Titre: Title:	The assessment of user knowledge with a bayesian framework and its comparison with item response theory
Auteur: Author:	Xiaoming Pu
Date:	2005
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Pu, X. (2005). The assessment of user knowledge with a bayesian framework and its comparison with item response theory [Master's thesis, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/7672/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/7672/
Directeurs de recherche: Advisors:	Michel C. Desmarais
Programme: Program:	Unspecified

UNIVERSITÉ DE MONTRÉAL

THE ASSESSMENT OF USER KNOWLEDGE WITH A BAYESIAN FRAMEWORK AND ITS COMPARISON WITH ITEM RESPONSE THEORY

XIAOMING PU
DÉPARTEMENT DE GÉNIE INFORMATIQUE
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÈES
(GÉNIE INFORMATIQUE)
NOVEMBRE 2005

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-16842-4
Our file Notre référence
ISBN: 978-0-494-16842-4

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé:

THE ASSESSMENT OF USER KNOWLEDGE WITH A BAYESIAN FRAMEWORK AND ITS COMPARISON WITH ITEM RESPONSE THEORY

présenté par: PU Xiaoming

en vue de l'obtention du diplôme de: Maîtrise ès sciences appliquées

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

M. FERNANDEZ José, Ph.D., président

M. DESMARAIS Michel, Ph.D., membre et directeur de recherche

M. BILODEAU Guillaume-Alexandre, Ph.D., membre

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank, first and foremost, my thesis advisor, Dr. Michel C. Desmarais. His advice, both as a professor and as a friend, were and always will be invaluable. Moreover, I would like to thank the entire *génie informatique* department faculty and staff for all the support and encouragement (and toleration) they have shown me throughout the years.

Thanks also goes to all those who offered me valuable supports and assistances, Peyman Meshkinfam, Alejandro Villarreal Morales, Tamer Rafla, Rafiou Oketoukoun, Shunkai Fu and Lei Ma, and to Yuan Yao for her constant love and support.

This study has been supported by the National Research Council of Canada.

Thank you,
Xiaoming Pu

ABSTRACT

Intelligent learning environments require fine grained assessment of the user's knowledge state. Models of the users knowledge are necessary for applications such as student study guides (Khuwaja, Desmarais, & Cheng, 1996a), adaptive hypertext and intelligent textbooks (Schwarz, Brusilovsky, & Weber, 1996), or intelligent tutoring systems (Mayo & Mitrovic, 2001; VanLehn, Lynch, Schulze, Shapiro, Shelby, Taylor, Treacy, Weinstein, & Wintersgill, 2005).

A large body of research has been devoted to building models for assessing knowledge efficiently. The psychometric field is the earliest to tackle this problem. In particular, the Item Response Theory (IRT) is a classic approach to skill modeling introduced over four decades ago. It has been applied to what is probably the earliest computer adaptive interface: Computer Adaptive Testing (CAT). In the last decade, the user modeling and intelligent learning environment fields have cast interest in the use of graphical probabilistics models and Bayesian networks to address skill assessment. These techniques offer the advantage of producing very fine grained assessment not typically available with psychometric techniques such as IRT. However, probabilistic graph models can be more complex to build and calibrate than IRT.

A probabilistic network approach, named POKS (Desmarais, Maluf, & Liu, 1996; Desmarais & Pu, 2005a, 2005b; Desmarais, Fu, & Pu, 2005), was developed to provide a fine grained assessment while providing a simpler framework than Bayesian network (BN). Not only is the computational complexity simpler, it also allows the model calibration with fewer data cases. However, it makes strong assumptions which can be violated and lead to inaccuracies and it remains an empirical question to determine the extent of the potential assumptions violation on its performance.

This study is part of a wider research program to better assess the strength and weaknesses of POKS in comparison with other approaches. It focuses on the comparison of POKS with IRT. The comparison was performed over two simulation studies and within the CAT framework, since this is what IRT was designed for. The first simulation is based on a 34 items test on the knowledge of the UNIX shell commands and the second simulation is based on a 160 items French language test. In both cases, the simulation consists in choosing the most informative item based on the Fisher information and the information gain criteria, and feeding the actual answer to the knowledge assessment technique. The result is then compared with the actual answers and the process is repeated from the first to the last test item and for each subject.

Experimental results show that both approaches can classify examinees as master or non master effectively and efficiently, with relatively comparable performance. However, more significant differences are found for a second task that consists in predicting individual question item outcome. Implications of these results for adaptive testing and student modeling are discussed, as well as the limitations and advantages of POKS, namely the issue of integrating concepts into its structure.

As intelligent learning environments evolve and become more popular, knowledge assessment models and techniques will become more pervasive.

Keywords: Student modeling, CAT, Bayesian modeling, POKS, IRT

CONDENSÉ EN FRANÇAIS

Introduction

Les environnements d'apprentissage intelligents visent à adapter le contenu présenté en fonction des besoins individuels de l'étudiant et de son niveau actuel de connaissances.

Un modèle étudiant est un composant essentiel de ces applications. Le modèle étudiant établit et maintient une information à jour, notamment ses intérêts, ses buts et, bien entendu, son niveau de connaissance, auquel nous référerons par ses *compétences.* Les informations fournies par le modèle d'étudiant sont employées pour guider les systèmes d'apprentissage et répondre d'une manière adaptative.

Les environnements intelligents d'apprentissage exigent souvent du modèle étudiant un niveau détaillé de l'état de la connaissance. Il importe donc de développer des techniques permettant d'établir un tel diagnostic de façon rapide et fiable.

Les tests adaptatifs (TA, ou "Computer Adaptive Testing" en anglais) sont probablement les premiers exemples d'environnements adaptatifs d'apprentissage. Le principe derrière un TA est d'ajuster les questions à la connaissance du candidat en tenant compte des réponses aux questions précédentes. Il s'agit d'une boucle qui, en la simplifiant à deux étapes, consiste à évaluer la connaissance de l'étudiant puis à choisir et à présenter la question la plus appropriée en fonction du succès ou de l'échec à cet item, boucle que l'on nomme souvent la boucle TA.

La théorie originale derrière le TA est celle de la réponse aux items (TRI), formalisée il y a déjà quatre décennies. Plus récemment, différentes approaches bayésiennes ont été également appliquées pour modéliser les compétences d'un candidat à partir de ses réponses à des items.

Cette étude se concentre sur le lien entre les deux champs, d'un côté le TRI et de l'autre les techniques de modélisation bayésienne de l'étudiant, le dernier étant un des principales techniques utilisées pour créer des modèles utilisateurs. Nous décrivons chaque approche et effectuons une évaluation comparative des performances entre la TRI et une approche de modélisation bayésienne nommée POKS (Desmarais et al., 1996).

Concepts de base

La TRI est l'approche la plus répandue pour effectuer des tests adaptatifs. Dans le contexte du TA, la TRI modélise le lien entre les compétences du candidat et la probabilité de succès à une réponse donnée. Ce lien correspond à ce que l'on nomme la courbe caractéristique de l'article (ICC).

L'un des modèles d'ICC les plus largement répandus est le modèle logistique à deux-paramètre (2-PL). Ce modèle possède des propriétés mathématiques fort utiles d'un point de vue pratique. Une fois les paramètres de la courbe ICC déterminés pour tous les items d'un test, il devient possible de dériver le niveau le plus probable des compétences du candidat à partir d'un ensemble donné de réponses à des items. Typiquement, les compétences sont estimées avec une méthode de maximum de vraisemblance, mais un certain nombre de méthodes ont été proposées et étudiées pour cette fin.

D'autre part, la modélisation bayésienne fournit un cadre mathématique alternatif par lequel nous pouvons calculer la probabilité d'un certain événement étant donné l'occurrence d'un ensemble d'un ou plusieurs autres événements. La méthode la plus directe est fondées sur la table complète des probabilité conditionnelle conjointes, mais elle s'avère impraticable dans la plupart des cas à cause du très grand nombre de données nécessaires pour calibrer de telles tables. Il existe un certain nombre de techniques pour contourner ce problème. Ces moyens varient selon leurs hypothèses ou selon les contraintes qu'ils imposent aux probabilités conditionnelles dans un modèle.

Les modèles de graphes bayésiens et, en particulier, le cadre des réseaux bayésiens (BN), sont parmi les approches les plus répandues pour la modélisation bayésienne. Ils permettent de modéliser uniquement des probabilités conditionnelles appropriées et reposent sur un cadre mathématique solide pour la mise à jour les probabilités basée sur l'occurrence d'un événement dans le réseau. En outre, l'identification de la structure des probabilités conditionnelles, la topologie du réseau elle-même, peut être dérivée des données empiriques. Nous révisons différentes techniques de graphes bayésiennes dans cette étude, notamment l'approche POKS.

Théorie de la réponse aux items (TRI)

La théorie de réponse aux items est une approche classique pour le modèle de l'étudiant. Elle a été employée dans des applications de TA depuis plusieurs décennies.

La courbe caractéristique d'un item (*Item Characteristic Curve*, ICC) décrit le rapport entre la chance de succès d'un candidat à un item de test donné et son niveau de compétences. Deux familles des fonctions mathématiques sont généralement employées pour qualifier l'ICC : le modèle normal (*normal ogive model*) et le modèle logistique.

Le modèle normal est basé sur l'observation empirique que la distribution des compétences d'individus suit une une courbe normale. Beaucoup de chercheurs ont justifié l'utilisation du modèle normal d'ICC sur cette base pragmatique et la pratique au cours des années démontre la justesse de celui-ci. Le modèle normal original prend deux paramètres, a et b, qui sont respectivement le facteur de discrimination de l'item et son degré de difficulté.

La fonction logistique est une très proche approximation de la distribution normale. En outre, elle possède des avantages mathématiques qui facilitent grandement le calcul par rapport au modèle normal lors de l'estimation de la compétence d'un individu. Par conséquent, le modèle logistique de l'ICC est maintenant plus souvent employé dans la pratique. Le modèle logistique et le modèle normal partagent les mêmes paramètres et leurs valeurs sont interchangeables, moyennant un facteur multiplicatif. En plus des paramètres a et b, un troisième paramètre représentant la chance, c, est utilisé et se nomme le modèle logistique à trois paramètres (3-PL).

Par définition, des modèles d'ICC (normal et logistique) stipulent que la probabilité de succès à un item donné est indépendante de son succès aux autres items étant donné son niveau de compétence. Ce principe peut être énoncé comme une indépendance locale entre les items individuels dans un test. C'est ce que l'on nomme l'hypothèse de l'indépendance locale dans la TRI et qui stipule l'indépendance conditionnelle entre les items.

