab	2	0.002				0.000				0.000				0.021				0.022			
Open	3	0.000	0.000			0.000	0.000			0.000	0.000			0.002	0.304			0.386	0.249		
RD	4	0.000	0.000	0.000		0.000	0.000	0.000		0.000	0.000	0.000		0.001	0.000	0.000		0.285	0.008	0.051	
IP	5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.062	0.444	0.000	0.037	0.753	0.237	0.011

Wilcoxon rank two-tailed tests.

Table E-12. One-way ANOVA - Innovation signalling and size

	Size														
	Very small		**Small**		**Medium**		**Large**								
	$\overline{x}$	δ	$\overline{x}$	δ	$\overline{x}$	δ	$\overline{x}$	δ	K-W	M-W					
	N_1=960		N_2=525		N_3=58		N_4=51			1 vs 2	1 vs 3	1 vs 4	2 vs 3	2 vs 4	3 vs 4
All	0.086	0.094	0.086	0.081	0.103	0.091	0.150	0.154	0.000	0.459	0.022	0.000	0.053	0.000	0.124
Innov	0.047	0.084	0.035	0.050	0.037	0.042	0.092	0.197	0.004	0.022	0.310	0.005	0.077	0.000	0.037
Collab	0.027	0.067	0.030	0.048	0.027	0.027	0.029	0.043	0.000	0.000	0.000	0.001	0.026	0.106	0.324
Open	0.016	0.064	0.014	0.051	0.026	0.059	0.029	0.052	0.000	0.397	0.000	0.000	0.000	0.000	0.083
RD	0.083	0.102	0.090	0.094	0.112	0.128	0.150	0.156	0.000	0.017	0.007	0.000	0.059	0.001	0.069
IP	0.012	0.046	0.013	0.048	0.021	0.109	0.034	0.118	0.006	0.185	0.032	0.001	0.088	0.005	0.198

Size 1: Very small; 2: Small; 3: Medium; 4: Large. Kruskal-Wallis two-tailed tests for AllRatio and one-tailed test for the other variables. Performed with the signalling variables ratio.

Table E-12. One-way ANOVA - Innovation signalling and size (continued)

		Size															
		Very small				**Small**				**Medium**				**Large**			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Innov	1																
Collab	2	0.000				0.015				0.191				0.001			
Open	3	0.000	0.000			0.000	0.000			0.025	0.041			0.002	0.788		
RD	4	0.000	0.000	0.000		0.000	0.000	0.000		0.000	0.000	0.000		0.000	0.000	0.000	
IP	5	0.000	0.000	0.127	0.000	0.000	0.000	0.899	0.000	0.000	0.000	0.126	0.000	0.000	0.005	0.023	0.000

Wilcoxon rank two-tailed tests.

Table E-13. One-way ANOVA - Innovation signalling and age

	Age									
	Early stage (1)		**Young (2)**		**Mature (3)**					
	$\overline{x}$	δ	$\overline{x}$	δ	$\overline{x}$	δ	K-W	M-W		
	N_1=530		N_2=615		N_3=449			1 vs 2	1 vs 3	2 vs 3
								Sig.		
All	0.084	0.087	0.089	0.091	0.094	0.101	0.106	0.163	0.037	0.421
Innov	0.046	0.076	0.043	0.074	0.045	0.092	0.961	0.775	0.858	0.960
Collab	0.027	0.070	0.028	0.054	0.028	0.053	0.000	0.000	0.000	0.500
Open	0.015	0.055	0.011	0.038	0.024	0.083	0.000	0.328	0.000	0.000
RD	0.080	0.097	0.091	0.103	0.096	0.110	0.011	0.022	0.005	0.443
IP	0.012	0.051	0.012	0.047	0.016	0.065	0.003	0.749	0.002	0.004

Kruskal-Wallis and Mann-Whitney two-tailed tests. Performed with the signalling variables ratio.

Table E-13. One-way ANOVA - Innovation signalling and age (continued)

	Age											
		Early stage				**Young**				**Mature**		
	1	2	3	4	1	2	3	4	1	2	3	4
Innov	1											
Collab	2	0.000				0.000				0.000			
Open	3	0.000	0.000			0.000	0.000			0.000	0.000		
RD	4	0.000	0.000	0.000		0.000	0.000	0.000		0.000	0.000	0.000	
IP	5	0.000	0.000	0.477	0.000	0.000	0.000	0.597	0.000	0.000	0.000	0.018	0.000

Wilcoxon rank two-tailed tests. Performed with the signalling variable ratio.

Table E-14. One-way ANOVA - Innovation signalling and country

	Country				
	Finland		**Canada**		M-W
	$\overline{x}$	δ	$\overline{x}$	δ	1 vs 2
	N_1=828		N_2=766		Sig.
All	0.071	0.077	0.108	0.104	0.000
Innov	0.035	0.064	0.055	0.094	0.000
Collab	0.027	0.061	0.029	0.058	0.010
Open	0.006	0.036	0.027	0.076	0.000
RD	0.070	0.081	0.109	0.119	0.000
IP	0.010	0.045	0.017	0.062	0.000

Country 1: Finland; 2: Canada. Mann-Whitney two-tailed tests. Performed with the signalling variable ratio.

Table E-14. One-way ANOVA - Innovation signalling and country (continued)

		Country							
		Finland				**Canada**			
		1	2	3	4	1	2	3	4
Innov	1								
Collab	2	0.000				0.000			
Open	3	0.000	0.000			0.000	0.000		
RD	4	0.000	0.000	0.000		0.000	0.000	0.000	
IP	5	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000

Wilcoxon rank two-tailed tests.

How to read regression results

All regression results are presented in the following pages. A summary with both the ANOVA and regression results is reported from Table E-36 to Table E-41. Following next is an example of the interpretation of the results.

Table E-15 introduces the first model with the sector, the size and the knowledge intensity level (Model a) along with the interaction between the sector and the knowledge intensity (Model b) for *countAll*, *countInnov* and *countCollab*. Each categorical variable is compared with an omitted group[80]. For example, in the regression 1a in the negative binomial section, some evidence of innovation and strong evidence of innovation are compared with no evidence of innovation. As some innovation has a coefficient of 0.158 and an incidence rate ratio (IRR) of 1.171., it implies it has 17.1% higher incidence rate than the firms with no evidence of innovation. Moreover, as the regression results only show the variations in incidence rate compared with the omitted group, predicted margin effects are performed with Wald tests which allow to assess the differences between the other categories. Table E-18 to Table E-23 shows the signal predicted in terms of occurrences for a target group of firms that remain constant from one variable to the other[81]. The predicted value, the differences and their correspond significance are reported. To continue our example with the innovation evidence, Table E-18 reports for the target group significant predicted signals of 62.17 for firms with no evidence, 72.82 for firms with some evidence and 62.84 for firms with strong evidence. Then, Delta 1 shows the differences with the first group, Delta 2 the differences with the second group and so on. The predicted values are the ones associated with the target group selected, but the *p-values* associated with the differences are based on the regression results. In this case, there are significant differences between some evidence and no evidence (Delta 1 =10.65) and between some evidence and strong evidence (Delta 2 =-9.98) which suggest that firms with some evidence yield more signal than all the other categories.

In 2a, both some evidence (IRR=1.174) and strong evidence seems to have a relative impact (IRR=1.163) on innovation culture and there is no significant difference between some and strong

[80] The omitted groups are *dNoInnovEvidence*, *dAverageTMkt, dEnerEnvironNat, dLowKI, dSmall, dEarlyStage, dFinland* and *d2014.*

[81] The target group for *countAll*, *countInnov*, *countCollab*, *countOpen*, *countRD* are all the same with *dSomeInnovEvidence, dAverageTMkt, dDigMediaICT, dHighKI, dVerySmall, dYoung, dCanada* and *d2014*. For *countIP*, the group is *dSomeInnovEvidence*, *dLongTMkt, dLifeSciHealth, dHighKI, dSmall, dMature, dCanada* and *d2014*as it optimizes the number of nonzero observations.

evidence which can be seen in Table E-20 and Table E-21. For collaboration, strong innovation evidence seems to have a relative negative impact on collaboration (IRR=0.807). From Table E-22, there is also a significant difference between some and strong innovation evidence (Delta 2 = -1.95).

The time to market does not seem to have any impact on *countAll*, *countInnov* and *countCollab*. There are also significant relative impacts from different sectors showing Life science and health (IRR_{all}=2.526, IRR_{innov}=2.534, IRR_{collab} =1.953), Digital media and ICT (IRR_{all}=3.390, IRR_{innov}=4.206, IRR_{collab} =4.362) and Other (IRR_{all}=2.3342, IRR_{innov}=2.869, IRR_{collab} =2.877) from the omitted group (Natural resources, energy & environment). For innovation overall, there are also significant differences between sectors with the exception of Life sciences with Other and Materials and manufacturing with Natural resources, Energy and environment. In terms of predicted values of the target group, there is Digital media and ICT (72.82) followed with Life science and health (54.20), followed with other (50.22), materials & manufacturing (24.71) and finally, Natural resources, energy & environment (21.39). Almost the same results are found in Table E-20 for innovation culture, but with the difference that Life science and health and other exchanged positions in terms of ranking. For collaboration, from Table E-22 the same ranking as previously reported with open innovation is found and all the sectors have significant differences among each other.

As expected, there seems to be higher impact on the signal obtained from firms with higher knowledge intensity level for all *countAll*, *countInnov* and *countCollab.* Next, medium firms seem to be consistently reporting less innovation-related signalling across all factors compared with small firms. The differences on the overall innovation and innovation culture signal are the same with only differences with medium firms from all other sizes. For collaboration, there are also differences between small (10.89) and very small firms (8.37), but medium and large firms do not seem to be different. Furthermore, the age does not impact the signal for all *countAll*, *countInnov* and *countCollab.* There are consistent differences when it comes to Canadian firms versus Finish firms with Canadian firms having consistent stronger signal.

Assessing the periods, it seems that 2014 was the period that yields the strongest signal. Finally, as expected, the sentence ratio contributes directly on the signal obtained, the closer the ratio is to the maximum, the more there is an aggregated signal.

In all models, the sentence ratio decreases the log odds of being an excessive zero by at least 1.075 for every incremental unit. Large firms versus small firms decrease the log odds of being an excessive zero by 15.85 for *countAll.* With *countInnov* and *countCollab,* knowledge intensity seems to decrease the odd ratio. Furthermore, both mature and young firms versus early-stage decrease the log odds of being an excessive zero on *countCollab.*

Table E-16 presents the zero-inflated negative binomial regression results of *countOpen*, *countRD* and *CountIP*. It is followed with their marginal effects from Table E-24 to Table E-29. The second model with the sectorial category is shown in Table E-17 followed with the marginal effects from Table E-30 to Table E-35.

Table E-15. Zero-inflated negative binomial with countAll, countInnov and countCollab – Model a and b

	countAll		countInnov		countCollab	
	1a	**1b**	**2a**	**2b**	**3a**	**3b**
Negative binomial						
Innovation evidence (*dNoInnovEvidence* omitted)						
	[1.171]	[1.175]	[1.174]	[1.173]	[1.056]	[1.057]
dSomeInnovEvidence	0.158***	0.161***	0.161*	0.159*	0.055	0.056
	(0.063)	(0.064)	(0.099)	(0.098)	(0.108)	(0.108)
	[1.011]	[1.000]	[1.163]	[1.159]	[0.811]	[0.807]
dStrongInnovEvidence	0.011	0.000	0.151*	0.148*	-0.210**	-0.215**
	(0.072)	(0.072)	(0.105)	(0.107)	(0.127)	(0.129)
Time market (*dAverageTMkt* omitted)						
	[1.011]	[1.014]	[1.025]	[1.022]	[0.863]	[0.860]
dLongTMkt	0.011	0.014	0.024	0.022	-0.147	-0.150
	(0.055)	(0.055)	(0.081)	(0.081)	(0.103)	(0.103)
	[0.945]	[0.945]	[1.053]	[1.046]	[0.970]	[0.962]
dShortTMkt	-0.057	-0.057	0.052	0.045	-0.031	-0.039
	(0.059)	(0.059)	(0.091)	(0.09)	(0.111)	(0.108)
Knowledge level (*dLowKI* omitted)						
	[1.743]		[1.308]		[1.255]	
dHighKI	0.556***		0.269***		0.227**	
	(0.061)		(0.093)		(0.129)	

Table E-15. Zero-inflated negative binomial with countAll, countInnov and countCollab – Model a and b (continued)

	countOpen		countRD		countIP	
	1a	**1b**	**2a**	**2b**	**3a**	**3b**
Sector (*dEnerEnvironNat* omitted)						
	[1.161]		[1.207]		[1.396]	
dMaterialMnfg	0.150		0.188		0.334*	
	(0.123)		(0.227)		(0.23)	
	[2.526]		[2.534]		[1.953]	
dLifeSciHealth	0.926***		0.930***		0.669***	
	(0.121)		(0.182)		(0.228)	
	[3.390]		[4.206]		[4.362]	
dDigMediaICT	1.221***		1.436***		1.473***	
	(0.117)		(0.206)		(0.203)	
	[2.342]		[2.869]		[2.877]	
dOther	0.851***		1.054***		1.057***	
	(0.118)		(0.242)		(0.207)	
Sector x Knowledge intensity (*dEnerEnvironNat x dLowKI* omitted)						
		[1.547]		[1.385]		[2.080]
dMaterialMnfg* x *dLowKI		0.436***		0.326		0.732*
		(0.182)		(0.419)		(0.501)
		[2.069]		[1.517]		[1.609]
dMaterialMnfg* x *dHighKI		0.727***		0.417		0.476*
		(0.136)		(0.378)		(0.328)

Table E-15. Zero-inflated negative binomial with countAll, countInnov and countCollab – Model a and b (continued)

	countAll		**countInnov**		**countCollab**	
	1a	**1b**	**2a**	**2b**	**3a**	**3b**
		[2.173]		[2.655]		[2.050]
dLifeSciHealth* x *dLowKI		0.776***		0.976**		0.718*
		(0.246)		(0.469)		(0.521)
		[4.986]		[3.230]		[2.362]
dLifeSciHealth* x *dHighKI		1.607***		1.173***		0.859***
		(0.125)		(0.31)		(0.318)
		[4.137]		[4.561]		[3.760]
dDigMediaICT* x *dLowKI		1.420***		1.517***		1.324***
		(0.143)		(0.365)		(0.336)
		[6.113]		[5.331]		[5.473]
dDigMediaICT* x *dHighKI		1.810***		1.674***		1.700***
		(0.129)		(0.357)		(0.308)
		[1.925]		[1.277]		[1.196]
dEnerEnvironNat* x *dHighKI		0.655***		0.245		0.179
		(0.158)		(0.283)		(0.344)
		[2.347]		[2.604]		[2.705]
dOther* x *dLowKI		0.853***		0.957***		0.995***
		(0.143)		(0.387)		(0.325)

Table E-15. Zero-inflated negative binomial with countAll, countInnov and countCollab – Model a and b (continued)

		countAll		countInnov		countCollab	
		1a	**1b**	**2a**	**2b**	**3a**	**3b**
			[4.551]		[3.892]		[3.570]
	dOther x dHighKI		1.515***		1.359***		1.273***
			(0.133)		(0.4)		(0.327)
		[2.351]	[2.358]	[2.158]	[2.165]	[1.971]	[1.971]
Sentence ratio	***SentenceRatio***	0.855***	0.858***	0.769***	0.772***	0.678***	0.679***
		(0.019)	(0.019)	(0.038)	(0.039)	(0.044)	(0.045)
Size (*dSmall* omitted)							
		[0.972]	[0.976]	[0.986]	[0.990]	[0.780]	[0.789]
	dVerySmall	-0.028	-0.024	-0.014	-0.010	-0.248***	-0.237**
		(0.055)	(0.055)	(0.1)	(0.101)	(0.104)	(0.105)
		[0.682]	[0.691]	[0.607]	[0.605]	[0.582]	[0.591]
	dMedium	-0.382***	-0.370**	-0.500***	-0.502***	-0.541***	-0.526***
		(0.14)	(0.144)	(0.18)	(0.181)	(0.16)	(0.161)
		[0.932]	[0.930]	[1.184]	[1.199]	[0.699]	[0.700]
	dLarge	-0.070	-0.072	0.169	0.181	-0.358**	-0.356**
		(0.14)	(0.142)	(0.211)	(0.208)	(0.195)	(0.197)
Age (*dEarlyStage* omitted)							
		[1.094]	[1.091]	[0.910]	[0.910]	[0.965]	[0.947]
	dYoung	0.090	0.087	-0.094	-0.094	-0.036	-0.055
		(0.062)	(0.062)	(0.097)	(0.097)	(0.128)	(0.134)

Table E-15. Zero-inflated negative binomial with countAll, countInnov and countCollab – Model a and b (continued)

	countAll		countInnov		countCollab	
	1a	**1b**	**2a**	**2b**	**3a**	**3b**
	[1.126]	[1.118]	[0.896]	[0.890]	[0.874]	[0.876]
dMature	0.118	0.112	-0.110	-0.116	-0.135	-0.132
	(0.075)	(0.075)	(0.111)	(0.107)	(0.158)	(0.16)
Country (*dFinland* omitted)						
	[1.886]	[1.823]	[1.825]	[1.837]	[2.161]	[2.169]
dCanada	0.635***	0.600***	0.601***	0.608***	0.771***	0.774***
	(0.087)	(0.091)	(0.141)	(0.136)	(0.218)	(0.226)
Period (*d2014* omitted)						
	[0.682]	[0.694]	[0.681]	[0.676]	[0.556]	[0.549]
d2010_2011	-0.382***	-0.365***	-0.385**	-0.391**	-0.586**	-0.600**
	(0.107)	(0.109)	(0.167)	(0.165)	(0.244)	(0.25)
	[0.684]	[0.712]	[0.812]	[0.806]	[0.574]	[0.569]
d2012_2013	-0.380***	-0.340***	-0.208	-0.215	-0.555**	-0.564**
	(0.097)	(0.099)	(0.148)	(0.134)	(0.225)	(0.231)
	[0.776]	[0.798]	[0.742]	[0.745]	[0.828]	[0.829]
d2015_2016	-0.253***	-0.226***	-0.299***	-0.294***	-0.189	-0.187
	(0.083)	(0.084)	(0.116)	(0.113)	(0.159)	(0.158)
	[0.004]	[0.004]	[0.002]	[0.002]	[0.004]	[0.004]
Constant	-5.489***	-5.579***	-6.138***	-6.150***	-5.460***	-5.428***
	(0.223)	(0.22)	(0.486)	(0.601)	(0.488)	(0.52)

Table E-15. Zero-inflated negative binomial with countAll, countInnov and countCollab – Model a and b (continued)

	countAll		countInnov		countCollab	
	1a	**1b**	**2a**	**2b**	**3a**	**3b**
Inflate						
Knowledge level (*dLowKI* omitted)						
	[0.422]	[0.443]	[0.308]	[0.314]	[0.389]	[0.375]
dHighKI	-0.864	-0.814	-1.179**	-1.157**	-0.945**	-0.981**
	(0.815)	(0.853)	(0.498)	(0.518)	(0.39)	(0.417)
Sector (*dEnerEnvironNat* omitted)						
	[2.033]	[1.969]	[2.153]	[2.189]	[1.495]	[1.669]
dMaterialMnfg	0.709	0.678	0.767	0.783	0.402	0.512
	(1.856)	(1.807)	(3.246)	(3.598)	(0.897)	(0.854)
	[0.360]	[0.321]	[0.783]	[0.827]	[0.116]	[0.143]
dLifeSciHealth	-1.021	-1.137	-0.244	-0.190	-2.154	-1.946
	(1.709)	(1.672)	(2.134)	(2.352)	(2.066)	(2.399)
	[0.199]	[0.193]	[0.289]	[0.300]	[0.641]	[0.656]
dDigMediaICT	-1.617	-1.645	-1.242	-1.203	-0.445	-0.422
	(2.01)	(1.966)	(2.864)	(3.178)	(0.679)	(0.666)
	[0.537]	[0.478]	[0.866]	[0.850]	[0.596]	[0.598]
dOther	-0.623	-0.739	-0.144	-0.163	-0.518	-0.513
	(2.065)	(2.04)	(3.436)	(3.8)	(0.729)	(0.716)

Table E-15. Zero-inflated negative binomial with countAll, countInnov and countCollab – Model a and b (continued)

		countAll		countInnov		countCollab	
		1a	**1b**	**2a**	**2b**	**3a**	**3b**
		[0.285]	[0.282]	[0.371]	[0.372]	[0.334]	[0.341]
Sentence ratio	***SentenceRatio***	-1.254***	-1.265***	-0.992***	-0.988***	-1.098***	-1.075***
		(0.200)	(0.209)	(0.262)	(0.27)	(0.215)	(0.233)
Size (*dSmall* omitted)							
		[0.731]	[0.727]	[0.472]	[0.477]	[1.054]	[1.082]
	dVerySmall	-0.313	-0.318	-0.751	-0.740	0.053	0.079
		(0.985)	(0.977)	(0.82)	(0.889)	(0.398)	(0.409)
		[3.009]	[2.899]	[0.165]	[0.171]	[0.156]	[0.113]
	dMedium	1.102	1.064	-1.802	-1.768	-1.861	-2.181
		(1.16)	(1.17)	(3.917)	(4.133)	(3.091)	(3.7)
		[0.000]	[0.000]	[0.711]	[0.782]	[0.893]	[0.852]
	dLarge	-15.484***	-17.530***	-0.340	-0.246	-0.114	-0.160
		(0.777)	(0.804)	(4.889)	(4.616)	(1.787)	(1.917)
Age (*dEarlyStage* omitted)							
		[1.531]	[1.675]	[0.721]	[0.720]	[0.514]	[0.486]
	dYoung	0.426	0.516	-0.327	-0.329	-0.665*	-0.722*
		(1.001)	(1.064)	(0.573)	(0.612)	(0.382)	(0.413)
		[0.919]	[0.998]	[0.513]	[0.492]	[0.330]	[0.339]
	dMature	-0.085	-0.002	-0.667	-0.710	-1.108*	-1.081*
		(1.502)	(1.604)	(0.795)	(0.84)	(0.645)	(0.623)

Table E-15. Zero-inflated negative binomial with countAll, countInnov and countCollab – Model a and b (continued)

	countAll		countInnov		countCollab	
	1a	**1b**	**2a**	**2b**	**3a**	**3b**
Country (*dFinland* omitted)						
	[0.550]	[0.522]	[0.535]	[0.528]	[4.352]	[3.921]
dCanada	-0.599	-0.650	-0.625	-0.639	1.471	1.366
	(1.017)	(1.074)	(1.341)	(1.36)	(1.033)	(1.2)
Period (*d2014* omitted)						
	[2.912]	[3.135]	[0.312]	[0.374]	[0.352]	[0.370]
d2010_2011	1.069	1.143	-1.163	-0.984	-1.045	-0.993
	(1.406)	(1.431)	(2.251)	(2.017)	(1.104)	(1.234)
	[2.750]	[2.905]	[2.033]	[2.008]	[0.344]	[0.366]
d2012_2013	1.012	1.066	0.709	0.697	-1.068	-1.004
	(1.470)	(1.510)	(1.133)	(1.093)	(0.975)	(1.116)
	[6.021]	[6.165]	[1.919]	[1.973]	[0.834]	[0.869]
d2015_2016	1.795	1.819	0.652	0.679	-0.182	-0.140
	(1.469)	(1.515)	(0.905)	(0.91)	(0.742)	(0.790)
	[726.9]	[754.9]	[1,839.5]	[1,744.2]	[6,314.9]	[5,480.9]
Constant	6.589***	6.627***	7.517	7.464	8.751***	8.609***
	(1.396)	(1.37)	(4.746)	(5.087)	(1.658)	(1.698)
Nb observations	1594	1594	1594	1594	1594	1594
Nonzero obs	1448	1448	1150	1150	931	931
Zero obs	146	146	444	444	663	663

Table E-15. Zero-inflated negative binomial with countAll, countInnov and countCollab – Model a and b (continued)

	countAll		countInnov		countCollab	
	1a	**1b**	**2a**	**2b**	**3a**	**3b**
Nb groups	1569	1569	1569	1569	1569	1569
McFadden R^2	0.145	0.145	0.137	0.137	0.126	0.126
McFadden adjusted R^2	0.140	0.140	0.129	0.129	0.116	0.116
χ^2	2617.83***	2576.14***	687.23***	701.99***	371.33***	390.20***
Loglikelihood (full model)	-6438.90	-6432.91	-4271.94	-4270.30	-3402.13	-3400.09
Loglikelihood (constant only)	-7399.84	-7399.84	-4719.93	-4719.93	-3652.15	-3652.15
Deviance (ZINB vs nbreg)	36.505***	35.234***	47.483***	44.600***	84.522***	85.445***

Note: Incidence rate ratio in square brackets. Standard errors in parentheses. ***, **, * show significance at the 1%, 5% and 10% level respectively. All alphas are positive and significant. Mean VIF is 2.13.