Sous l'hypothèse de l'indépendance locale, la probabilité d'un ensemble de réponses à des items étant donné un niveau donné de compétence se définit comme un simple produit des probabilités conditionnelles individuelles. Par l'intermédiaire de la technique de l'évaluation de maximum de la vraisemblance (*Maximum Likelihood Estimation*, MLE), il est alors possible de trouver la compétence du candidat ou les paramètres a et b les plus probables en fonction d'un échantillon de réponses aux items. Du point de vue mathématique, l'évaluation de maximum de vraisemblance correspond à un processus de calcul de racines. Les approches basées sur la méthode de Newton-Raphson sont habituellement des solutions applicables dans ce contexte. De la perspective du modèle d'étudiant, l'évaluation des paramètres des items constitue la solution à la création du modèle.

Une implantation du modèle TRI a été réalisé pour la présente étude basé sur le modèle logistique à deux paramètres (discrimination et difficulté).

Ordres partiels de connaissances (Partial Order Knowledge Structure, POKS)

L'ordre partiel de la connaissance (Partial Order Knowledge Structure, POKS, Desmarais et al, 1996) est une approche de modélisation bayésienne qui repose sur plusieurs hypothèses fortes pour réduire la complexité dans la modélisation bayésienne. L'approche POKS permet la modélisation bayésienne de la structure de la connaissance en créant des liens entre les items d'un test, en accord avec la théorie de l'espace de la connaissance (*Knowledge Spaces*, Falmagne et al., 1990). Un des buts de cette étude est d'explorer la validité de la méthode POKS pour modéliser et prédire la compétence malgré la simplicité du modèle bayésien utilisé.

Le réseau POKS ne contient que des items et aucun noeud concepts n'est inclus. Imposer cette règle facilite l'inférence de la structure POKS et élimine tout effort d'ingénierie humaine pour construire le réseau. Le recours à l'apprentissage à partir de données rend ainsi l'approche plus comparable à la TRI par rapport aux autres approches de modélisation bayésiennes qui exigeraient une étape d'ingénierie de la connaissance. Les mêmes données peuvent être employées pour les approches de la POKS et la TRI sans aucune manipulation ou transformation, permettant de ce fait une comparaison des deux approches sur une base égale.

L'algorithme d'induction du réseau POKS se fonde sur une analyse par paire des items entre eux. Une telle analyse permet d'identifier l'ordre dans lequel les personnes maîtrisent les items de connaissance. Elle est inspirée de la théorie des espaces de la connaissance qui déclare que l'ordre d'acquisition de compétence peut être modélisé par un graphe ET/OU. Pour notre fin, nous imposons une plus forte hypothèse dans lequel l'ordre d'acquisition de compétence peut être modélisé par un graphique acyclique dirigé (*Directed Acyclic Graph*, DAG), ou "ordre partiel". Cette hypothèse nous permet de limiter l'algorithme de l'induction du réseau POKS à l'analyse par paire uniquement. Il s'agit ici d'une autre forme d'indépendance locale présumée entre les items de connaissance.

Étant donné de l'hypothèse d'indépendance de POKS, la mise à jour de probabilité à partir de l'observation d'une nouvelle évidence (une réponse à un item) peut être réalisée par un calcul de probabilité postérieures simple. Ceci permet POKS d'employer le théorème de Bayes et nous l'utilisons dans sa forme de ratio de chance (*Odds ratio*).

L'induction de réseau POKS repose sur trois tests paramétriques entre deux noeuds (conformément au traitement par paire). Ainsi, pour tester la relation $A \rightarrow B$, les deux premiers tests vérifient la force des probabilités conditionnelles $P(B|A)$ et $P(\neg A|\neg B)$ en utilisant une distribution binomiale, tandis qu'un troisième test vérifie le degré d'indépendance entre A et B par un test χ^2.

Une fois le réseau créé, la probabilité de chaque noeud représente le modèle de l'étudiant et elle est mise à jour à chaque nouvelle évidence observée.

Choix de l'item

Dans le TA, le choix de l'item à présenter est une étape commune à la TRI comme à POKS. Dans le contexte du TA, le choix de l'item consiste à identifier le niveau des compétences du candidat avec le maximum de précision et en utilisant le moindre nombre d'items ou, en d'autres termes, de choisir l'item le plus informatif à chaque fois selon l'état de connaissance.

Notre étude présente deux méthodes de choix d'item qui sont employées dans les simulations réalisées. Il s'agit de l'approche de l'information de Fisher et l'approche du gain d'information.

La fonction de l'information de Fisher est une méthode très connue et utilisée en statistique. Pour la TRI, elle correspond plus ou moins à choisir un item dont le point d'inflexion de l'ICC est près de la compétence évaluée.

Quant au gain de l'information, cette méthode vise à diminuer l'entropie globale du test. L'item favorisé par l'approche du gain de l'information est celui qui maximise la réduction prévue de l'entropie.

Simulation et résultats

Nous comparons l'approche POKS à l'approche de la TRI avec deux paramètres (2-PL). Les résultats de simulations du processus de TA et la performance respective des deux approches pour prédire le niveau de compétence sont rapportés.

Les simulations sont effectuées sur deux ensembles de données pour les deux approches : un test de 34 items portant sur les commandes UNIX administré à 48 individus, et un test de 160 items portant sur la langue française administré à 41 candidats. Chaque item peut prendre deux valeurs, réussi ou non.

Les simulations avec POKS et avec la TRI-2PL sont toutes deux réalisées avec l'information de Fisher et la technique de réduction d'entropie. Les résultats sont comparés aux réponses réelles selon deux métriques: (1) la capacité de classifier correctement le répondant selon qu'il réussit ou non le test (avec des seuils de 50%, 60% et 70%) et (2) la capacité de prédire la réussite pour chaque question prise individuellement.

Les résultats expérimentaux démontrent que les deux approches peuvent classifier effectivement et efficacement des candidats, avec une performance relativement comparable. Cependant, des différences plus significatives sont trouvées pour la prévision des réussites aux questions individuelles.

Ce résultat démontre que POKS réussit, malgré la simplicité de son cadre bayésien, à performer au même niveau qu'une approche reconnue comme la TRI. Sa meilleure performance pour la prédiction des réponses individuelles n'est pas surprenante compte tenu que la TRI n'a pas la prétention de prédire à ce niveau de détail, mais démontre néanmoins le potentiel de POKS pour effectuer un diagnostic détaillé de la connaissance qui n'est pas possible avec la TRI à moins d'utiliser une approche plus complexe et multidimensionnelle.

Discussion et conclusion

Étant donné que l'approche POKS offre un potentiel intéressant pour le diagnostic détaillé du niveau de connaissance et qu'elle est, du point de vue computationel, beaucoup plus simple que l'approche de la TRI, on peut conclure que POKS constitue un candidat intéressant pour les environnements d'apprentissages adaptatifs et intelligents.

TABLE DES MATIÈRES

LIST OF TABLES

LIST OF FIGURES

LIST OF NOTATIONS AND SYMBOLS

θ	Ability level
$\hat{\theta}$	Estimate of ability level
u_i	Dichotomous response to item i in IRT ($u_i = 1$ is a success, otherwise $u_i = 0$)
X_i	Dichotomous response to item i in POKS ($x_i = 1$ is a success, otherwise $X_i = 0$)
$P(u_i = 1 \| \theta)$	Probability of correct response to item i in IRT
$P(X_i = 1)$	Probability of correct response to item i in POKS
$\boldsymbol{u} = [u_1, u_2, \ldots, u_n]$	Response vector of an examinee.
$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{u_1} \\ \boldsymbol{u_2} \\ \vdots \\ \vdots \\ \boldsymbol{u_N} \end{bmatrix}$	Response matrix for examinees $1, 2, \ldots, N$
$L_i(u_i \| \theta)$	Likelihood function for observed response u_i
$L(u_1, u_2, \ldots, u_n \| \theta)$	Likelihood function for observed response $u_1, u_2, \ldots, u_n$
$I_i(\theta)$	Item information function of item i
$I(\theta) = \prod_{i}^{n} I_i(\theta)$	Test information function

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Intelligent Learning Environments

Intelligent learning environments are meant to adapt to the needs and knowledge of individual students. In order to allow learning instruction to be individually designed, intelligent learning environments first capture the student's understanding of the subject, then use this information to determine the difficulty of material and any necessary remediation that best fit his current needs or knowledge level[1]. For example, this could imply adapting a tutor's didactic content and strategy, or adapting hyperlinks of a documentation system, or some query retrieval result, etc. Example of intelligent learning environments are:

- student study guides (Khuwaja, Desmarais, & Cheng, 1996b)
- adaptive hypertext and intelligent textbooks (Schwarz et al., 1996)
- intelligent tutoring systems (Mayo & Mitrovic, 2001; VanLehn et al., 2005)
- web based adaptive hyper-textbook, hyper-media and course-ware (Brusilovsky, Schwarz, & Weber, 1996; Brusilovsky, Eklund, & Schwarz, 1997, 1998)

In summary, we use the words "intelligent learning environment" for those adaptive learning environments that are capable of tailoring the learning instructions and materials, or learning progress upon individual student's need and current knowledge level.

[1] Only masculine forms are used in this study, e.g. he/his. However they are intended to, and not limited to both masculine and feminine forms.

1.2 Student Modeling

Student model represents the learning system's belief about the learner's knowledge. The student model is an essential component in intelligent learning environments that are responsive to individual student's needs and profiles. Student model builds and maintains the system's understanding of the student. The information provided by student model are used to guide the learning systems to respond adaptively.

Intelligent learning environments often require fine grained assessment of the students's knowledge state. As the needs of assessing students in terms of mastery level with respect to one or more knowledge units grow rapidly in intelligent learning environment, the importance of fast and reliable diagnostic assessment becomes a key issue.

There are many techniques for generating student models. Most of them are computationally complex and expensive, for example, Item Response Theory, and Bayesian networks.

1.3 Computer Adaptive Testing (CAT)

Computer adaptive testing (CAT) applications are probably the earliest examples of the use of intelligent user modeling techniques and adaptive learning environments (see Eggen, 2004). For example, Graduate Record Examination (GRE) is a standardized test administered through CAT[2].

The principle behind CAT is to adjust test questions to the examinee's knowledge by taking into account how examinee answered previous questions. Taking the same CAT, a low-ability examinee and a high-ability examinee will see quite different sets of questions: the low-ability examinee will mainly see relatively easy questions, and the

[2] GRE is the test taken in order to get into graduate school in the United States. It is administered by the Educational Testing Service (ETS) `http://www.ets.org`

high-ability examinee will see more difficult questions. Both individuals may answer the same percentage of questions correctly, but because the high-ability examinee can answer more difficult questions correctly, he or she will get a higher score.

Computer adaptive tests are usually carried out in an iterative fashion which is often called CAT loop. The CAT loop process generally consists of the steps below,

1. Estimate an examinee's most likely *score* from his observed responses to previously presented questions (if no responses history currently available, take the population's average *score* as the initial estimate).

2. Stop if some termination conditions are met, otherwise continue to *Step 3.*

3. Find the most informative item according to the estimated *score.*

4. Present the question to the student and record his answer to this question, then return to *Step 1.*

This CAT loop process continues with the test gradually locating the person's competence level. The *score* that serves as an estimate of competence gets more accurate with each question given. The test ends when the accuracy of that estimate reaches a statistically acceptable level (or when a maximum number of items has been presented).

1.4 Comparison of Two Approaches in CAT

The original theory behind CAT is the Item Response Theory (IRT), a framework introduced by Birnbaum (1968) and Lord and Novick (1968), and refined by a number of other researchers since its introduction (see van der Linden & Hambleton, 1997; Hambleton, Swaminathan, & Rogers, 1991). More recently, the Bayesian modeling approach has also been applied to model an examinee's ability based on test item responses. This interest in Bayesian modeling has come not only from researchers in educational testing, such as Rudner (2002) and Mislevy and Gitomer (1995), but also from researchers in adaptive interfaces and user modeling (see, for example Conati, Gertner, & VanLehn, 2002).

This study focuses on the link between the two fields, namely IRT and the Bayesian student modeling techniques, which is one of the major probabilistic user modeling techniques. We compare each approach and conduct a comparative performance evaluation between IRT and one such Bayesian modeling approach named POKS.

POKS (Desmarais et al., 1996) is a specific Bayesian modeling approach particularly well suited for the comparison with IRT approach. Because, akin to the IRT approach, it does not necessarily require a knowledge engineering effort to build the model but, instead, relies on statistical techniques to build and calibrate the model. Indeed, by relying solely on observable nodes to build a graph model of item-to-item relations, there is no need to define latent skills behind each test item. The process then becomes very similar to IRT for which no knowledge engineering effort is required as a single latent skill (hidden node) is assumed for every test item.

Section 1.4 provides the basics of the IRT and Bayesian modeling approaches. It is followed by more detailed descriptions of the specific IRT and POKS techniques compared in this study (see Section 2.3.4 and Section 3.6).

CHAPTER 2

OVERVIEW OF BASIC CONCEPTS

This section provides a general overview of basic background theories and concepts of two student modeling approaches for the comparison of IRT and Bayesian modeling in CAT. A more detailed and technical review of the specific frameworks compared in this study is given in the following sections.

2.1 Computer Adaptive Testing and Item Response Theory

The prevalent means of conducting computer adaptive testing (CAT) is based on the Item Response Theory (IRT). IRT establishes a conceptual model which bridges students' ability with the chances of success in test questions. Such underlying ability of interest is referred as *latent trait* in IRT literature, (e.g. knowledge, skill, ability, etc.), therefore Item Response Theory is also known as Latent Trait Theory.

In the testing context,

- *Examinee* is the synonym of test taker.
- *Item* is the test question presented to an examinee.
- *Item response* is the answer received from examinee to given test question (item).

  Two types of item response can be obtained from items by definition.

  - *Dichotomous* item responses (also called binary responses) are those obtained from: (a) items that are scored correct or incorrect in achievement tests (e.g. multiple choice); or (b) items that are dichotomously scored according to a scoring key in an attitude, or personality scale (e.g. true/false, agree/disagree).

- *Polytomous* items response (also called graded response, Likert, Likert-type, or ordinal item responses) are those involve more than two scoring options such as a five-point strongly agree to strongly disagree scale on a personality or attitude measure.

This study only considers the case of dichotomous item response. In IRT, item response to item i is usually denoted as u_i, and $\boldsymbol{u}$ for the response vector to items $i = 1, 2, \dots, n$ [3].

- *Probability of correct response*

One usually thinks of an individual either getting an item right or wrong, however it is more useful to think in the percentage (frequency) of success when we have to deal with the test subjects and items repeatedly. Probability of correct response can be elaborated in two hypothetical scenarios.

First, suppose we have the universe of people with some common ability level take a given item. Some of those people will get this item right, others will get it wrong. Thus, the probability of correct response is the value of average percentage of correct response when the number of people approaches infinite (see Lord & Novick, 1968).

In the second scenario, suppose we have a set of items of the same intrinsic characteristics (see Section 3.1), and we assume an examinee's responses to those items will be independent of each other (e.g. no interference between

[3] The notation of item response in POKS is in preference to X_i.

items, no skills learned from test, no fatigue). Similarly, this examinee will get some items right and some others wrong. Likewise, the probability of correct response is the value of average percentage of answering correctly when the number of items goes infinite.

Note that the above two distinct scenarios only help to explain the probability of correct response intuitively. Therefore, there is no difference in how one interpreting it in practical use.

- *Ability* is the "score" on the scale of latent trait in IRT. *Proficiency* is another word interchangeable with *ability* in IRT. They both refer to those unobservable latent traits which determine the item response in the test. Given an item, the ability is the only factor that will account for the probability of correct response (see below). Ability is denoted as θ in IRT.

The classical IRT models assume that a single ability level accounts for the examinee's performance (correct/wrong answers to the presented questions). IRT models the relationship between examinee's ability and his probability of correct response into a theoretical function named item characteristic curve (ICC). The ICC functions normally have an "S" shaped curve which implies that the higher the ability level one examinee poses (e.g. more knowledged) the higher chance he will succeed in that question. Each test item can have its own ICC item parameters. Several types of ICC exist. One of the most widely used ICC is the two-parameter logistic model (2-PL), it has desirable mathematical properties for practical use.

The shape of 2-PL ICC is modeled by two parameters: the item's *difficulty* level and the item's *discrimination* power. Figure 2.1 illustrates the typical 2-PL ICC curve corresponding to an item of difficulty $b = 0$ (average difficulty) and discrimination $a = 1$. These two parameters, difficulty and discrimination, can be estimated from data of each test item. Typically, parameters are estimated by a maximum likelihood approaches (Baker, 1992).

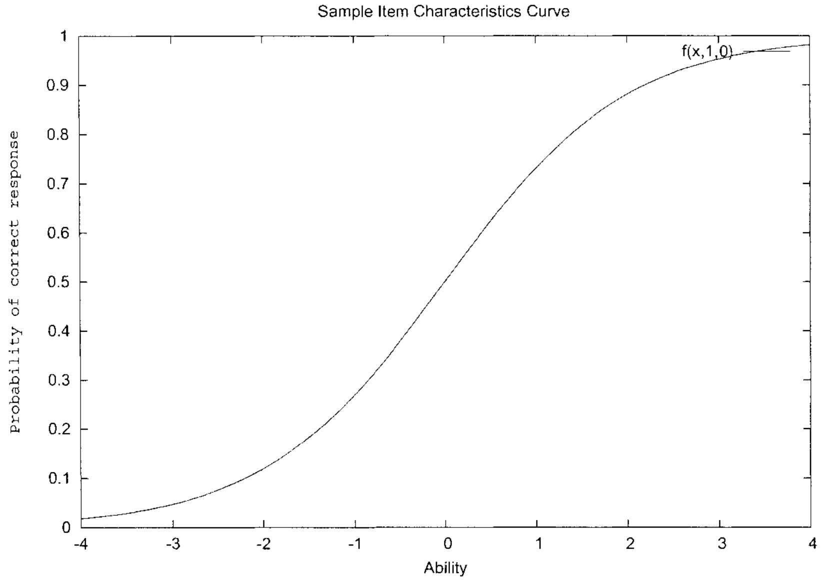

Figure 2.1: A typical ICC curve, with item parameter $a = 1, b = 0$

Once the parameters of the ICC curve are determined for all test items, it becomes possible to derive the examinee's most likely ability level from a given set of item responses. Typically, the ability is estimated with a maximum likelihood model, but a number of methods have been proposed and investigated for this task (Baker, 1992). Section 2.3.4 provides more details on the IRT ability estimation.

2.2 Bayesian Modeling Approaches to CAT and Student Modeling

We will return to the IRT approach in Section 3.6 to provide the details on the specific algorithms used in this study. Let us now turn to the Bayesian approach to student modeling and describe how this approach is applied to CAT.

2.2.1 Bayesian modeling and Bayesian networks

Bayesian modeling provides a mathematical framework by which we can compute the probability of a certain event given the occurrence of a set of one or more events. For example, in CAT, one could compute the conditional probability of mastery of a test item given the previous responses by using samples where such a response pattern was found. This simple but impractical approach relies on the full joint conditional probability table. The problem with this straightforward approach is, obviously, that the number of conditional probabilities grows exponentially with the number of items. The approach quickly becomes impractical because of limited data. For example, computing the probability of correctly answering a specific test item given the answers to the last 10 items would entail constructing a conditional probability table of 2^{10} entries, if we assume each item can take two values, $\{success, failure\}$. A reliable empirical estimate of this set of conditional probabilities would require thousands of data cases, whereas a subjective estimate is deemed too tedious and unreliable.

There exist a number of means to avoid relying on the full joint conditional probability distribution to perform Bayesian inference. These means will vary according to their assumptions, or according to the constraints they impose on the conditional probabilities in a model.

The Bayesian graph models, and in particular the Bayesian networks (BN) framework, are amongst the most prevalent approaches to Bayesian modeling. They allows the modeling of only the relevant conditional probabilities and they can rest on a sound mathematical scheme to update the probabilities upon the occurrence of an event in the network (see Heckerman, 1995). Furthermore, the identification of the relevant subset of conditional probabilities, the topology of the network itself, can be derived from empirical data (see Heckerman, 1995; Cheng, Greiner, Kelly, Bell, & Liu, 2002; Liu & Desmarais, 1997). A simple example of such BN can be found in Vomlel (2004a, 2004b, 2002).

To reduce the number of conditional probabilities to only the relevant ones while maintaining consistent probability updates from new evidence, BN structures define clear semantics of conditional probabilities and independence relations between nodes in the network. It states that the probability of a node X_i, given the evidence from the nodes' parents $pa(X_i)$, is independent of all nodes, except its descendants. Assuming that the vector $X_1, \ldots, X_i$ represents a specific combination of responses to test items and concepts mastery for a given individual, it follows from the above definition of a BN that the probability of this vector is:

$$P(X_1, \ldots, X_k) = \prod_{i=1}^{k} (X_i | pa(X_i)) \tag{2.1}$$

where $pa(X_i)$ represents the set of parent nodes of X_i in the BN.