Table E-16. Zero-inflated negative binomial with countOpen, countRD and countIP – Model a and b

	countOpen		countRD		countIP	
	4a	**4b**	**5a**	**5b**	**6a**	**6b**
Negative binomial						
Innovation evidence (*dNoInnovEvidence* omitted)						
	[1.097]	[1.144]	[1.193]	[1.206]	[0.967]	[0.951]
dSomeInnovEvidence	0.093	0.134	0.176***	0.187***	-0.034	-0.051
	(0.198)	(0.189)	(0.074)	(0.078)	(0.188)	(0.178)
	[0.998]	[1.040]	[0.964]	[0.944]	[1.035]	[1.010]
dStrongInnovEvidence	-0.002	0.039	-0.037	-0.058	0.034	0.010
	(0.25)	(0.239)	(0.084)	(0.087)	(0.225)	(0.214)
Time market (*dAverageTMkt* omitted)						
	[0.837]	[0.897]	[1.000]	[1.006]	[1.293]	[1.254]
dLongTMkt	-0.178	-0.109	0.000	0.006	0.257*	0.226*
	(0.166)	(0.16)	(0.066)	(0.065)	(0.171)	(0.175)
	[1.280]	[1.313]	[0.904]	[0.906]	[0.826]	[0.844]
dShortTMkt	0.247	0.272	-0.101	-0.099	-0.192	-0.170
	(0.185)	(0.182)	(0.071)	(0.071)	(0.18)	(0.181)
Knowledge level (*dLowKI* omitted)						
	[0.714]		[1.771]		[1.945]	
dHighKI	-0.336**		0.572***		0.665***	
	(0.202)		(0.069)		(0.227)	

Table E-16. Zero-inflated negative binomial with countOpen, countRD and countIP – Model a and b (continued)

	countOpen		countRD		countIP	
	4a	**4b**	**5a**	**5b**	**6a**	**6b**
Sector (*dEnerEnvironNat* omitted)						
	[3.356]		[0.965]		[1.419]	
dMaterialMnfg	1.211***		-0.035		0.350	
	(0.402)		(0.161)		(0.31)	
	[2.537]		[2.640]		[2.794]	
dLifeSciHealth	0.931***		0.971***		1.027***	
	(0.251)		(0.154)		(0.295)	
	[6.813]		[2.621]		[4.269]	
dDigMediaICT	1.919***		0.964***		1.451***	
	(0.229)		(0.15)		(0.459)	
	[7.137]		[1.951]		[2.720]	
dOther	1.965***		0.668***		1.001***	
	(0.268)		(0.147)		(0.309)	
Sector x Knowledge intensity (*dEnerEnvironNat x dLowKI* omitted)						
		[13.295]		[1.594]		[1.247]
dMaterialMnfg* x *dLowKI		2.587***		0.467**		0.221
		(0.61)		(0.261)		(0.77)
		[2.090]		[1.828]		[2.438]
dMaterialMnfg* x *dHighKI		0.737*		0.603***		0.891**
		(0.494)		(0.225)		(0.536)

Table E-16. Zero-inflated negative binomial with countOpen, countRD and countIP – Model a and b (continued)

	countOpen		countRD		countIP	
	4a	**4b**	**5a**	**5b**	**6a**	**6b**
		[3.881]		[1.778]		[1.257]
dLifeSciHealth* x *dLowKI		1.356***		0.576***		0.229
		(0.523)		(0.238)		(0.962)
		[2.412]		[5.856]		[5.444]
dLifeSciHealth* x *dHighKI		0.881**		1.767***		1.694***
		(0.418)		(0.192)		(0.557)
		[6.620]		[3.396]		[5.488]
dDigMediaICT* x *dLowKI		1.890***		1.223***		1.703***
		(0.458)		(0.215)		(0.583)
		[6.663]		[5.063]		[6.396]
dDigMediaICT* x *dHighKI		1.897***		1.622***		1.856***
		(0.383)		(0.204)		(0.584)
		[1.014]		[2.058]		[1.725]
dEnerEnvironNat* x *dHighKI		0.014		0.722***		0.545
		(0.394)		(0.233)		(0.566)
		[10.095]		[2.065]		[2.293]
dOther* x *dLowKI		2.312***		0.725***		0.830*
		(0.426)		(0.211)		(0.553)

Table E-16. Zero-inflated negative binomial with countOpen, countRD and countIP – Model a and b (continued)

		countOpen		countRD		countIP	
		4a	**4b**	**5a**	**5b**	**6a**	**6b**
			[6.346]		[4.063]		[4.792]
	dOther* x *dHighKI		1.848***		1.402***		1.567***
			(0.432)		(0.216)		(0.533)
		[2.212]	[2.182]	[2.368]	[2.331]	[1.842]	[1.849]
Sentence ratio	***SentenceRatio***	0.794***	0.780***	0.862***	0.846***	0.611***	0.615***
		(0.088)	(0.09)	(0.021)	(0.031)	(0.122)	(0.114)
Size (*dSmall* omitted)							
		[0.600]	[0.638]	[0.936]	[0.967]	[1.651]	[1.620]
	dVerySmall	-0.510***	-0.450***	-0.066	-0.033	0.501***	0.482***
		(0.18)	(0.184)	(0.065)	(0.077)	(0.183)	(0.186)
		[0.328]	[0.374]	[0.573]	[0.601]	[2.107]	[2.159]
	dMedium	-1.114***	-0.984***	-0.557***	-0.510***	0.745	0.770*
		(0.34)	(0.345)	(0.19)	(0.205)	(0.62)	(0.564)
		[0.920]	[0.928]	[0.939]	[1.010]	[0.650]	[0.673]
	dLarge	-0.083	-0.074	-0.063	0.009	-0.431	-0.395
		(0.362)	(0.363)	(0.182)	(0.204)	(0.389)	(0.343)
Age (*dEarlyStage* omitted)							
		[0.874]	[0.826]	[1.138]	[1.144]	[1.176]	[1.127]
	dYoung	-0.135	-0.191	0.129*	0.134*	0.162	0.120
		(0.178)	(0.177)	(0.073)	(0.071)	(0.236)	(0.246)

Table E-16. Zero-inflated negative binomial with countOpen, countRD and countIP – Model a and b (continued)

		countOpen		countRD		countIP	
		4a	**4b**	**5a**	**5b**	**6a**	**6b**
		[0.995]	[1.020]	[1.199]	[1.184]	[1.898]	[1.794]
	dMature	-0.005	0.019	0.182**	0.169*	0.641***	0.584**
		(0.227)	(0.225)	(0.089)	(0.101)	(0.247)	(0.243)
Country (*dFinland* omitted)							
		[1.410]	[1.469]	[1.607]	[1.542]	[2.662]	[2.474]
	dCanada	0.344	0.384	0.474***	0.433***	0.979***	0.906***
		(0.311)	(0.324)	(0.104)	(0.115)	(0.252)	(0.244)
Period (*d2014* omitted)							
		[0.893]	[0.884]	[0.719]	[0.741]	[0.593]	[0.608]
	d2010_2011	-0.113	-0.123	-0.330**	-0.300**	-0.523*	-0.497*
		(0.339)	(0.344)	(0.142)	(0.14)	(0.308)	(0.3)
		[1.228]	[1.225]	[0.661]	[0.696]	[0.463]	[0.511]
	d2012_2013	0.205	0.203	-0.414***	-0.362***	-0.770***	-0.671**
		(0.279)	(0.298)	(0.115)	(0.136)	(0.286)	(0.274)
		[1.612]	[1.622]	[0.768]	[0.813]	[0.625]	[0.649]
	d2015_2016	0.477*	0.484*	-0.264**	-0.207	-0.470	-0.432*
		(0.263)	(0.271)	(0.11)	(0.131)	(0.289)	(0.232)
		[0.001]	[0.000]	[0.002]	[0.002]	[0.000]	[0.000]
Constant		-7.567***	-7.843***	-6.120***	-6.126***	-7.966***	-7.822***
		(0.994)	(1.068)	(0.263)	(0.29)	(1.181)	(1.245)

Table E-16. Zero-inflated negative binomial with countOpen, countRD and countIP – Model a and b (continued)

		countOpen		countRD		countIP	
		4a	**4b**	**5a**	**5b**	**6a**	**6b**
Inflate							
Knowledge intensity		[0.512]	[0.502]	[0.276]	[0.440]	[1.175]	[0.939]
	dHighKI	-0.669**	-0.690**	-1.286**	-0.821	0.162	-0.062
		(0.335)	(0.33)	(0.649)	(1.019)	(0.82)	(0.831)
Sector (*dEnerEnvironNat* omitted)							
		[9.780]	[9.239]	[0.191]	[0.622]	[1.445]	[1.393]
	dMaterialMnfg	2.280***	2.223***	-1.653*	-0.476	0.368	0.332
		(0.72)	(0.761)	(0.849)	(1.332)	(1.511)	(1.636)
		[1.238]	[1.450]	[0.017]	[0.075]	[0.819]	[0.874]
	dLifeSciHealth	0.213	0.371	-4.089**	-2.589	-0.200	-0.134
		(0.748)	(0.761)	(2.016)	(3.175)	(1.693)	(1.638)
		[0.394]	[0.421]	[0.000]	[0.043]	[15.346]	[15.532]
	dDigMediaICT	-0.931	-0.866	-16.242***	-3.140	2.731*	2.743
		(0.685)	(0.735)	(2.884)	(2.79)	(1.575)	(1.72)
		[2.539]	[2.747]	[0.086]	[0.233]	[2.547]	[2.437]
	dOther	0.932	1.010	-2.459***	-1.455	0.935	0.891
		(0.617)	(0.657)	(0.804)	(1.333)	(1.166)	(1.244)
		[0.358]	[0.356]	[0.452]	[0.427]	[0.244]	[0.227]
Sentence ratio	***SentenceRatio***	-1.029***	-1.033***	-0.794***	-0.852***	-1.412*	-1.485**
		(0.138)	(0.139)	(0.171)	(0.194)	(0.83)	(0.716)

Table E-16. Zero-inflated negative binomial with countOpen, countRD and countIP – Model a and b (continued)

	countOpen		countRD		countIP	
	4a	**4b**	**5a**	**5b**	**6a**	**6b**
Size (*dSmall* omitted)						
	[0.470]	[0.481]	[0.608]	[1.044]	[17.766]	[25.622]
dVerySmall	-0.756**	-0.732**	-0.498	0.043	2.877	3.243
	(0.329)	(0.34)	(0.627)	(1.069)	(4.554)	(3.969)
	[0.191]	[0.183]	[0.657]	[0.863]	[155.513]	[280.809]
dMedium	-1.656	-1.698	-0.420	-0.147	5.047	5.638
	(1.482)	(1.728)	(1.322)	(2.157)	(6.556)	(5.535)
	[0.950]	[0.945]	[0.000]	[5.882]	[0.000]	[0.000]
dLarge	-0.052	-0.057	-17.467***	1.772	-36.456***	-39.735***
	(0.869)	(0.942)	(2.134)	(1.593)	(2.59)	(2.601)
Age (*dEarlyStage* omitted)						
	[0.984]	[0.932]	[0.909]	[1.030]	[2.154]	[1.999]
dYoung	-0.016	-0.071	-0.095	0.030	0.768	0.693
	(0.315)	(0.323)	(0.853)	(0.653)	(0.838)	(0.875)
	[1.206]	[1.168]	[1.026]	[0.756]	[7.787]	[8.205]
dMature	0.187	0.155	0.026	-0.280	2.052	2.105
	(0.42)	(0.43)	(0.834)	(1.602)	(1.655)	(1.537)

Table E-16. Zero-inflated negative binomial with countOpen, countRD and countIP – Model a and b (continued)

		countOpen		countRD		countIP	
		4a	**4b**	**5a**	**5b**	**6a**	**6b**
Country (*dFinland* omitted)							
		[0.079]	[0.078]	[0.000]	[0.250]	[3.141]	[3.550]
	dCanada	-2.539***	-2.546***	-19.722***	-1.385	1.145	1.267
		(0.585)	(0.59)	(2.178)	(1.711)	(1.627)	(1.474)
Period (*d2014* omitted)							
		[3.332]	[3.228]	[4.46E+08]	[7.481]	[0.809]	[0.781]
	d2010_2011	1.203	1.172	19.916***	2.012	-0.212	-0.247
		(0.744)	(0.759)	(2.505)	(1.688)	(1.201)	(1.099)
		[1.169]	[1.218]	[1.95E+08]	[2.908]	[0.144]	[0.130]
	d2012_2013	0.156	0.197	19.091***	1.068	-1.938	-2.041
		(0.668)	(0.679)	(2.286)	(2.106)	(1.863)	(1.721)
		[1.786]	[1.872]	[1.031]	[1.658]	[0.568]	[0.493]
	d2015_2016	0.580	0.627	0.030	0.506	-0.565	-0.708
		(0.566)	(0.582)	(1.07)	(2.203)	(2.137)	(1.649)
		[2.97E+04]	[2.88E+04]	[1.18E+03]	[395.361]	[452.699]	[648.698]
Constant		10.299***	10.269***	7.071***	5.980***	6.115**	6.475**
		(1.487)	(1.51)	(1.574)	(1.569)	(3.002)	(3.113)
Nb observations		1594	1594	1594	1594	1594	1594
	Nonzero obs	454	454	1317	1317	337	337
	Zero obs	1140	1140	277	277	1257	1257

Table E-16. Zero-inflated negative binomial with countOpen, countRD and countIP – Model a and b (continued)

	countOpen		countRD		countIP	
	4a	**4b**	**5a**	**5b**	**6a**	**6b**
Nb groups	1569	1569	1569	1569	1569	1569
McFadden R^2	0.172	0.172	0.143	0.143	0.126	0.126
McFadden adjusted R^2	0.155	0.155	0.137	0.137	0.101	0.101
χ^2	210.12***	211.42***	2206.63***	1981.74***	152.23***	153.32***
Loglikelihood (full model)	-1798.74	-1794.49	-5282.87	-5273.39	-1327.77	-1325.90
Loglikelihood (constant only)	-1903.00	-1903.00	-6006.07	-6006.07	-1399.79	-1399.79
Deviance (ZINB vs nbreg)	156.286***	140.753***	30.543***	30.804**	48.204***	47.327***

Note: Incidence rate ratio in square brackets. ***, **, * show significance at the 1%, 5% and 10% level respectively. Standard errors in parentheses. All alphas are positive and significant. Mean VIF is 2.13.

Table E-17. Zero-inflated negative binomial regressions – Model c

	countAll	countInnov	countCollab	countOpen	countRD	countIP
	1c	2c	3c	4c	5c	6c
Negative binomial						
Innovation evidence (*dNoInnovEvidence* omitted)						
	[1.184]	[1.200]	[0.964]	[0.931]	[1.204]	[0.964]
dSomeInnovEvidence	0.169***	0.182**	-0.037	-0.072	0.185**	-0.037
	(0.067)	(0.096)	(0.231)	(0.196)	(0.088)	(0.231)
	[1.011]	[1.208]	[1.102]	[0.831]	[0.949]	[1.102]
dStrongInnovEvidence	0.011	0.189**	0.097	-0.186	-0.053	0.097
	(0.075)	(0.105)	(0.407)	(0.246)	(0.092)	(0.407)
Time market (*dAverageTMkt* omitted)						
	[1.047]	[1.075]	[1.157]	[0.892]	[1.006]	[1.157]
dLongTMkt	0.046	0.072	0.146	-0.114	0.006	0.146
	(0.061)	(0.086)	(0.182)	(0.175)	(0.073)	(0.182)
	[1.001]	[1.092]	[0.809]	[1.258]	[0.939]	[0.809]
dShortTMkt	0.001	0.088	-0.212	0.229	-0.063	-0.212
	(0.061)	(0.084)	(0.182)	(0.195)	(0.079)	(0.182)
Sectorial category (*dSpecializedSupplier*)						
	[0.599]	[0.980]	[0.929]	[3.228]	[0.575]	[0.929]
dSupplierDominated	-0.513***	-0.020	-0.074	1.172***	-0.553***	-0.074
	(0.073)	(0.112)	(0.877)	(0.248)	(0.142)	(0.877)

Table E-17. Zero-inflated negative binomial regressions – Model c (continued)

		countAll	countInnov	countCollab	countOpen	countRD	countIP
		1c	2c	3c	4c	5c	6c
		[1.648]	[2.171]	[1.960]	[3.313]	[1.515]	[1.960]
	dKIBS	0.500***	0.775***	0.673***	1.198***	0.415***	0.673***
		(0.061)	(0.093)	(0.259)	(0.21)	(0.089)	(0.259)
		[1.088]	[1.367]	[0.767]	[1.211]	[1.105]	[0.767]
	dMassProd	0.084	0.313*	-0.265	0.191	0.100	-0.265
		(0.214)	(0.242)	(0.498)	(0.316)	(0.21)	(0.498)
		[2.338]	[2.085]	[1.775]	[2.071]	[2.273]	[1.775]
Sentence ratio	***SentenceRatio***	0.849***	0.735***	0.574	0.728***	0.821***	0.574
		(0.02)	(0.035)	(0.456)	(0.091)	(0.037)	(0.456)
Age (*dEarlyStage* omitted)							
		[1.045]	[0.864]	[0.913]	[0.736]	[1.159]	[0.913]
	dYoung	0.044	-0.146*	-0.091	-0.306*	0.147*	-0.091
		(0.058)	(0.086)	(0.528)	(0.185)	(0.078)	(0.528)
		[1.018]	[0.805]	[1.307]	[0.809]	[1.145]	[1.307]
	dMature	0.017	-0.217**	0.268	-0.212	0.136	0.268
		(0.065)	(0.098)	(0.759)	(0.221)	(0.119)	(0.759)
Country (*dFinland* omitted)							
		[2.084]	[2.034]	[2.991]	[1.548]	[2.136]	[2.991]
	dCanada	0.734***	0.710***	1.096	0.437	0.759***	1.096
		(0.077)	(0.111)	(0.971)	(0.284)	(0.104)	(0.971)

Table E-17. Zero-inflated negative binomial regressions – Model c (continued)

	countAll	**countInnov**	**countCollab**	**countOpen**	**countRD**	**countIP**
	1c	2c	3c	4c	5c	6c
Period (*d2014* omitted)						
	[0.571]	[0.568]	[0.473]	[0.767]	[0.543]	[0.473]
d2010_2011	-0.560***	-0.566***	-0.748	-0.265	-0.610***	-0.748
	(0.111)	(0.167)	(0.487)	(0.364)	(0.151)	(0.487)
	[0.593]	[0.655]	[0.361]	[0.937]	[0.499]	[0.361]
d2012_2013	-0.523***	-0.423***	-1.020***	-0.065	-0.696***	-1.020***
	(0.096)	(0.136)	(0.333)	(0.269)	(0.164)	(0.333)
	[0.759]	[0.679]	[0.768]	[1.471]	[0.758]	[0.768]
d2015_2016	-0.276***	-0.387***	-0.263	0.386	-0.277*	-0.263
	(0.086)	(0.112)	(0.509)	(0.259)	(0.147)	(0.509)
	[0.013]	[0.007]	[0.004]	[0.002]	[0.008]	[0.004]
Constant	-4.315***	-4.959***	-5.651	-6.494***	-4.833***	-5.651
	(0.181)	(0.345)	(5.252)	(0.967)	(0.41)	(5.252)
Inflate						
Sectorial category (*dSpecializedSupplier*)						
	[0.000]	[6.992]	[10.319]	[2.042]	[1.358]	[10.319]
dSupplierDominated	-19.993**	1.945	2.334	0.714*	0.306	2.334
	(8.716)	(1.251)	(8.31)	(0.383)	(2.614)	(8.31)

Table E-17. Zero-inflated negative binomial regressions – Model c (continued)

		countAll	countInnov	countCollab	countOpen	countRD	countIP
		1c	2c	3c	4c	5c	6c
		[0.000]	[1.374]	[11.351]	[0.226]	[0.280]	[11.351]
	dKIBS	-14.058*	0.317	2.429	-1.488***	-1.272	2.429
		(7.755)	(1.256)	(7.24)	(0.413)	(1.797)	(7.24)
		[368.064]	[0.738]	[1.623]	[1.444]	[4.913]	[1.623]
	dMassProd	5.908	-0.304	0.484	0.368	1.592	0.484
		(3.991)	(3.491)	(5.168)	(1.031)	(2.32)	(5.168)
		[0.000]	[0.327]	[0.479]	[0.345]	[0.383]	[0.479]***
Sentence ratio	***SentenceRatio***	-10.727*	-1.119***	-0.736***	-1.063***	-0.961	-0.736
		(5.75)	(0.206)	(0.269)	(0.147)	(0.68)	(0.269)
Age (*dEarlyStage* omitted)							
		[0.070]	[0.742]	[1.113]	[0.862]	[1.702]	[1.113]
	dYoung	-2.664	-0.299	0.107	-0.148	0.532	0.107
		(2.022)	(0.529)	(1.569)	(0.349)	(1.865)	(1.569)
		[0.000]	[0.773]	[1.373]	[1.137]	[0.954]	[1.373]
	dMature	-15.124*	-0.258	0.317	0.129	-0.047	0.317
		(9.14)	(0.598)	(1.617)	(0.422)	(3.221)	(1.617)
Country (*dFinland* omitted)							
		[0.000]	[1.132]	[1.610]	[0.053]	[0.321]	[1.610]
	dCanada	-11.567**	0.124	0.476	-2.936***	-1.137	0.476
		(4.546)	(0.974)	(5.545)	(0.577)	(0.808)	(5.545)

Table E-17. Zero-inflated negative binomial regressions – Model c (continued)

	countAll	countInnov	countCollab	countOpen	countRD	countIP
	1c	2c	3c	4c	5c	6c
Period (*d2014* omitted)						
	[0.035]	[0.093]	[0.876]	[3.996]	[8.087]	[0.876]
d2010_2011	-3.353	-2.378	-0.132	1.385	2.090	-0.132
	(5.377)	(1.827)	(2.804)	(0.872)	(3.078)	(2.804)
	[2.64E+10]	[0.661]	[0.303]	[1.208]	[2.053]	[0.303]
d2012_2013	23.998**	-0.415	-1.195	0.189	0.719	-1.195
	(9.584)	(1.102)	(4.069)	(0.724)	(6.108)	(4.069)
	[2.70E+16]	[1.032]	[1.636]	[2.062]	[4.786]	[1.636]
d2015_2016	37.833**	0.031	0.492	0.724	1.566	0.492
	(17.025)	(0.818)	(1.057)	(0.546)	(4.908)	(1.057)
	[1.86E+22]	[369.164]	[61.734]	[3.78E+04]	[56.778]	[61.734]
Constant	51.276*	5.911***	4.123	10.540***	4.039*	4.123
	(28.318)	(1.317)	(7.984)	(1.396)	(2.193)	(7.984)
Nb observations	1594	1594	1594	1594	1594	1594
Nonzero obs	1448	1448	1150	1150	931	931
Zero obs	146	146	444	444	663	663
Nb groups	1569	1569	1569	1569	1569	1569
McFadden R^2	0.132	0.123	0.116	0.160	0.128	0.117
McFadden adjusted R^2	0.128	0.118	0.109	0.148	0.124	0.100
χ^2	2422.36***	576.70***	66.98***	136.86***	788.81***	66.98***

Table E-17. Zero-inflated negative binomial regressions – Model c (continued)

	countAll	countInnov	countCollab	countOpen	countRD	countIP
	1c	2c	3c	4c	5c	6c
Loglikelihood (full model)	-6543.21	-4336.84	-1340.26	-1824.54	-5376.38	-1340.26
Loglikelihood (constant only)	-7410.66	-4739.43	-1402.07	-1912.44	-6022.24	-1402.07
Deviance (ZINB vs nbreg)	41.33***	40.75***	76.39***	116.06***	17.52*	39.90***

Note: Incidence rate ratio in square brackets. ***, **, * show significance at the 1%, 5% and 10% level respectively. Standard errors in parentheses. All alphas are positive and significant. Mean VIF is 1.73.