For CAT and student modeling, the application of BN and graph models generally consists in modeling the conditional probabilities as a hierarchy of concepts with items as leaf nodes. Figure 2.2 illustrates a very simple graph structure that, in fact, represents an IRT model. It contains a unique concept node, θ, and a set of item nodes, $X_1, \ldots, X_n$. The semantics of this networks would state, for example, that the probability of node X_1 is independent of the probability of node X_2 given the ability θ. This definition translates into:

$$P(X_1 | \theta, X_2) = P(X_1 | \theta) P(X_2) \tag{2.2}$$

However, the probability that skill θ is mastered depends on the responses to all item nodes. We return with more details on the IRT model in section Section 2.3.4.

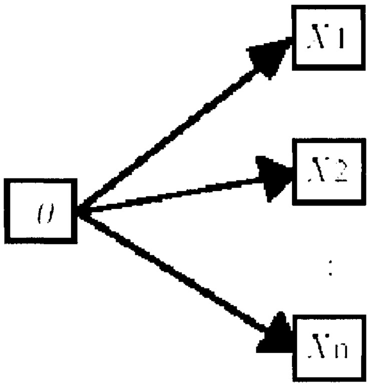

Figure 2.2: BN structure of an IRT model, where θ is the examinee's ability and $\{X_1, X_2, \ldots, X_n\}$ are the test items.

One of the major advantages of graph models over IRT is that the assessment of the probability of mastery to a test item does not rely on a single trait, namely the examinee's ability level. High level concepts embedded in a graph model constitute a powerful means of representing a variety of ability dimensions. For example, Figure 2.2 can be augmented by defining multiple θ over a set of test items, which, in turn, can be organized as a hierarchy or as a directed graph structure with high level θ representing global skills. Moreover, misconceptions can also be included in the structure.

The flexibility and representational power of graph models and their derivatives have been recognized and applied to student modeling by a number of researchers in the last decade, for example, Reye (2004; Conati et al., 2002; Jameson, 1995; Millán, Trella, Pérez-de-la-Cruz, & Conejo, 2000; Mislevy & Gitomer, 1995; Desmarais et al., 1996; Martin & Vanlehn, 1995; Zapata-Rivera & Greer, 2004). They have also been applied more specifically to CAT systems (Vomlel, 2004b; Millán & Pérez-de-la-Cruz, 2002; Collins, Greer, & Huang, 1996; VanLehn & Martin, 1997). We will review some of this work in the remainder of this section. The specific Bayesian modeling approach used in the current study will be further described in Section 3.6.

2.2.2 Student graph models

Vanlehn, Martin, Conati and a number of collaborators were amongst the most early and active users of BN for student assessment (Martin & Vanlehn, 1995). In the latest of a series of three tutors embedding a BN, *Andes Tutor* (Conati et al., 2002; VanLehn & Niu, 2001) incorporates a BN composed of a number of different types of nodes (rules, context-rules, fact, goal nodes). Each node can take a value of "mastered" or "non-mastered" with a given probability. Probabilities can be computed from Bayes posterior probability rule, or in a deterministic binary form (e.g. $P(X = 1) \rightarrow P(Y = 1)$), or in logical "and" and "or" relations with an arbitrary amount of a noise factor that makes these relations non-deterministic. These relations are named leaky-or and noisy-and (see Neapolitan, 1998). Most conditional probabilities in the network are subjectively assessed.

In *Hydrive*, Mislevy and Gitomer (1995) used a BN for assessing a student's competence at troubleshooting an aircraft hydraulics system. The BN is also engineered through careful modeling of the domain knowledge in a hierarchy of abilities. Node categories are not necessarily binary, as each node has its own set of values such as $\{weak, strong\}$ or $\{expert, good, ok, weak\}$. Conditional probabilities are first posited by expert judgment and further refined with a data set of 40 subjects.

The work of Collins (1996) is amongst the first to create a CAT with a Bayesian network. They use the notion of granularity hierarchies to define the BN. Granularity hierarchies are essentially aggregations of concepts or skills into a hierarchy, akin to Mislevy and Gitomer (1995) where the leaves are test items and the root represents the whole subject domain. The BN tested are knowledge engineered from expert knowledge and different topologies are compared. Since the system is a CAT, the choice of the next item is adapted to optimize ability assessment. It is based on a utility measure that yields the item with highest discrimination factor, that is, the item whose difference in ability estimate, is the highest between a correct and an incorrect answer.

In his unpublished Master thesis, Collins (1996) compares a BN model with an IRT model. He calibrates a BN model with conditional probabilities obtained from an extensive pool of 6000 data cases for a test of 440 items, which in fact corresponds to 44 items replicated 10 times to simulate a large test[4]. Comparison of the BN approach with an IRT model revealed that, after approximately 20 items, the BN approach is more effective in classifying examinees as master or non-master than the two IRT-based algorithms they compared it with, namely *EXPSRT* and *EXPSRT-2* Collins (1996). However, it is not clear what impact the replication of the original 44 items can have on these results and how much this could favor one approach over the other. For example, the non adaptive paper and pencil test proved more accurate than the IRT and BN approaches, which is unexpected and could be explained by this replication[5].

In a more recent CAT system, Millán and Pérez-de-la-Cruz (2002) defined a hierarchical BN with three layers: concepts, topics, and subjects. A fourth layer links test items to concepts. They used different means of computing the updated probabilities according to the layer. The concepts, topics, and subjects layers use a summation formula to yield an updated probability. New probabilities are a function of weights assigned to evidence nodes according to their importance, which can be a factor of

[4] Note that the BN only contained the 44 original nodes, not 440.

[5] The POKS approach used in the current study would be highly influenced by the replication of items. Replicated items would be aggregated into fully connected nodes, in effect merging them into the equivalent of a single node.

time devoted to a certain topic in a course, for example. At the items level, the probabilities that a concept is mastered is a function of test items. That probability is computed from a conditional probability with parameters modeled from adopted ICC function such as the one in IRT. They tested the accuracy of their approach with simulated students and a test of 60 questions and found a relatively good performance for assessing mastery of each of 17 different concepts with error rates varying from 3% to 10%.

2.2.3 Learning graph models from data

In contrast with most Bayesian student model approaches, Vomlel (2004b) has conducted experiments with empirically derived BN. This work is, to our knowledge, the only experiment using empirical data to construct BN, although it does involve some knowledge engineering effort for categorizing concepts and test items into a hierarchy. Vomlel used *HUGIN* PC algorithm (Jensen, Kjæul, Lang, & Madsen, 2002) to calibrate a number of network topologies from 149 data cases of a 20 questions arithmetic tests administered to high school students. The basic topology of the network was constrained based on a knowledge engineering of the domain with experts, but *HUGIN* was used to refine or define parts of the BN's structure. The BN was composed of a few skills and student misconceptions. Some of the different BN structures tested incorporated hidden nodes that were created by *HUGIN*'s BN induction algorithm. Conditional probabilities were all calibrated from empirical data. The results show that an adaptive test with such BN can correctly identify the skills with an accuracy of approximately 90% after the 9th question and performs significantly better than a fixed question order test.

2.3 Considerations for Comparing IRT and Bayesian Modeling

When comparing IRT with Bayesian modeling, the question of how the model is built and calibrated (or learned) is a crucial one, as the two approaches differ significantly on that issue. IRT modeling is entirely based on calibration from data and has limited modeling flexibility, whereas Bayesian modeling offers much more flexibility but it involves knowledge engineering efforts that can also be limiting for many applications. These issues are central to the practical use of student modeling and we discuss them in more details in this section.

2.3.1 Automated approach considerations

The IRT models are empirically derived from test data and student expertise is solely defined by a set of observable test items, which usually take on two possible values: mastered or non-mastered[6]. IRT does not rely on any subjective assessment, nor on the ability of a domain expert knowledge engineer, as it requires no human intervention to build the model. The same can be said about POKS with item only node structures. Such algorithmic techniques, for which the model is learned or induced from empirical data, have important advantages that stem from their amenability to complete automation:

- It avoids the so called "domain expert bottleneck" and is thus more scalable.
- It is not subject to human biases and expert ability to build domain models.
- It lends itself to automated updates when new items are added to a test (a very common situation for qualification tests where items need to be regularly

[6] Besides mastered and non-mastered, a third category is often used, *undecided*, and any number of categories can be defined in theory.

renewed).

- It allows dynamic re-calibration of the model as new data is gathered.

The last two features are highly regarded by practitioners since test content is often subject to frequent updates which impose a strong burden for the maintenance of CAT test content.

2.3.2 Graph model considerations

What IRT lacks is the ability to make detailed cognitive assessment such as identifying specific concepts or misconceptions. In the original IRT, there is no provision for dividing the knowledge domain into different concepts that can be assessed individually, except by segmenting a large test into smaller ones, or by using what is known as multidimensional IRT (MIRT) models (Reckase, 1997; McDonald, 1997). But as we move towards MIRT, then some knowledge engineering effort is required to identify the dimensions and to classify items according to each of them. It becomes a graph model with multiple hidden nodes.

Our review of Bayesian student modeling revealed that the prevalent approach is to follow knowledge engineering techniques to build sophisticated graphical models with multiple levels of hidden nodes. Such models are often structured into a hierarchical decomposition of concepts into more and more specific skills, with items as leaf nodes. In some variants, misconceptions, multi-parents nodes, and sibling links can add yet more cognitive assessment and representation power to such structures. This is an essential feature of many intelligent learning environments that rely on fine grained student modeling.

However, this flexibility comes at the cost of modeling efforts to define the structure by domain experts, who must also be knowledgeable in Bayesian modeling. Beyond the structural definition, the problem of calibrating hidden nodes relations and nodes with multiple parent relations is paramount because of the lack of sufficient data cases (Jameson, 1995). Complex graph models often involve simplifications and approximations such as leaky-AND/OR gates (Martin & Vanlehn, 1995; Conati et al., 2002) and weighted means (Millán & Pérez-de-la-Cruz, 2002) thereby weakening the validity and accuracy of the model.

As a consequence of the above obstacles, complex graph models leave little room for automation and its benefits. Although recent developments has shown that small networks of a few tens of nodes can be reliably derived from empirical data of a few thousand cases (Cheng et al., 2002) this is still impractical in student modeling and the automated construction of a BN network remains a difficult problem that involves complex algorithms and considerable computing resources. In practice, heuristics and some degree of expert intervention are required for building a BN. With the exception of Vomlel (2004b), who has used the combination of a network topology induction system with knowledge engineered adjustments to the structure, Bayesian student models do not allow automated model building. When used, empirical data serves the sole purpose of calibrating conditional probabilities, and yet, many also use subjectively estimated parameters.

2.3.3 Item node structures

Item node structures are networks with links among item themselves, as opposed to hierarchical structures of concepts with items as leaf nodes. They are particularly subject to the difficulties of using Bayesian graph models because the number of nodes can be large (e.g. in the French language test used for this study, we have 160 item nodes) and their structure is not as apparent as when dealing with concepts. Nevertheless, the Theory of Knowledge Spaces (Falmagne, Koppen, Villano, Doignon, & Johannesen, 1990) states that items do have a structure and that it can be used for making knowledge assessment. But the obstacles to using a knowledge engineering approach and the amount of data required for the precise calibration of Bayesian networks makes such approach impractical.

We will see in Section 3.6 that POKS addresses these obstacles by reverting to binary relations, which allows calibration with small data sets, and using strong assumptions. That approach makes POKS amenable to algorithmic model construction and calibration. However, the issue of detailed cognitive assessment remains since concepts have to be included to provide fine grained assessment. We return to it in Section 6.9.3.

2.3.4 Item selection

There is a common concern to IRT and POKS in adaptive testing, i.e. how to choose the item that will be presented to the examinee based on his performances or profiles. The criteria of selecting such "optimal" item depend on the goals and achievements that an adaptive testing is meant for.

In CAT context, the goal of item selection can be generally expressed as, using the least number of items to identify the examinee's ability level with maximum precision, or in other words, choosing the most informative item.

Note that choosing the most informative item is only one of many alternative strategies. The choice could also be determined by other considerations, such as the need to randomize and diversify the items presented across examinees, or to adapt item difficulty to ability level. Moreover, the choice of the items administered could be outside the control of the system. For example, the system could be in a non-intrusive, observational mode, as it is often the case in advice giving interfaces.

However, in the context of the current study, we will follow the usual CAT goal of assessing ability with the least number of questions. There are numerous measures for finding the most informative items, such as

- the IRT item information function (Birnbaum, 1968), also known as Fisher information;
- the minimum information cost (Lewis & Sheehan, 1990);
- the information gain (entropy reduction), or finally;
- the relative entropy, also known as Kuller-Leiber distance (see Eggen, 1998).

The readers are referred to Rudner (2002) for a comparative studies of some of these measures. Section 4.6 discusses two item selection methods used in this study, Fisher information and information gain.

Summary

This section reviews the basic background theories and concepts of two student modeling approaches in this comparison study. IRT is the prevalent CAT approaches for student modeling, it has been widely used in the practical use over decades. Bayesian modeling is relative new in the field of CAT, however its advantage of being a graph model makes Bayesian network a promising tool for the fine grained knowledge assessment. The modeling considerations for comparing IRT and Bayesian modeling approaches outline the common concerns in model creation and calibration for these two approaches. However, a specific Bayesian modeling approach, POKS is able to accommodate the differences and put the comparison on the same footing. The detailed techniques of IRT and POKS are illustrated in Section 2.3.4 and Section 3.6 respectively.

CHAPTER 3

ITEM RESPONSE THEORY

Item Response Theory (IRT) is a classic approach to model students' skills. It has been used in CAT application over decades. This section describes IRT and its underlying approaches of estimating examinee's ability and item parameters.

The item characteristic curve (ICC) is the backbone of IRT which describes the relationship between an examinee's internal mental state (ability level) and his external behavior (the probability of correct response to presented questions). Two types of ICC model, normal ogive model and logistic model are examined in Section 3.1.

It is useful to elicit some further information regarding the overall test from individual test items' ICC. Section 3.2 explains the idea of local independence between individual item's ICC in a test. Under the assumption of local independence in IRT, Section 3.3 elaborates on the likelihood function for two categories of response pattern (responses received from a single examinee or all examinees). The likelihood functions provide the overall information regarding the responses pattern of items in a test.

Section 3.4 and Section 3.5 investigate the two building blocks of IRT framework, ability estimation and item parameter estimation procedures. Both procedures employ the technique of maximum likelihood estimation to likelihood functions.

Section 3.6 responds to a common concern in IRT and CAT, how precise the examinee's ability estimate is. By defining test information function, the amount of information obtained from administering a set of items or the whole test is given in substantial number for further comparison and analysis.

3.1 Item Characteristic Curve (ICC)

The item characteristic curve (ICC) reflects the intrinsic characteristics of a certain item, and gives the relationship between the probabilities of a correct response across a range of ability. Thus, modeling ICC concerns the determination of the function form and the parameters incorporated in the function.

Section 3.1.1 and Section 3.1.2 examine two ICC models and their corresponding mathematical properties, normal ogive model and logistic model respectively.

3.1.1 Normal ogive model

In Terman's (1916) extension work of Binet-Simon Intelligence Scale, he fitted ICC by graphical means in order to address the relationship between two variables, the proportion of correct responses from empirical data and a criterion variable which is considered to be an unobserved hypothetical variable (roughly equivalent to latent trait in IRT). While fitting smooth functions to the observed proportion of correct response can be done by graphical means, the resulting curves lack mathematical rigor. If appropriate mathematical functions could be found that both fit the observed data and have reasonable mathematical properties, the theoretical aspect of the ICC could be advanced. Terman (1916) has shown that the empirically obtained item characteristic curves have the appearance of a cumulative distribution function ('S' shaped). In addition, only two properties of the item characteristic curves, difficulty and discrimination factors, are needed to describe an item's technical characteristics.

Since the normal distribution (also called Gaussian distribution) is a keystone of statistical theory, it is not surprising that the normal ogive has been used as the ICC model. Richardson (1936), Ferguson (1942), and Finney (1944) have justified the use of the normal ogive as an ICC model on pragmatic ground. Therefore, in typical sets of item response data, the normal ogive has proved to be workable model (Baker, 1992).

Normal ogive function is defined as:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \tag{3.1}$$

where e is the natural logarithm constant.

Therefore, the two-parameter normal ogive ICC model is defined as:

$$\begin{aligned} P(\theta) = \Phi[L(\theta)] &= \int_{-\infty}^{L(\theta)} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \\ &= \int_{-\infty}^{a(\theta-b)} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \end{aligned} \tag{3.2}$$

where $L(\theta) = a(\theta - b)$, θ is ability level, b is the difficulty parameter, a is discrimination parameter.

3.1.2 Logistic model

A function which very nearly coincides with the normal ogive model, and which has advantages of mathematical convenience in several areas of application, is the logistic distribution function.

Logistic distribution function is defined as:

$$\Psi(x) = \frac{1}{1 + e^{-x}} \tag{3.3}$$

Haley (1952) has shown that, for all $-\infty < x < \infty$,

$$|\Phi(x) - \Psi(1.702x)| < 0.01 \tag{3.4}$$

which implies that the difference between normal ogive function (Equation 3.1) and logistic function (Equation 3.3) is very small, i.e. less than 0.01 for every set of item parameter values. Therefore, the two ICC models give very similar results for most practical work.

The two-parameter logistic (2-PL) ICC model is defined as

$$P(\theta) = \Psi[DL(\theta)] = \frac{1}{1+e^{-DL(\theta)}} = \frac{1}{1+e^{-Da(\theta-b)}} \tag{3.5}$$

where $L(\theta) = a(\theta - b)$, θ is ability level, a and b are item parameters whose roles are generally the same as those in the two-parameter normal ogive model (Equation 3.2), and $D \approx 1.7$ is a transforming scale factor between normal ogive and logistic model (see Equation 3.4).

Equation 3.3 has some favorable mathematical properties, such as the simple form of 1st and 2nd order derivatives, logistic ICC model (Equation 3.3) is often treated as a mathematically convenient, close approximation to the classical normal ogive model.

Under some circumstances, an examinee presenting extremely low ability level could answer an item correctly just by guessing. In order to model such impact, a third parameter *guessing* factor c is introduced to 2-PL model (Birnbaum, 1968).

The three-parameter logistic (3-PL) model is defined as

$$P(\theta) \;=\; c + (1-c)\frac{1}{1+e^{-Da(\theta-b)}} \tag{3.6}$$

where b is the difficulty parameter, a is discrimination parameter, c is guessing parameter, θ is ability level.

Sometimes one may wish to restrict the discriminant power of the items, therefore, the discrimination parameter a can be dropped from Equation 3.5.

The one-parameter logistic (1-PL) ICC, also called Rasch model, is defined as

$$P(\theta) \quad = \quad \frac{1}{1+e^{-(\theta-b)}} \tag{3.7}$$

which is exactly the 2-PL model with a discrimination parameter a fixed at unity (see Birnbaum, 1968).

In summary, 1-PL, 2-PL and 3-PL ICC models are

$$\begin{cases} P(\theta) = c + \dfrac{(1-c)}{1+e^{-Da(\theta-b)}} & \text{3-PL} \\ P(\theta) = \dfrac{1}{1+e^{-Da(\theta-b)}} & \text{2-PL} \\ P(\theta) = \dfrac{1}{1+e^{-(\theta-b)}} & \text{1-PL or Rasch model} \end{cases} \tag{3.8}$$

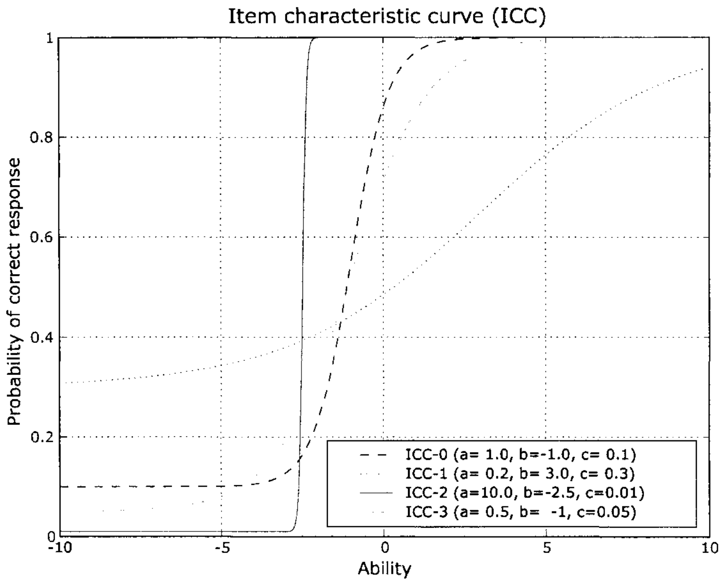

Figure 3.3: Logistic ICC curves

Figure 3.3 depicts several logistic item characteristic curves with various a, b, c values which jointly determine the shape of curves. Some general properties of logistic ICC models can be inferred from Figure 3.3

- Parameter c is the probability that a person completely lacking in ability ($\theta = -\infty$) will answer the item correctly. It is also called the lower asymptote. If an item cannot not be answered correctly by guessing then $c = 0$.