Table E-18. Count overall innovation signal marginal effects – Model 1a

Variable		Signal	Δ1	Δ2	Δ3	Δ4
Innovation evidence						
dNoInnovEvidence	1	62.17***				
dSomeInnovEvidence†	2	72.82***	10.65***			
dStrongInnovEvidence	3	62.84***	0.67	-9.98***		
Time to market						
dLongTMkt	1	73.66***				
dAverageTMkt†	2	72.82***	-0.84			
dShortTMkt	3	68.79***	-4.86	-4.03		
Knowledge intensity		***				
dLowKI	1	41.72***				
dHighKI†	2	72.82***	31.10***			
Sectors						
dMaterialMnfg	1	24.71***				
dLifeSciHealth	2	54.20***	29.49***			
dDigMediaICT†	3	72.82***	48.11***	18.62***		
dEnerEnvironNat	4	21.39***	-3.32	-32.81***	-51.43***	
dOther	5	50.22***	25.51***	-3.98	-22.60***	28.83 ***
Size						
dVerySmall†	1	72.82***				
dSmall	2	74.86***	2.04			
dMedium	3	50.93***	-21.89**	-23.93***		
dLarge	4	69.89***	-2.93	-4.97	18.96*	
Age						
dEarlyStage	1	66.56***				
dYoung†	2	72.82***	6.26			
dMature	3	74.92***	8.36	2.10		
Country						
dFinland	1	38.57***				
dCanada†	2	72.82***	34.25***			

Table E-18. Count overall innovation signal marginal effects – Model 1a (continued)

Variable		Signal	Δ1	Δ2	Δ3	Δ4
Period						
d2010_2011	1	49.58***				
d2012_2013	2	49.68***	0.10			
d2014†	3	72.82***	23.24***	23.14***		
d2015_2016	4	56.23***	6.65	6.55	-16.59***	

† represents the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta (Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-19. Count overall innovation signal marginal effects – Model 1b

Variable		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
Innovation evidence											
dNoInnovEvidence	1	58.08***									
dSomeInnovEvidence†	2	68.25***	10.18***								
dStrongInnovEvidence	3	58.08***	0.00	-10.17***							
Time to market											
dLongTMkt	1	69.19***									
dAverageTMkt†	2	68.25***	-0.93								
dShortTMkt	3	64.49***	-4.69	-3.76							
Sector x Knowledge intensity											
dMaterialMnfg x *dLowKI*	1	16.88***									
dMaterialMnfg x *dHighKI*	2	22.88***	6.00**								
dLifeSciHealth x *dLowKI*	3	24.20***	7.32*	1.32							
dLifeSciHealth x *dHighKI*	4	55.64***	38.75***	32.76***	31.44***						
dDigMediaICT x *dLowKI*	5	46.13***	29.25***	23.25***	21.93***	-9.51**					
dDigMediaICT x *dHighKI*†	6	68.25***	51.37***	45.38***	44.06***	12.62**	22.13***				

Table E-19. Count overall innovation signal marginal effects – Model 1b (continued)

Variable		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
dEnerEnvironNat x *dLowhKI*	7	11.04	-5.84***	-11.84***	-13.16***	-44.59***	-35.09***	-57.21***			
dEnerEnvironNat x *dHighKI*	8	21.40***	4.52	-1.48	-2.80	-34.23***	-24.73***	-46.85***	10.36***		
dOther x *dLowKI*	9	26.08***	9.20***	3.20	1.88	-29.56***	-20.05***	-42.18***	15.04***	4.68	
dOther x *dHighKI*	10	50.74***	33.86***	27.86***	26.54***	-4.90	4.61	-17.52***	39.70***	29.34***	24.66***
Size											
dVerySmall†	1	68.25									
dSmall	2	69.87***	1.62								
dMedium	3	48.12	-20.13***	-21.75***							
dLarge	4	65.11	-3.15	-4.76	-16.99*						
Age											
dEarlyStage	1	62.58***									
dYoung†	2	68.25***	5.67								
dMature	3	70.00***	7.42	1.74							
Country											
dFinland	1	37.41***									
dCanada†	2	68.25***	30.84***								

Table E-19. Count overall innovation signal marginal effects – Model 1b (continued)

Variable		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
Period											
d2010_2011	1	47.28									
d2012_2013	2	48.48***	1.20								
d2014†	3	68.25***	20.98***	19.77***							
d2015_2016	4	54.15***	6.87	5.67	-14.11***						

† represents the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-20. Count innovation culture signal marginal effects - Model 2a

Variable		Signal	Δ1	Δ2	Δ3	Δ4
Innovation evidence						
dNoInnovEvidence	1	11.97***				
dSomeInnovEvidence†	2	14.06***	2.09*			
dStrongInnovEvidence	3	13.92***	1.95*	-0.14		
Time to market						
dLongTMkt	1	14.41***				
dAverageTMkt†	2	14.06***	-0.35			
dShortTMkt	3	14.81***	0.40	0.75		
Knowledge intensity		***				
dLowKI	1	10.57***				
dHighKI†	2	14.06***	3.49***			
Sectors						
dMaterialMnfg	1	3.85***				
dLifeSciHealth	2	8.37***	4.52***			
dDigMediaICT†	3	14.06***	10.21***	5.69***		
dEnerEnvironNat	4	3.28***	-0.57	-5.09***	-10.78***	
dOther	5	9.45***	5.60***	1.08	-4.61***	6.17 ***
Size						
dVerySmall†	1	14.06***				
dSmall	2	14.14***	0.08			
dMedium	3	8.69***	-5.37***	-5.45***		
dLarge	4	16.81***	2.75	2.68	8.13**	
Age						
dEarlyStage	1	15.41***				
dYoung†	2	14.06***	-1.35			
dMature	3	13.88***	-1.53	-0.19		
Country						
dFinland	1	7.66***				
dCanada†	2	14.06***	6.40***			

Table E-20. Count innovation culture signal marginal effects - Model 2a (continued)

Variable		Signal	Δ1	Δ2	Δ3	Δ4
Period						
d2010_2011	1	9.62***				
d2012_2013	2	11.33***	1.71			
d2014†	3	14.06***	4.44**	2.73		
d2015_2016	4	10.36***	0.74	-0.97	-3.70***	

† represent the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-21. Count innovation culture signal marginal effects – Model 2b

Variable		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
Innovation evidence											
dNoInnovEvidence	1	11.87***									
dSomeInnovEvidence†	2	13.92***	2.05*								
dStrongInnovEvidence	3	13.76***	1.89*	-0.16							
Time to market											
dLongTMkt	1	14.23***									
dAverageTMkt †	2	13.92***	-0.31								
dShortTMkt	3	14.56***	0.33	0.64							
Sector x Knowledge intensity											
dMaterialMnfg x *dLowKI*	1	3.09***									
dMaterialMnfg x *dHighKI*	2	3.78**	0.69								
dLifeSciHealth x *dLowKI*	3	6.54***	3.45*	2.76							
dLifeSciHealth x *dHighKI*	4	8.32***	5.23***	4.55***	1.78						
dDigMediaICT x *dLowKI*	5	11.71***	8.62***	7.93	5.17	3.39***					
dDigMediaICT x *dHighKI*†	6	13.92***	10.83***	10.14***	7.38*	5.60***	2.21				

Table E-21. Count innovation culture signal marginal effects – Model 2b (continued)

Variable		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
dEnerEnvironNat x *dLowhKI*	7	2.43	-0.66	-1.35	-4.11**	-5.89***	-9.28***	-11.49***			
dEnerEnvironNat x *dHighKI*	8	3.28***	0.19	-0.50	-3.27	-5.05***	-8.44***	-10.65***	0.84		
dOther x *dLowKI*	9	6.41***	3.32***	2.63***	-0.14	-1.92	-5.31	-7.52***	3.97***	3.13***	
dOther x *dHighKI*	10	10.02***	6.93***	6.24***	3.48	1.70	-1.69	-3.90***	7.59***	6.75***	3.62***
Size											
dVerySmall†	1	13.92									
dSmall	2	13.95***	0.03								
dMedium	3	8.55	-5.37***	-5.40***							
dLarge	4	16.78	2.85	2.83	-8.22***						
Age											
dEarlyStage	1	15.25***									
dYoung†	2	13.92***	-1.33								
dMature	3	13.65***	-1.60	-0.27							
Country											
dFinland	1	7.53***									
dCanada†	2	13.92***	6.39***								

Table E-21. Count innovation culture signal marginal effects – Model 2b (continued)

Variable		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
Period											
d2010_2011	1	9.46									
d2012_2013	2	11.14***	1.68								
d2014†	3	13.92***	4.46**	2.78							
d2015_2016	4	10.30***	0.84	-0.84	-3.62***						

† represents the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-22. Count collaboration signal marginal effects - Model 3a

Variable	Signal		Δ1	Δ2	Δ3	Δ4
Innovation evidence						
dNoInnovEvidence	1	7.93***				
dSomeInnovEvidence†	2	8.37***	0.45			
dStrongInnovEvidence	3	6.43***	-1.50**	-1.95***		
Time to market						
dLongTMkt	1	7.23***				
dAverageTMkt †	2	8.37***	1.15			
dShortTMkt	3	8.12***	0.89	-0.25		
Knowledge intensity		***				
dLowKI	1	4.64**				
dHighKI †	2	8.37***	3.73**			
Sectors						
dMaterialMnfg	1	1.96**				
dLifeSciHealth	2	4.86***	2.90*			
dDigMediaICT†	3	8.37***	6.41***	3.51***		
dEnerEnvironNat	4	1.66***	-0.30*	-3.20***	-6.71***	
dOther	5	5.63***	3.67***	0.77**	-2.74***	3.97 ***
Size						
dVerySmall†	1	8.37***				
dSmall	2	10.89***	2.52**			
dMedium	3	8.19***	-0.19**	-2.70***		
dLarge	4	7.84*	-0.54	-3.05**	-0.35	
Age						
dEarlyStage	1	6.87***				
dYoung†	2	8.37***	1.50			
dMature	3	8.42***	1.55	0.04		
Country						
dFinland	1	4.93***				
dCanada†	2	8.37***	3.44***			

Table E-22. Count collaboration signal marginal effects - Model 3a (continued)

Variable			Signal	Δ1	Δ2	Δ3	Δ4
Period							
	d2010_2011	1	5.69***				
	d2012_2013	2	5.88***	0.20			
	d2014†	3	8.37***	2.69**	2.49**		
	d2015_2016	4	7.27***	1.58*	1.39**	-1.10	

† represent the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-23. Count collaboration signal marginal effects – Model 3b

Variable		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
Innovation evidence											
dNoInnovEvidence	1	8.38***									
dSomeInnovEvidence†	2	8.86***	0.48								
dStrongInnovEvidence	3	6.76***	-1.62*	-2.10***							
Time to market											
dLongTMkt	1	7.63***									
dAverageTMkt †	2	8.86***	1.24								
dShortTMkt	3	8.52***	0.90	-0.34							
Sector x Knowledge intensity											
dMaterialMnfg x *dLowKI*	1	1.35									
dMaterialMnfg x *dHighKI*	2	1.86**	0.51								
dLifeSciHealth x *dLowKI*	3	3.72**	2.38	1.86							
dLifeSciHealth x *dHighKI*	4	4.80***	3.45	2.94*	1.07						
dDigMediaICT x *dLowKI*	5	4.25**	2.91*	2.39***	0.53	-0.54**					
dDigMediaICT x *dHighKI*†	6	8.86***	7.52***	7.00***	5.14**	4.07***	4.61**				

Table E-23. Count collaboration signal marginal effects – Model 3b (continued)

Variable		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
dEnerEnvironNat x *dLowhKI*	7	0.90*	-0.45*	-0.96*	-2.82*	-3.89***	-3.35***	-7.96***			
dEnerEnvironNat x *dHighKI*	8	1.71***	0.36	-0.16	-2.02	-3.09***	-2.55***	-7.16***	0.80		
dOther x *dLowKI*	9	3.19**	1.85	1.33***	-0.53	-1.60	-1.06*	-5.67***	2.29***	1.49***	
dOther x *dHighKI*	10	5.92***	4.57*	4.05***	2.19	1.12*	1.66	-2.95***	5.01***	4.21***	2.72*
Size											
dVerySmall†	1	8.86***									
dSmall	2	11.46***	2.59**								
dMedium	3	8.64***	-0.22**	-2.82***							
dLarge	4	8.33**	-0.54	-3.13**	0.31						
Age											
dEarlyStage	1	7.35***									
dYoung†	2	8.86***	1.52								
dMature	3	8.90***	1.55	0.04							
Country											
dFinland	1	5.06***									
dCanada†	2	8.86***	3.80***								

Table E-23. Count collaboration signal marginal effects – Model 3b (continued)

Variable		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
Period											
d2010_2011	1	5.82									
d2012_2013	2	6.03***	0.22								
d2014†	3	8.86***	3.05**	2.83**							
d2015_2016	4	7.61***	1.79*	1.58**	-1.26						

† represents the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-24. Count OI signal marginal effects - Model 4a

Variable		Signal	Δ1	Δ2	Δ3	Δ4
Innovation evidence						
dNoInnovEvidence	1	1.36**				
dSomeInnovEvidence†	2	1.50***	0.13			
dStrongInnovEvidence	3	1.36***	0.00	-0.14		
Time to market						
dLongTMkt	1	1.25***				
dAverageTMkt †	2	1.50***	0.24			
dShortTMkt	3	1.92***	0.66***	0.42		
Knowledge intensity		***				
dLowKI	1	2.02**				
dHighKI †	2	1.50***	-0.52**			
Sectors						
dMaterialMnfg	1	0.38**				
dLifeSciHealth	2	0.51***	0.13			
dDigMediaICT†	3	1.50***	1.11**	0.98***		
dEnerEnvironNat	4	0.21***	-0.18***	-0.31***	-1.29***	
dOther	5	1.30***	0.91**	0.78***	-0.20	1.09 ***
Size						
dVerySmall†	1	1.50***				
dSmall	2	2.39***	0.89***			
dMedium	3	0.84*	-0.66**	-1.55***		
dLarge	4	2.21*	0.71	-0.18	1.37***	
Age						
dEarlyStage	1	1.71***				
dYoung†	2	1.50***	-0.22			
dMature	3	1.69***	-0.02	0.19		
Country						
dFinland	1	0.73***				
dCanada†	2	1.50***	0.77			

Table E-24. Count OI signal marginal effects - Model 4a (continued)

Variable		Signal	Δ1	Δ2	Δ3	Δ4
Period						
d2010_2011	1	1.23***				
d2012_2013	2	1.83***	0.60			
d2014†	3	1.50***	0.27	-0.33		
d2015_2016	4	2.34***	1.12*	0.52	0.84*	

† represent the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-25. Count OI signal marginal effects – Model 4b

Variable		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
Innovation evidence											
dNoInnovEvidence	1	1.33**									
dSomeInnovEvidence†	2	1.52***	0.19								
dStrongInnovEvidence	3	1.38***	0.05	-0.14							
Time to market											
dLongTMkt	1	1.36***									
dAverageTMkt †	2	1.52***	0.16								
dShortTMkt	3	1.99***	0.63**	0.47							
Sector x Knowledge intensity											
dMaterialMnfg x *dLowKI*	1	1.18									
dMaterialMnfg x *dHighKI*	2	0.27**	-0.91***								
dLifeSciHealth x *dLowKI*	3	0.73*	-0.45**	0.46							
dLifeSciHealth x *dHighKI*	4	0.50***	-0.67***	0.23	-0.22						
dDigMediaICT x *dLowKI*	5	1.45**	0.28*	1.19**	0.73	0.95**					
dDigMediaICT x *dHighKI*†	6	1.52***	0.34*	1.25***	0.79*	1.01***	0.06				

Table E-25. Count OI signal marginal effects – Model 4b (continued)

Variable		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
dEnerEnvironNat x *dLowhKI*	7	0.20**	-0.98***	-0.07*	-0.53***	-0.30**	-1.25***	-1.32***			
dEnerEnvironNat x *dHighKI*	8	0.22***	-0.96***	-0.05**	-0.51***	-0.28***	-1.23***	-1.30***	0.02		
dOther x *dLowKI*	9	1.59***	0.42	1.33***	0.87**	1.09***	0.14***	0.08*	1.39***	1.37***	
dOther x *dHighKI*	10	1.20***	0.02*	0.93***	0.47	0.70***	-0.25	-0.32	1.00***	0.98***	-0.39*
Size											
dVerySmall†	1	1.52***									
dSmall	2	2.29***	0.77***								
dMedium	3	0.91*	-0.61*	-1.38***							
dLarge	4	2.13*	0.62	-0.16	-1.22***						
Age											
dEarlyStage	1	1.83***									
dYoung†	2	1.52***	-0.31								
dMature	3	1.85**	0.02	0.34							
Country											
dFinland	1	0.72***									
dCanada†	2	1.52***	0.80								

Table E-25. Count OI signal marginal effects – Model 4b (continued)

Variable		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
Period											
d2010_2011	1	1.24***									
d2012_2013	2	1.84***	0.60								
d2014†	3	1.52***	0.28	-0.33							
d2015_2016	4	2.38***	1.14*	0.54	0.87*						

† represents the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-26. Count RD signal marginal effects - Model 5a

Variable		Signal	Δ1	Δ2	Δ3	Δ4
Innovation evidence						
dNoInnovEvidence	1	23.51***				
dSomeInnovEvidence†	2	28.05***	4.54***			
dStrongInnovEvidence	3	22.65***	-0.86	-5.39***		
Time to market						
dLongTMkt	1	28.04***				
dAverageTMkt †	2	28.05***	0.01			
dShortTMkt	3	25.34***	-2.70	-2.71		
Knowledge intensity		***				
dLowKI	1	15.83***				
dHighKI †	2	28.05***	12.21***			
Sectors						
dMaterialMnfg	1	10.33***				
dLifeSciHealth	2	28.25***	17.92***			
dDigMediaICT†	3	28.05***	17.72***	-0.21		
dEnerEnvironNat	4	10.70***	0.37	-17.55***	-17.35***	
dOther	5	20.87***	10.54***	-7.38***	-7.17***	10.17 ***
Size						
dVerySmall†	1	28.05***				
dSmall	2	29.96***	1.91			
dMedium	3	17.17***	-10.88***	-12.79***		
dLarge	4	28.13***	0.08	-1.83	10.96**	
Age						
dEarlyStage	1	24.65***				
dYoung†	2	28.05***	3.39**			
dMature	3	29.56***	4.91**	1.52		
Country						
dFinland	1	17.45***				
dCanada†	2	28.05***	10.59***			

Table E-26. Count RD signal marginal effects - Model 5a (continued)

Variable		Signal	Δ1	Δ2	Δ3	Δ4
Period						
d2010_2011	1	20.15***				
d2012_2013	2	18.54***	-1.62			
d2014†	3	28.05***	7.89***	9.51***		
d2015_2016	4	21.53***	1.38	2.99	-6.52***	

† represent the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-27. Count RD signal marginal effects – Model 5b

Variable		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
Innovation evidence											
dNoInnovEvidence	1	22.25***									
dSomeInnovEvidence†	2	26.83***	4.58***								
dStrongInnovEvidence	3	20.99***	-1.25	-5.84***							
Time to market											
dLongTMkt	1	26.99***									
dAverageTMkt †	2	26.83***	-0.16								
dShortTMkt	3	24.31***	-2.68*	-2.52							
Sector x Knowledge intensity											
dMaterialMnfg x *dLowKI*	1	8.03***									
dMaterialMnfg x *dHighKI*	2	9.48***	1.45								
dLifeSciHealth x *dLowKI*	3	9.38***	1.34	-0.10							
dLifeSciHealth x *dHighKI*	4	31.00***	22.96***	21.52***	21.62***						
dDigMediaICT x *dLowKI*	5	17.96***	9.93***	8.48***	8.58***	-13.03***					
dDigMediaICT x *dHighKI*†	6	26.83***	18.80***	17.35***	17.45***	-4.16	8.87***				

Table E-27. Count RD signal marginal effects – Model 5b (continued)

Variable		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
dEnerEnvironNat x *dLowhKI*	7	4.89***	-3.15**	-4.59***	-4.49***	-26.11***	-13.07***	-21.94***			
dEnerEnvironNat x *dHighKI*	8	10.53***	2.49	1.05	1.15	-20.47***	-7.43***	-16.30***	5.64***		
dOther x *dLowKI*	9	10.74***	2.71*	1.26	1.37	-20.25***	-7.22***	-16.09***	5.86***	0.22	
dOther x *dHighKI*	10	21.38***	13.34***	11.90***	12.00***	-9.62***	3.42*	-5.45***	16.49***	10.85***	10.63***
Size											
dVerySmall†	1	26.83***									
dSmall	2	27.75***	0.92								
dMedium	3	16.67***	-10.16**	-11.08***							
dLarge	4	27.80***	0.97	0.05	-11.13**						
Age											
dEarlyStage	1	23.46***									
dYoung†	2	26.83***	3.37**								
dMature	3	27.79***	4.33**	0.96							
Country											
dFinland	1	17.31***									
dCanada†	2	26.83***	9.52***								

Table E-27. Count RD signal marginal effects – Model 5b (continued)

Variable		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
Period											
d2010_2011	1	19.67***									
d2012_2013	2	18.62***	-1.05								
d2014†	3	26.83***	7.16**	8.21***							
d2015_2016	4	21.80***	2.13	3.18	-5.03						

† represents the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-28. Count IP signal marginal effects - Model 6a

Variable		Signal	Δ1	Δ2	Δ3	Δ4
Innovation evidence						
dNoInnovEvidence	1	1.89***				
dSomeInnovEvidence†	2	1.82***	-0.06			
dStrongInnovEvidence	3	1.95***	0.07	0.13		
Time to market						
dLongTMkt†	1	1.82***				
dAverageTMkt	2	1.41***	-0.41*			
dShortTMkt	3	1.16***	-0.66***	-0.25		
Knowledge intensity						
dLowKI	1	0.95***				
dHighKI †	2	1.82***	0.88***			
Sectors						
dMaterialMnfg	1	0.88***				
dLifeSciHealth†	2	1.82***	0.95**			
dDigMediaICT	3	1.20	0.33***	-0.62		
dEnerEnvironNat	4	0.64***	-0.23	-1.18***	-0.56***	
dOther	5	1.54**	0.66***	-0.29	0.33*	0.89***
Size						
dVerySmall	1	1.34				
dSmall†	2	1.82***	0.48***			
dMedium	3	0.31	-1.03	-1.52		
dLarge	4	1.28***	-0.06***	-0.54	0.97**	
Age						
dEarlyStage	1	1.03***				
dYoung	2	1.19***	0.17			
dMature†	3	1.82***	0.80***	0.63**		
Country						
dFinland	1	0.72***				
dCanada†	2	1.82***	1.10***			