- Parameter b is a location parameter, it determines the position of the curve along the ability scale. The more difficult the item, the further the curve is to the right. The curve has its inflection point at $\theta = b$. When there is no guessing, b is the ability level where the probability of a correct answer is 0.5. When there is guessing, b is the ability level where the probability of a correct answer is $(1.0 - c)/2$.

- Parameter a is proportional to the slope of the curve at the inflection point. This slope is actually is $0.425a(1 - c)$. Thus a represents the discriminating power of an item, the degree to which item response varies with ability level.

One special case of interest is that an item could have perfect discrimination. The ICC of such an item is a vertical line at some point along the ability scale. A close approximation of ideal perfect discrimination is achieved by having relatively large a value, e.g. see curve with $a = 10.0$ in Figure 3.3.

3.2 Local Independence Assumption

Recall the normal ogive (Equation 3.2) and logistic (Equation 3.6) ICC models, they assert that the probability of success on an item depends on item parameters and examinee ability θ, and on nothing else. If the model is true, an examinee's ability

θ is the only one required to determine his probability of success on a specific item. In other words, if we know the examinee's ability, knowledge of his success or failure on any other items will change nothing to the probability of success to a given item. The principle just stated is Lazarsfeld's assumption (Lazarsfeld, 1959) of local independence.

Stated formally,

$$P(u_i = 1|\theta) = P(u_i = 1|\theta, u_j, u_k, \ldots) \qquad (i \neq j, k, \ldots) \tag{3.9}$$

A mathematically equivalent statement of local independence is that the probability of success on all items is equal to the product of the separate probabilities of success. For just three items i, j, k, for example

$$P(u_i = 1, u_j = 1, u_k = 1|\theta) = P(u_i = 1|\theta)P(u_j = 1|\theta)P(u_k = 1|\theta)$$

Local independence is an important assumption that has been taken as a cornerstone of IRT. Item Response Theory posits local independence, or conditional independence of item response given item parameters and examinee ability parameters. By taking the assumption of local independence, one can elicit the likelihood function in a product form (see Section 3.3).

Many researchers have challenged the validity of local independence assumption in the context of practical adaptive testing and this assumption is still a debatable issue in IRT (see Mislevy & Chang, 2000; Jiao & Kamata, 2003). In the scope of this study, we take it for granted.

3.3 Likelihood Function

We have given a simple example of joint probability of item response pattern received from three items in Section 3.2. In this section, we will give the general forms of joint probability of item response pattern to n items, for both one examinee and all examinees cases. The function of joint probability of item response pattern are called likelihood function in IRT.

3.3.1 Joint probability of item responses for one examinee

Suppose we have observed all the item responses u_i $(i = 1, 2, \ldots, n)$ for one examinee and for a given θ. This is the case when one examinee takes n items in the test.

For a single item i, the conditional probability given θ of a single item response u_i is defined as

$$L(u_i|\theta) = \begin{cases} P_i(u_i = 1|\theta) & if \ u_i \ = \ 1, \\ P_i(u_i = 0|\theta) = 1 - P_i(u_i = 1|\theta) & if \ u_i \ = \ 0 \end{cases} \tag{3.10}$$

This can be written more compactly in various way because $u_i = 1$ or 0. We shall write

$$L(u_i|\theta) = P_i^{u_i} Q_i^{1-u_i} \tag{3.11}$$

where $P_i = P_i(u_i = 1|\theta)$, $Q_i = 1 - P_i$, and $u_i = 1$ or 0

Because of local independence, success on one item is statistically independent of success on other items. Therefore, the joint probability of all item responses, given

θ, is the product of the distributions for the separate items:

$$L(\boldsymbol{u}|\theta; \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \equiv L(u_1, u_2, \ldots, u_n|\theta) = \prod_{i=1}^{n} P_i^{u_i} Q_i^{1-u_i} \tag{3.12}$$

where:

$P_i = P_i(u_i = 1|\theta)$ is the probability of correct response to item i, and $Q_i = 1 - P_i$,

u_i is the response to item i, 1 for success and 0 otherwise,

$\boldsymbol{u}$ is the vector of responses $\{u_1, u_2, \ldots, u_n\}$,

$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ are vectors of item parameters a_i, b_i, c_i $(i = 1, , \ldots, n)$, and

θ is ability level of the examinee.

3.3.2 Joint probability of item responses for all examinees

The joint probability of the N different $\boldsymbol{u}$ for all examinees is the product of the separate probability. This joint probability is then

$$L(\boldsymbol{U}|\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \equiv L(\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_N}|\boldsymbol{\theta}) = \prod_{j=1}^{N} \prod_{i=1}^{n} P_{ij}^{u_{ij}} Q_{ij}^{u_{ij}} \tag{3.13}$$

where:

$P_{ij} = P_i(\theta_j)$ is the probability of correct response to item i for examinee j,

u_{ij} is the response to item i for examinee j, $(i = 1, 2, \ldots, n; j = 1, 2, \ldots, N)$

$\boldsymbol{U}$ is the matrix of responses,

$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ are vectors of item parameters a_i, b_i, c_i $(i = 1, , \ldots, n)$, and

$\boldsymbol{\theta}$ is the vector of ability for N examinees, $\{\theta_1, \theta_2, \ldots, \theta_N\}$.

We return to Equation 3.13 in Section 3.5 on JMLE.

3.4 Ability Estimation

The essential goal of IRT is to obtain a measure of an examinee's ability level based on the observation of his responses to the n items administered.

Section 3.3.1 suggests that Equation 3.12 (the joint probability of observed responses) can be solved for θ, the unknown ability level, if item responses are observed and item parameters are known from pre-testing. Therefore, it is possible to infer the examinee's ability level from his observed responses by finding an ability estimate $\hat{\theta}$ that maximizes the likelihood function in Equation 3.12. The ability value obtained in this manner is called maximum likelihood estimator $\hat{\theta}$.

3.4.1 Maximum likelihood estimation (MLE)

The process of maximizing the likelihood function in Equation 3.12 with respect to ability level variable is maximum likelihood estimation (MLE).

Maximum likelihood estimation is a popular statistical method used to make inferences about parameters of the underlying probability distribution of a given data set. By definition (Harris & Stocker, 1998), a likelihood function $L(t)$ is the probability or probability density for the occurrence of a sample configuration, $x_1, x_2, \ldots, x_n$ given that the probability density $f(x|t)$ with parameter t is known: $L(t) = \prod_{i}^{n} f(x_i|t) = f(x_1|t) \cdots f(x_n|t)$; and a maximum likelihood estimator $\hat{t}$ is a value of the parameter t such that the likelihood function $L(t)$ is a maximum.

The numeric means to obtain maximum likelihood estimate of parameter t corresponds to finding the roots by solving the equation $\partial L(t)/\partial t = 0$, where t represents the variable of interest.

Suppose a given examinee j responds to the n items of a test, and the responses are dichotomously scored, $u_{ij} = 0, 1$, where i designates the item $1 \leq i \leq n$, and j

designates the examinee $1 \leq j \leq N$, yielding a vector of item responses of length n denoted by $\boldsymbol{u_j} = (u_{1j}, u_{2j}, \ldots, u_{nj}|\theta_j)$. Under the local independence assumption, the u_{ij} are statistically independent. Thus, the probability of the vector of item responses for a given examinee is given by the likelihood function

$$L = P(\boldsymbol{u_j}|\theta) = \prod_{i=1}^{n} P_i(\theta_j)^{u_{ij}} Q_i(\theta_j)^{1-u_{ij}} \tag{3.14}$$

To simplify the notation, let $P_i(\theta_j) = P_{ij}$ and $Q_i(\theta_j) = Q_{ij}$; then

$$L = P(\boldsymbol{u_j}|\theta) = \prod_{i=1}^{n} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}} \tag{3.15}$$

Taking the natural logarithm of the likelihood function yields

$$l \equiv \log L = \log P(\boldsymbol{u_j}|\theta) = \sum_{i=1}^{n} [u_{ij} \log P_{ij} + (1 - u_{ij}) \log Q_{ij}] \tag{3.16}$$

Since parameters for all n items are assumed to be known here, P_{ij} are functions of item characteristic curve and only derivatives of the log-likelihood with respect to a given examinee's ability parameter will need to be taken in order to solve the problem of max : $\log P(\boldsymbol{u_j}|\theta)$.

$$\frac{\partial l}{\partial \theta_j} = \sum_{i=1}^{n} \left[\frac{u_{ij}}{P_{ij}} \cdot \frac{\partial P_{ij}}{\partial \theta_j} + \frac{1 - u_{ij}}{Q_{ij}} \cdot \frac{\partial Q_{ij}}{\partial \theta_j} \right] \tag{3.17}$$

The derivatives of P_{ij} and Q_{ij} with respect to the ability parameter will be dependent on the item characteristic curve model employed. For purpose of this presentation, these derivatives will be left in their definition form.

The maximum likelihood estimation procedure seeks the value of $\hat{\theta}_j$ which maximizes the likelihood function. It is equivalent to seek the maximum of log-likelihood function. A general solution to such maximum problem is finding the value of θ_j that satisfies the condition $\partial l/\partial\theta_j = 0$. The Newton-Raphson technique is usually used to obtain the estimates of an ability parameter via an iterative procedure. Thus, the second-order partial derivatives of the likelihood function with respect to the ability parameter will also be needed. For a given examinee, a Newton-Raphson equation can be established and solved iteratively for the maximum likelihood estimate of ability. This equation is as follows

$$[\hat{\theta}_j]_{k+1} = [\hat{\theta}_j]_k - \left[\frac{\partial^2 l}{\partial {\theta_j}^2}\right]_k^{-1} \cdot \left[\frac{\partial l}{\partial \theta_j}\right]_k \tag{3.18}$$

When this Newton-Raphson procedure has been performed, an ability estimate $\hat{\theta}_j$ for the examinee j is obtained. In this study, the Newton-Raphson equations for ability parameter estimation are presented under 2-PL ICC of interest (see 3.4.2).