Table E-28. Count IP signal marginal effects - Model 6a (continued)

Variable			Signal	Δ1	Δ2	Δ3	Δ4
Period							
	d2010_2011	1	1.10***				
	d2012_2013	2	0.90***	-0.20			
	d2014†	3	1.82***	0.73*	0.92***		
	d2015_2016	4	1.18***	0.08	0.28	-0.65	

† represent the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-29. Count IP signal marginal effects – Model 6b

Variable	Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
Innovation evidence										
dNoInnovEvidence	1	1.94***								
dSomeInnovEvidence†	2	1.85***	-0.10							
dStrongInnovEvidence	3	1.96***	0.02	0.12						
Time to market										
dLongTMkt†	1	1.85***								
dAverageTMkt	2	1.47***	-0.37*							
dShortTMkt	3	1.24***	-0.60**	-0.23						
Sector x Knowledge intensity										
dMaterialMnfg x *dLowKI*	1	0.41*								
dMaterialMnfg x *dHighKI*	2	0.80***	0.39							
dLifeSciHealth x *dLowKI*	3	0.43***	0.02	-0.37						
dLifeSciHealth x *dHighKI*†	4	1.85***	1.44***	1.05***	1.42**					
dDigMediaICT x *dLowKI*	5	0.90	0.49**	0.10**	0.48*	-0.95				
dDigMediaICT x *dHighKI*	6	1.09	0.68***	0.29***	0.66**	-0.76	0.18			

Table E-29. Count IP signal marginal effects – Model 6b (continued)

Variable		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
dEnerEnvironNat x *dLowhKI*	7	0.34*	-0.07	-0.46**	-0.09	-1.51***	-0.57***	-0.75***			
dEnerEnvironNat x *dHighKI*	8	0.58***	0.17	-0.22	0.16	-1.27***	-0.32***	-0.50***	0.25		
dOther x *dLowKI*	9	0.70***	0.29	-0.10	0.27	-1.15***	-0.20**	-0.39***	0.36*	0.12	
dOther x *dHighKI*	10	1.47***	1.06**	0.67***	1.04*	-0.38	0.57	0.38	1.13***	0.89***	0.77***
Size											
dVerySmall	1	1.21									
dSmall†	2	1.85***	0.64***								
dMedium	3	0.23	-0.99	-1.62*							
dLarge	4	1.32***	0.11***	-0.52	-1.10**						
Age											
dEarlyStage	1	1.09***									
dYoung	2	1.22***	0.13								
dMature†	3	1.85***	0.76***	0.63**							
Country											
dFinland	1	0.72***									
dCanada†	2	1.82***	1.10***								

Table E-29. Count IP signal marginal effects – Model 6b (continued)

Variable		Signal	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
Period											
d2010_2011	1	1.14***									
d2012_2013	2	1.00***	-0.14								
d2014†	3	1.85***	0.71*	0.85**							
d2015_2016	4	1.24***	0.10	0.24*	-0.61						

† represents the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-30. Count overall innovation signal marginal effects –Model 1c

Variable		Signal	Δ1	Δ2	Δ3
Innovation evidence					
dNoInnovEvidence	1	58.13***			
dSomeInnovEvidence†	2	68.80***	10.68***		
dStrongInnovEvidence	3	58.77***	0.65	-10.03***	
Time to market					
dLongTMkt	1	72.01***			
dAverageTMkt †	2	68.80***	-3.21		
dShortTMkt	3	68.86***	-3.15	0.06	
Sectorial category					
dSupplierDominated	1	25.00***			
dSpecializedSupplier	2	41.74***	16.75***		
dKIBS†	3	68.80***	43.81***	27.06***	
dMassProd	4	45.42***	20.42***	3.68	-23.38**
Age					
dEarlyStage	1	65.81***			
dYoung†	2	68.80***	2.99		
dMature	3	66.97***	1.16	-1.83	
Country					
dFinland	1	33.02***			
dCanada†	2	68.80***	35.79***		
Period					
d2010_2011	1	39.29***			
d2012_2013	2	40.79***	1.50		
d2014†	3	68.80***	29.51***	28.02***	
d2015_2016	4	52.21***	12.92***	11.42***	-16.59***

† represents the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-31. Count innovation culture signal marginal effects – Model 2c

Variable		Signal	Δ1	Δ2	Δ3
Innovation evidence					
dNoInnovEvidence	1	12.02***			
dSomeInnovEvidence†	2	14.43***	2.40**		
dStrongInnovEvidence	3	14.52***	2.50**	0.09	
Time to market					
dLongTMkt	1	15.51***			
dAverageTMkt †	2	14.43***	-1.08		
dShortTMkt	3	15.75***	0.24	1.32	
Sectorial category					
dSupplierDominated	1	5.68***			
dSpecializedSupplier	2	6.71***	1.03		
dKIBS†	3	14.43***	8.75***	7.72***	
dMassProd	4	9.24***	3.56*	2.53*	-5.19**
Age					
dEarlyStage	1	16.49***			
dYoung†	2	14.43***	-2.06**		
dMature	3	13.41***	-3.08**	-1.02	
Country					
dFinland	1	7.12***			
dCanada†	2	14.43***	7.31***		
Period					
d2010_2011	1	8.47***			
d2012_2013	2	9.57***	1.10		
d2014†	3	14.43***	5.96***	4.86***	
d2015_2016	4	9.79***	1.32	0.22	-4.64***

† represents the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-32. Count collaboration signal marginal effects –Model 3c

Variable	Signal		Δ1	Δ2	Δ3
Innovation evidence					
dNoInnovEvidence	1	8.63***			
dSomeInnovEvidence†	2	9.34***	0.70		
dStrongInnovEvidence	3	7.27***	-1.36*	-2.06***	
Time to market					
dLongTMkt	1	8.49***			
dAverageTMkt †	2	9.34***	0.85		
dShortTMkt	3	9.40***	0.91	0.06	
Sectorial category					
dSupplierDominated	1	3.21***			
dSpecializedSupplier	2	4.29***	1.09		
dKIBS†	3	9.34***	6.13***	5.04***	
dMassProd	4	3.68***	0.47**	-0.62*	-5.66***
Age					
dEarlyStage	1	7.57***			
dYoung†	2	9.34***	1.76		
dMature	3	9.30***	1.73	-0.04	
Country					
dFinland	1	5.41***			
dCanada†	2	9.34***	3.93***		
Period					
d2010_2011	1	5.25***			
d2012_2013	2	5.51***	0.26		
d2014†	3	9.34***	4.09***	3.83***	
d2015_2016	4	7.95***	2.70***	2.44***	-1.39

† represents the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-33. Count OI signal marginal effects – Model 4c

Variable	Signal		Δ1	Δ2	Δ3
Innovation evidence					
dNoInnovEvidence	1	2.38***			
dSomeInnovEvidence†	2	2.22***	-0.17		
dStrongInnovEvidence	3	1.98***	-0.40	-0.24	
Time to market					
dLongTMkt	1	1.98***			
dAverageTMkt †	2	2.22***	0.24		
dShortTMkt	3	2.79***	0.81**	0.57	
Sectorial category					
dSupplierDominated	1	1.53***			
dSpecializedSupplier	2	0.57***	-0.96***		
dKIBS†	3	2.22***	0.69	1.65***	
dMassProd	4	0.64***	-0.89***	0.07	-1.58***
Age					
dEarlyStage	1	2.99***			
dYoung†	2	2.22***	-0.77*		
dMature	3	2.40***	-0.59	0.18	
Country					
dFinland	1	0.75***			
dCanada†	2	2.22***	1.47		
Period					
d2010_2011	1	1.47***			
d2012_2013	2	2.06***	0.58		
d2014†	3	2.22***	0.74	0.16	
d2015_2016	4	3.09***	1.62**	1.04**	0.88

† represents the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-34. Count R&D signal marginal effects – Model 5c

Variable		Signal	Δ1	Δ2	Δ3
Innovation evidence					
dNoInnovEvidence	1	28.45***			
dSomeInnovEvidence†	2	34.25***	5.80**		
dStrongInnovEvidence	3	26.99***	-1.46	-7.26***	
Time to market					
dLongTMkt	1	34.44***			
dAverageTMkt †	2	34.25***	-0.20		
dShortTMkt	3	32.15***	-2.30	-2.10	
Sectorial category					
dSupplierDominated	1	12.87***			
dSpecializedSupplier	2	22.44***	9.57***		
dKIBS†	3	34.25***	21.38***	11.80***	
dMassProd	4	23.86***	10.99***	1.42	-10.38*
Age					
dEarlyStage	1	29.59***			-
dYoung†	2	34.25***	4.65**		-
dMature	3	33.90***	4.31	-0.35	-
Country					
dFinland	1	15.93***			-
dCanada†	2	34.25***	18.31***		-
Period					
d2010_2011	1	18.23***			-
d2012_2013	2	17.02***	-1.21		-
d2014†	3	34.25***	16.01***	17.22***	-
d2015_2016	4	25.69***	7.45***	8.67***	-8.56**

† represent the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-35. Count IP signal marginal effects –Model 6c

Variable		Signal	Δ1	Δ2	Δ3
Innovation evidence					
dNoInnovEvidence	1	1.49***			
dSomeInnovEvidence†	2	1.44***	-0.05		
dStrongInnovEvidence	3	1.64***	0.15	0.21	
Time to market					
dLongTMkt†	1	1.44***			
dAverageTMkt	2	1.24***	-0.20*		
dShortTMkt	3	1.00***	-0.43***	-0.24	
Sectorial category					
dSupplierDominated	1	0.43			
dSpecializedSupplier†	2	1.44***	1.01**		
dKIBS	3	0.84	0.41***	-0.59	
dMassProd	4	0.97	0.54	-0.47***	0.12***
Age					
dEarlyStage	1	1.17***			
dYoung	2	1.05***	-0.12		
dMature†	3	1.44***	0.27***	0.39**	
Country					
dFinland	1	0.53**			
dCanada†	2	1.44***	0.91***		
Period					
d2010_2011	1	0.70***			
d2012_2013	2	0.61	-0.08		
d2014†	3	1.44***	0.74*	-0.08***	
d2015_2016	4	0.96**	0.27	0.74	0.82

† represents the group for which the predicted value was calculated. Wald tests are used to evaluate the significance of the Delta(Δ) values: ***, **, * show significance at the 1%, 5% and 10% level respectively.

Table E-36. Results summary – Overall innovation

	Overall innovation signal	
Method	**One-way ANOVA**	**Regression**
Propositions		
Evidence of innovation results		
Proposition 1	Partially, firms with some evidence have a stronger signal than all the other firms. However, firms with strong innovation evidence have a stronger signal than the firms with no evidence.	Partially, firms with some evidence have the strongest corresponding signal, but firms with strong innovation evidence have more signalling strength than the firms with no evidence.
Time to market		
Proposition 2	Yes, firms with a longer time to market have higher ranks than the other firms. There is no difference between firms that have a shorter time to market and the average group.	No, there is no significant difference.
Knowledge intensity level		
Proposition 3	Yes, firms with high knowledge intensity have a stronger signal than the other firms.	Yes, firms with high knowledge intensity have a stronger signal than the other firms.
Sector		
Proposition 4	Partially, firms in Life science and health have a stronger signal than all the other sectors, digital media and other ICT firms have a stronger signal than all the remaining sector excepting Energy and environment.	Partially, firms in digital media and other ICT have a stronger signal than all the other sectors, Life science and health have a stronger signal than all the remaining sector excepting with Other.
Proposition 5	-	Yes, firms from the same sectors with high knowledge intensity have a stronger signal than the other firms of the same sector.