3.4.2 MLE with 2-PL ICC model

As mentioned in Section 3.1, the logistic models are favored over the normal ogive models for computational ease. This assertion stay upright here in the case of MLE. As the Newton Raphson procedure of MLE needs the information of 1st and 2nd derivatives of log-likelihood function, the logistic models stand out for the simplicity and feasibility in its mathematical expression for such 1st and 2nd derivatives.

Consequently, the equations for the maximum likelihood estimation of an examinee's ability will be derived below using 2-PL model. These equations also apply to 1-PL Rasch model by just substituting $a = 1.0$. The equations for 3-PL model are omitted here[7].

Again, the first and second derivatives of the log-likelihood with respect to ability

[7] First, the guessing parameter c in 3-PL model makes the MLE equations more complicated than 2-PL model. Secondly, we did not use 3-PL in this study, the reason is discussed in Section 3.5.2

need to be computed in 2-PL ICC.

The $\partial P_{ij}/\partial\theta_j$ and $\partial Q_{ij}/\partial\theta_j$ of 2-PL model are

$$\frac{\partial P_{ij}}{\partial\theta_j} = Da_iP_{ij}Q_{ij} \tag{3.19}$$

$$\frac{\partial Q_{ij}}{\partial\theta_j} = -Da_iP_{ij}Q_{ij} \tag{3.20}$$

where D is the constant and a_i is the discrimination parameter in 2-PL ICC (see Equation 3.5). Substituting these derivatives in Equation 3.17 yields the first derivative of the log-likelihood with respect to θ_j, and it is

$$\frac{\partial l}{\partial\theta_j} = \sum_{i=1}^{n} Da_i(u_{ij} - P_{ij}) \tag{3.21}$$

The second-order derivative of the log-likelihood function with respect to θ_j is

$$\frac{\partial^2 l}{\partial{\theta_j}^2} = \frac{\partial}{\partial\theta_j}\left[\sum_{i=1}^{n} Da_i(u_{ij} - P_{ij})\right] = -\sum_{i=1}^{n}(Da_i)^2P_{ij}Q_{ij} \tag{3.22}$$

Substituting Equation 3.21 for $\partial l/\partial\theta_j$ and Equation 3.22 for the $\partial^2 l/\partial\theta_j^2$ in Equation 3.18 yields

$$\begin{aligned}
[\hat{\theta}_j]_{k+1} &= [\hat{\theta}_j]_k + \left[\sum_{i=1}^{n}(Da_i)^2P_{ij}Q_{ij}\right]_k^{-1} \cdot \left[\sum_{i=1}^{n} Da_i(u_{ij} - P_{ij})\right]_k && (3.23)\\
&= [\hat{\theta}_j]_k + \left[\frac{\sum_{i=1}^{n} Da_i(u_{ij} - P_{ij})}{\sum_{i=1}^{n}(Da_i)^2P_{ij}Q_{ij}}\right]_k && (3.24)
\end{aligned}$$

which could be solved iteratively for the value of $\hat{\theta}_j$ for each examinee.

3.4.3 Mathematical properties of MLE

In general, the maximum likelihood estimation approach has desirable mathematical and optimality properties, for example (see NIST, 2003),

- It becomes minimum variance unbiased estimator as the sample size increases. By unbiased, we mean that if we take (a very large number of) random samples with replacement from a population, the average value of the parameter estimates will be theoretically exactly equal to the population value. By minimum variance, we mean that the estimator has the smallest variance, and thus the narrowest confidence interval, of all estimators of that type.
- It has approximate normal distributions and approximate sample variances that can be used to generate confidence bounds and hypothesis tests for the parameters.

However, it is noteworthy that maximum likelihood estimation approach has some common disadvantages.

- It can be heavily biased for small samples and the optimality properties may not apply for small samples (see Baker, 1992, ch. 3,4).
- It can be sensitive to the choice of starting values.

In summary, the estimation of a single examinee's ability is based on his vector of responses to n binary items and known values of item parameters. The mathematical details of the solution equations varies on the type of ICC used. In the case of 2-PL model, the estimation process was formulated as an iterative Newton-Raphson procedure.

The maximum likelihood procedure for the estimation of an examinee's ability is the second one of the two fundamental blocks underlying IRT test analysis procedures for dichotomously scored item (the first block is the item parameters estimation procedure, see Section 3.5). These two blocks will be incorporated in various ways to yield procedures that estimate both the item parameters and examinee parameters for a set of test results.

3.5 Item Parameters Estimation

Recall the approach of ability estimation, the ability estimate is obtained by finding an ability value that maximizes the likelihood function which is just the joint probability of item responses for one examinee $L(\boldsymbol{u}|\theta, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ (see Equation 3.12). The essence of this case is that the joint probability of item responses for one examinee is treated as a single variable function of ability level θ, and the remaining parameters, such as, examinee's responses and item parameters, are taken as known parameters.

Likewise, the IRT joint probability of item responses for all examinees $L(\boldsymbol{U}|\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ (see Equation 3.13) can be treated as a multivariate function of item parameters $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ and ability parameters $\boldsymbol{\theta}$ when all examinees' responses $\boldsymbol{U}$ are known. We can presumably think there exists a set of $\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{c}}, \hat{\boldsymbol{\theta}}$ that best fits our observation of item responses in the way of maximizing the likelihood function.

The IRT joint maximum likelihood estimation (JMLE) is an approach that estimates item parameters by finding $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ and $\boldsymbol{\theta}$ values that maximizes the likelihood function $L(\boldsymbol{U}|\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ (Birnbaum, 1968). Because the $\boldsymbol{\theta}$ values are simultaneously evaluated along with item parameters $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$, as a byproduct in this approach, it is so called "joint" maximum likelihood estimation.

3.5.1 Joint maximum likelihood estimation (JMLE)

For the sake of simplicity, the parameter estimate problem can be expressed in the following. We are given a matrix $\boldsymbol{U}$ consisting of the responses ($u_{ij} = 0, 1$) of each of N examinees to each of n items. To be general, we assume that these responses arise from a 3-PL model (see Equation 3.6). We need to infer the parameters of the model: a_i, b_i, c_i $(i = 1, 2, \ldots, n)$ and θ_j $(j = 1, 2, \ldots, N)$.

The maximum likelihood estimates are the parameter values that maximize the likelihood $L(\boldsymbol{U}|\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ given the observed $\boldsymbol{U}$. Maximum likelihood estimates are usually found from the roots of the likelihood equations (see Equation 3.13), which set the derivatives of the log-likelihood equal to zero.

Let $l \equiv \log L(\boldsymbol{U}|\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$, the log-likelihood equations are

$$\frac{\partial l}{\partial \theta_j} = \sum_{i=1}^{n} \frac{u_{ij} - P_{ij}}{P_{ij}Q_{ij}} \frac{\partial P_{ij}}{\partial \theta_j} = 0 \qquad (j = 1, 2, \ldots, N) \tag{3.25}$$

$$\frac{\partial l}{\partial a_i} = \sum_{j=1}^{N} \frac{u_{ij} - P_{ij}}{P_{ij}Q_{ij}} \frac{\partial P_{ij}}{\partial a_i} = 0 \qquad (i = 1, 2, \ldots, n) \tag{3.26}$$

$$\frac{\partial l}{\partial b_i} = \sum_{j=1}^{N} \frac{u_{ij} - P_{ij}}{P_{ij}Q_{ij}} \frac{\partial P_{ij}}{\partial b_i} = 0 \qquad (i = 1, 2, \ldots, n) \tag{3.27}$$

$$\frac{\partial l}{\partial c_i} = \sum_{j=1}^{N} \frac{u_{ij} - P_{ij}}{P_{ij}Q_{ij}} \frac{\partial P_{ij}}{\partial c_i} = 0 \qquad (i = 1, 2, \ldots, n) \tag{3.28}$$

In general, for the three-parameter logistic model, we have

$$\begin{aligned}\frac{\partial P_{ij}}{\partial \theta_j} &= \frac{Da_iQ_{ij}(P_{ij}-c_i)}{1-c_i}\\ \frac{\partial P_{ij}}{\partial a_i} &= \frac{D(\theta_j-b_i)Q_{ij}(P_{ij}-c_i)}{1-c_i}\\ \frac{\partial P_{ij}}{\partial b_i} &= \frac{-Da_iQ_{ij}(P_{ij}-c_i)}{1-c_i}\\ \frac{\partial P_{ij}}{\partial c_i} &= \frac{Q_{ij}}{1-c_i}\end{aligned}$$

These formulas are given here to show their particular character. The reader need not be concerned with the details.

The important characteristics of Equation 3.25 is that when the item parameters a_i, b_i, c_i $\quad (i = 1, 2, \ldots, n)$ are known, the ability estimate $\hat{\theta}_j$ for examinee j is found from just one equation out of the N equations (see Equation 3.25). The estimate $\hat{\theta}_j$ does not depends on other $\hat{\theta}$. When the examinee parameter θ are known, the three other parameter a_i, b_i, c_i for item i are estimated by solving just three equations out of Equation 3.26, 3.27, 3.28. the estimates for item i do no depend on the parameter of the other items.

This suggests an iterative procedure where we treat the trial values of $\hat{\theta}_j$, $(j = 1, 2, \ldots, N)$ as known while solving Equation 3.26, 3.27, 3.28 for the estimates $\hat{a}_i$, $\hat{b}_i$, $\hat{c}_i$, $(i = 1, 2, \ldots, n)$; then treat all item parameters a_i, b_i, c_i, $(i = 1, 2, \ldots, n)$ as known while solving Equation 3.25 for new trial values $\hat{\theta}_j$, $(j = 1, 2, \ldots, N)$. This is to be repeated until the numerical values converge. Because of the independence within each set of parameters estimates when the other set is fixed, this procedure is simpler and quicker than solving for all parameters at once.

3.5.2 Known problems of JMLE

The Birnbaum's JMLE approach for item parameters estimation explained in previous section looks straightforward. However, the numeric computing for such procedure is non-trivial. Some problems and difficulties of JMLE approached were reported by Baker (1992),

- Biased item parameters estimation when sample size is small. Studies on *LOGIST* and *BILOG* tools suggests a rule of thumb in practical use, "JMLE works best for large groups of examinees and tests with more than, say, 60 items and 1000 examinees".

- Estimation problems of 3-PL model. The c_i value is often overestimated while a_i is underestimated. When c_i is poorly estimated, there will be an impact on the estimation of remaining parameters a_i and b_i. The ability estimates $\hat{\boldsymbol{\theta}}$ will be indirectly affected when the item parameters are in error. In addition, an initial item parameter estimate must be within a certain neighborhood of a real parameter value for the Newton-Raphson iterative approach to converge.