Table E-36. Results summary – Overall innovation (continued)

	Overall innovation signal	
Method	**One-way ANOVA**	**Regression**
Sectorial category		
Proposition 6	Yes, firms from either mass production or value-added knowledge-based industries have a stronger signal than the firms classified as supplier-dominated.	Yes, firms from either mass production or value-added knowledge-based industries have a stronger signal than the firms classified as supplier-dominated.
Proposition 7	No, there is no significant difference between KIBS specialised-suppliers.	Yes, firms classified as KIBS have a stronger signal than those classified as specialised-suppliers.
Control effects		
Size	Yes, it seems there is a significant difference in signal rankings due to a firm's size, notably between firms below 100 employees and over 100 employees as there is no difference between very small firms and small firms and between medium and large firms.	Yes, firms of all sizes have higher than firms with medium size.
Age	No, there is no significant difference.	No, there is no significant difference.
Country	Yes, Canadian firms have a stronger signal than the Finnish firms.	Yes, Canadian firms seem to have a stronger signal.

Table E-37. Results summary – Innovation culture

	Innovation culture signal	
Method	**One-way ANOVA**	**Regression**
Propositions		
Evidence of innovation results		
Proposition 1a	Yes, firms with higher innovation evidence have a stronger signal.	Partially, firms with some or strong innovation evidence have a stronger signal than firms with no evidence. However, there is no difference between firms with some or strong innovation evidence.
Time to market		
Proposition 2a	Yes, firms with longer or shorter time to market have a stronger signal than the other firms. There is no difference between firms with longer or shorter time to market.	No, there is no significant difference.
Knowledge intensity level		
Proposition 3a	Yes, firms with high knowledge intensity have a stronger signal than the other firms.	Yes, firms with high knowledge intensity have a stronger signal than the other firms.
Sector		
Proposition 4a	Yes, firms in digital media and other ICT have a stronger signal than all the other sectors, Life science and health have a stronger signal than all the remaining sector.	Yes, firms in digital media and other ICT have a stronger signal than all the other sectors, Life science and health have a stronger signal than all the remaining sector, but not different than Other.
Proposition 5a	-	No, firms from the same sectors with high knowledge intensity does not seem to have a stronger signal than the other firms of the same sector, except for Other.

Table E-37. Results summary – Innovation culture (continued)

	Innovation culture signal	
Method	**One-way ANOVA**	**Regression**
Sectorial category		
Proposition 6a	Yes, firms from either mass production or value-added knowledge-based industries have a stronger signal than the firms classified as supplier-dominated.	Partially, firms classified as KIBS have a stronger signal than those classified as supplier-dominated, but there is no differences between specialised production and supplier-dominated.
Proposition 7a	Yes, firms classified as KIBS have a stronger signal than those classified as specialised-suppliers.	Yes, firms classified as KIBS have a stronger signal than those classified as specialised-suppliers.
Control effects		
Size	No, the larger firms have a stronger signal.	No, very small, small and large firms do not seem to have a difference in the strength of their signal but have a stronger signal than medium firms.
Age	No, there is no significant difference.	No, there is no significant difference.
Country	Yes, Canadian firms have a stronger signal.	Yes, Canadian firms seem to have a stronger signal.

Table E-38. Results summary – Collaboration

	Collaboration signal	
Method	**One-way ANOVA**	**Regression**
Propositions		
Evidence of innovation results		
Proposition 1b	Partially, firms with some evidence have a stronger signal. Firms with strong evidence are not ranked higher than firms with no evidence.	No, there is no difference between firms with no evidence and some evidence and both have a stronger signal than the firms with strong evidence.
Time to market		
Proposition 2b	Yes, firms with a longer time to market have a stronger signal than the average group but does not have a difference with those with shorter time to market.	No, there is no significant difference.
Knowledge intensity level		
Proposition 3b	Yes, firms with high knowledge intensity have a stronger signal than the other firms.	Yes, firms with high knowledge intensity have a stronger signal than the other firms.
Sector		
Proposition 4b	Partially, firms in Life science and health have a stronger signal than all the other sectors, digital media and other ICT have a stronger signal than all the remaining sector excepting Natural resources.	Partially, firms in digital media and other ICT have a stronger signal than all the other sectors, Life science and health have a stronger signal than all the remaining sector, but with Other which is significantly higher.
Proposition 5b	-	Partially, only firms in Digital media and ICT and Other with high knowledge intensity have a stronger signal than the other firms of their sector.

Table E-38. Results summary – Collaboration (continued)

	Collaboration signal	
Method	**One-way ANOVA**	**Regression**
Sectorial category		
Proposition 6b	Partially, firms from either mass production or value-added knowledge-based industries have a stronger signal than the firms classified as supplier-dominated, but scale-based.	Partially, firms classified as KIBS have a stronger signal than those classified as supplier-dominated, but there are no differences between specialised production and supplier-dominated.
Proposition 7b	No, there is no significant difference.	Yes, firms classified as KIBS have a stronger signal than specialised-suppliers.
Control effects		
Size	Partially, medium-size firms have a stronger signal than small and very small firms, but not with a significant difference with large firms.	Yes, small firms seem to have a stronger signal than very small, medium and large firm.
Age	Yes, early-stage firms are ranked lower than other firms.	No, there is no significant difference.
Country	Yes, Canadian firms seem to have a stronger signal.	Yes, Canadian firms seem to have a stronger signal.

Table E-39. Results summary – OI

	OI signal	
Method	**One-way ANOVA**	**Regression**
Propositions		
Evidence of innovation results		
Proposition 1c	Partially, there is no difference between some and strong evidence which both have a stronger signal no evidence.	No, there is no significant difference.
Time to market		
Proposition 2c	Yes, as both firms with longer and shorter time to market have a stronger signal than the other firms. There is no difference between firms with longer and shorter time to market.	Yes, firms with a shorter time to market have a stronger signal than the firms with a longer time to market have a stronger signal than the firms.
Knowledge intensity level		
Proposition 3c	No, firms with low knowledge have a stronger signal than the other firms.	No, firms with low knowledge intensity have a stronger signal than the other firms.
Sector		
Proposition 4c	Partially, firms in digital media and other ICT have a stronger signal than most sectors excepting Natural resources and Energy and environment. Firms in Life science and health have a stronger signal than manufacturing and other firms.	Partially, firms in digital media and other ICT have a stronger signal than most sectors excepting Other. Firms in Life science and health only have a stronger signal than Natural resources, Energy and environment.
Proposition 5c	-	No, Materials and manufacturing and Other with low KI have a stronger signal than with high KI.

Table E-39. Results summary – OI (continued)

	OI signal	
Method	**One-way ANOVA**	**Regression**
Sectorial category		
Proposition 6c	Partially, firms from either mass production or value-added knowledge-based industries have a stronger signal than the firms classified as supplier-dominated, but the ones as specialised-suppliers.	Partially, firms classified as mass production have a stronger signal than those classified as supplier-dominated, but supplier dominated firms seem to have a stronger signal than specialised production.
Proposition 7c	Yes, firms classified as KIBS have a stronger signal than those classified as specialised-suppliers.	Yes, firms classified as KIBS have a stronger signal than those classified as specialised-suppliers.
Control effects		
Size	Yes, large firms have a stronger signal.	No, small firms have a stronger signal than very small firms and medium firms. Very small firms have a stronger signal than medium firms and large firms have a stronger signal than medium firms.
Age	Yes, mature firms have a stronger signal.	No, there is no significant difference.
Country	Yes, Canadian firms have a stronger signal	No, there is no significant difference.

Table E-40. Results summary – R&D

	R&D signal	
Method	**One-way ANOVA**	**Regression**
Propositions		
Evidence of innovation results		
Proposition 1d	Partially, firms with some evidence have a stronger signal, but firms with strong innovation evidence have a stronger signal than the firms with no evidence.	Partially substantiated as firms with some evidence have the strongest corresponding signal.
Time to market		
Proposition 2d	Yes, firms with a longer time to market have a stronger signal than the other firms. There is no difference between the firms that have a shorter time to market and the average group.	No, there is no significant difference.
Knowledge intensity level		
Proposition 3d	Yes, firms with high knowledge intensity have a stronger signal than the other firms.	Yes, firms with high knowledge intensity have a stronger signal than the other firms.
Sector		
Proposition 4d	Partially, firms in Life science and health have a stronger signal than all the other sectors, while digital media and other ICT have a stronger signal than all the remaining sector excepting Energy and environment.	Yes, firms in Life science and health and digital media and other ICT have a stronger signal than all the other sectors. There is no difference between Life science and health and digital media and other ICT.
Proposition 5d	-	Partially, firms in most sectors with high knowledge intensity does seem to have a stronger signal than the other firms of the same sector, except for manufacturing.

Table E-40. Results summary – R&D (continued)

	R&D signal	
Method	**One-way ANOVA**	**Regression**
Sectorial category		
Proposition 6d	Yes, firms from either mass production or value-added knowledge-based industries have a stronger signal than the firms classified as supplier-dominated.	Yes, firms from either mass production or value-added knowledge-based industries have a stronger signal than the firms classified as supplier-dominated.
Proposition 7d	No, firms classified as specialised-suppliers have a stronger signal than those classified as KIBS.	Yes, firms classified as KIBS have a stronger signal than those classified as specialised-suppliers.
Control effects		
Size	Yes, large firms have a stronger signal.	No, there is no significant difference between very small, small and large firms, but they all have stronger signal than medium firms.
Age	Yes, early-stage firms are ranked lower than other firms.	Yes, early-stage firms seem to have significantly less signal strength.
Country	Yes, Canadian firms have a stronger signal.	Yes, Canadian firms seem to have a stronger signal.

Table E-41. Results summary – IP

	IP signal	
Method	**One-way ANOVA**	**Regression**
Propositions		
Evidence of innovation		
Proposition 1e	Partially, there is no difference between some and strong evidence which both have a stronger signal than no evidence.	No, there is no significant difference.
Time to market		
Proposition 2e	Yes, firms with a longer time to market have a stronger signal than the other firms. There is no difference between the firms that have a shorter time to market and the average group.	Yes, firms with a longer time to market have a stronger signal than the other firms. There is no difference between the firms that have a shorter time to market and the average group.
Knowledge intensity level		
Proposition 3e	Yes, firms with high knowledge intensity have a stronger signal than the other firms.	Yes, firms with high knowledge intensity have a stronger signal than the other firms.
Sector		
Proposition 4e	Partially, firms in Life science and health have a stronger signal than all the other sectors, manufacturing only have a stronger signal than firms in the sector classified as Other.	Partially, firms in Life science and health and Digital media and ICT have a stronger signal than manufacturing, Natural resources, energy, and environment. Manufacturing does not seem to have a stronger signal than any other sector.
Proposition 5e	-	Partially, firms in life science and other sectors with high knowledge intensity does seem to have a stronger signal than the other firms of the same sector.

Table E-41. Results summary – IP (continued)

	IP signal	
Method	**One-way ANOVA**	**Regression**
Sectorial category		
Proposition 6e	Yes, firms from either mass production or value-added knowledge-based industries have a stronger signal than the firms classified as supplier-dominated.	Yes, knowledge-based industries have a stronger signal than the firms classified as supplier-dominated.
Proposition 7e	Yes, firms classified as specialised-suppliers have a stronger signal than those classified as KIBS.	No, firms classified as KIBS have a stronger signal than those classified as specialised-suppliers.
Control effects		
Size	Partially, as large firms have a stronger signal than small and very small firms, but not with a difference with medium firms.	No, very small firms and medium firms have a stronger signal than small firms and large firms.
Age	Yes, mature firms have a stronger signal than the other firms.	Yes, mature firms seem to have a stronger signal than the other firms.
Country	Yes, Canadian firms seem to have a stronger signal.	Yes, Canadian firms seem to have a stronger signal.