- The Heywood Case. When the 2-PL and 3-PL models are employed in the JMLE procedure, a phenomenon occurs in certain data set: discrimination estimates for one or more item can become very large which, in turn, results in large values of the ability estimates for examinees answering those item correctly. In successive cycle, both the discrimination and the ability estimates go toward infinity, and the overall solution diverges, i.e. "blows up".

Since the item parameters estimation techniques is a broad issue in IRT, we only outline the sketch of JMLE in this section and leave out the details of numeric computing. Those details can be found at Birnbaum, 1968, ch. 17.9, Lord, 1980, ch. 12, and Baker, 1992, p. 84-113.

3.5.3 Alternatives to JMLE

A distinguishing characteristic of Birnbaum's (1968) joint maximum likelihood estimation (JMLE) paradigm is that examinee abilities are estimated along with the item parameters. In IRT, the item parameters are often referred to as "structural" parameters, which are fixed in number by the size of the test; the ability parameters of the examinees are the "incidental" parameters, which are the numbers depending on the sample size. Neyman and Scott (1948) showed that, when structural parameters are estimated simultaneously with the incidental parameters, the maximum likelihood estimates of structural parameters cannot be consistent as sample size increases[8], with the only exception for one-parameter logistic (Rasch) model.

Therefore, an estimation procedure for two- and three-parameter logistic IRT model that avoids the problem of inconsistent estimation of structural parameters has considerable value. The basic paper in this regard was due to Bock and Lieberman (1970), who developed a marginal maximum likelihood (MMLE) procedure for estimating item parameters. Unfortunately, the Bock and Lieberman approach posed a formidable computational task and was practical for only for very short tests. A subsequent reformation of this marginal maximum likelihood estimation (MMLE) approach by Bock and Aitkin (1981) has resulted in a procedure that is both theoretically acceptable and computationally feasible. Their reformulation, under certain conditions, is an instance of an EM algorithm (see Dempster, Laird, & Rubin, 1977). In general, the EM algorithm is an iterative procedure for finding maximum likelihood estimation of parameters, where the E stands for expectation step and the M

[8] A consistent estimator is an estimator that converges in probability to the quantity being estimated as the sample size grows.

for the maximization step. As a result, the combined name, MMLE/EM, will be used to name the Bock and Aitkin procedure for estimating item parameters. Readers are referred to (Baker, 1992, ch. 6) for more details.

3.6 Test Information Function

Section 3.4 has shown how to obtain a maximum likelihood estimate of an examinee's unknown ability. Given an examinee's ability estimate, it is also of interest to have some measure of how "precise" the estimate is. Test information function is the indicator of such estimate precision that is frequently used in IRT literature and CAT applications.

As its name suggests, the test information function gives a certain amount of "information" at a given ability level. Birnbaum (1968) has defined the test information function as

$$I(\theta) = \sum_{i=1}^{n} \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \tag{3.29}$$

where $P_i(\theta)$ is obtained by evaluating the item response characteristic curve model at θ, and $P_i'(\theta) = \partial P_i(\theta)/\partial\theta$. Note that Equation 3.29 involves only the ability level θ and the item response characteristic curves of the items in the test.

Birnbaum (1968) has shown that Equation 3.29 is the upper bound on the amount of information that can be yielded by any possible test scoring formula, thus providing another advantage to using the maximum likelihood estimate of ability (see Section 7.2).

Test information defined in Equation 3.29 has a close relationship to MLE sample variance. Cramer (1946) has shown that a maximum likelihood estimator $\hat{\theta}$ has a normal asymptotic distribution with mean θ and variance $\sigma^2 = 1/I(\theta)$, where $I(\theta)$

is the test information and $\sqrt{1/I(\theta)}$ is the standard error (see Section 7.2). Some additional mathematical conditions are necessary, but the usual ICC models meet them (Samejima, 1977). In the present context, the variance of interest $\sigma^2_{\hat{\theta}|\theta}$, is the variance of the conditional distribution of θ at a given ability level $\hat{\theta}$. Thus, the larger this variance, the less precise the estimate of θ and the less information one has about an examinee's unknown ability level.

From a test theory point of view, defining the test information function in terms of the variance of the conditional distribution of the maximum likelihood estimates of ability is a crucial concept. It provides substantive interpretation of the meaning of the amount of information. The greater the amount of information at a given ability level, the more closely the maximum likelihood estimates of ability cluster around the true but unknown ability level, hence, the more precise the estimate is.

We return to the test information when we discuss the Fisher information in Section 5.1.

Summary

This section describes one of the two poles in our comparison study, the Item Response Theory (IRT).

Item Response Theory generally assumes an single latent trait, examinee's ability level, behind observed sample of item responses. However, multiple latent traits can be found in multidimensional IRT models and thus they are out of the scope of this study. This latent trait (ability level) is treated as a conceptual index that intended to assess an examinee's ability or knowledge state in a single "*score*".

Item characteristic curve (ICC) quantifies the relationship between the probability of correct item response and examinee's ability level given fixed item parameters. ICC model was originally developed using the normal ogive model but the logistic model with the re-scaling provides virtually the same results while simplifying the computations greatly. Due to its amenability to mathematical analysis, logistic models (see Equation 3.8) become the de facto ICC models in both theoretical and practical use.

The local independence assumption in IRT states the conditional independence of item response given item parameters and examinee's ability parameter. The likelihood functions are derived from the local independence assumption. They link the joint probability of observed item response pattern to examinee's ability and the item parameters.

Via the maximum likelihood estimation technique, it is possible to infer examinee's ability or item parameters from observed sample responses data. The maximum likelihood estimation is generally abstracted as the root(s) finding process. Newton-Raphson based approaches are usually the applicable solutions in such situations. From student modeling perspective, item parameters estimation offers the model creation solution, whereas the ability estimation is the inference process from established toolbox. These are the two fundamental building block of IRT.

Test information function is of special interest because it gives the upper bound to the information that can be obtained by any method of scoring the test. Moreover, it provides a measure of how precise the ability estimate obtained from MLE is.

The other pole of in our comparison study is POKS. The next section will cover POKS in details.

CHAPTER 4

PARTIAL ORDER KNOWLEDGE STRUCTURE (POKS)

Approaches such as Bayesian networks (BN) are considered highly powerful modeling and inferencing techniques because they make few assumptions and they can represent complex relationships among variables with efficiency and parsimony. They can also be learned from training data. Yet, they generally lend themselves to a variety of sound and efficient inference computations. However, in spite of these qualities, BN may not be always be the most advantageous technique in comparison to simpler techniques that make strong assumptions. A simple Bayes posterior probability update approach under strong independence assumption, named POKS (Desmarais et al., 1996; Desmarais & Pu, 2005a, 2005b; Desmarais & Meshkinfam, 2005; Desmarais et al., 2005) is such alternative, and it is explained in this section.

Section 4.1 examines POKS underlying theory, namely, the Theory of Knowledge Spaces (Falmagne et al., 1990) first, which is based on item-to-item node network structure. The POKS network is constructed solely on item nodes, without any hidden nodes or nodes that come from knowledge engineering work.

Section 4.2 reviews the local independence among the item evidence node in POKS. Under the assumption of local independence in item evidence nodes, Section 4.3 explains the steps of how to build POKS inference network, i.e. the network induction process.

Following the network extracted from this induction procedure, we show how to make inference of examinee's knowledge state from observed information. Section 4.4 and Section 4.5 cover the techniques of updating item probability between nodes, or in POKS terms, the propagation of evidence. A small numeric example of evidence propagation is provided in Section 4.6.

4.1 Item-to-item Node Structures and the Theory of Knowledge Spaces

Probably the most distinctive characteristic of POKS is that it permits the inference of known or unknown items based on this structure. It derives from the work of Falmagne et al. (1990). Others such as Kambouri, Koppen, Villano, and Falmagne (1994) have worked towards using the structural characteristics of item-to-item structures to infer an individual's knowledge state.

Item-to-item relations have their cognitive grounding in the Theory of Knowledge Spaces (Falmagne et al., 1990) and they are termed as surmise relations. The meaning of such relation is essentially that we expect people to master these items in the reverse order of these relations. Figure 4.4 illustrates such type of relations with a simple example.

It can be seen that the example in Figure 4.4 comprises the following surmise relations: $a \rightarrow b \rightarrow d$ and $a \rightarrow c \rightarrow d$. However, no relation exists between b and c. For example, if a pupil succeeds item a, it will increase the estimated probability of success to items b, c, d. Conversely, failure to item d will decrease the estimated probability of success to items a, b, c. Finally, failure or success between items b and c will not affect the estimated probability of success to the node according to the Theory of Knowledge Spaces. [9]

However, POKS does not strictly conform to the Theory of Knowledge Spaces because it uses partial orders such as Figure 4.4, whereas *knowledge structures* use

[9] Note that this is not the case for the IRT theory, nor of the Bayesian modeling techniques reviewed in this study, which link every test items to one or more global abilities.

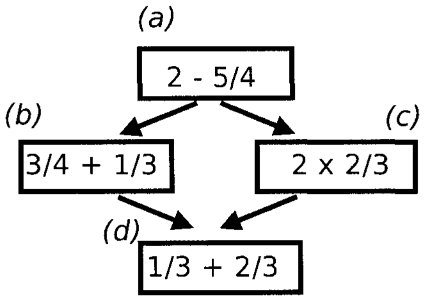

Figure 4.4: A simple example of knowledge structure

AND/OR graph. The difference is that partial orders define possible knowledge states closed under union and intersection, whereas *AND/OR* graphs define possible knowledge states closed under union only. Indeed, defining the knowledge state of an individual as a subset of a global set of knowledge items, Falmagne and his colleagues established that the set of possible knowledge states from a global set of items is constrained by closure under union: if we join two individuals' knowledge state, this is also a possible knowledge state (for details, see Falmagne et al., 1990). If, for example, we define X_b and X_c as two items that test different methods of solving a problem and that any one of these methods can be used in solving a third item X_a (but at least one must be used), this would be reflected in knowledge structures theory as an *OR* relation binding the three nodes and clearly expressing the alternative found in the relation. It is also clear that the intersection of two individuals, each mastering a single alternative method between X_b and X_c, would yield an invalid knowledge state: Someone who masters X_a but none of X_b and X_c (we ignore probabilistic issues here). In POKS, we would likely find weak surmise relations $X_a \rightarrow X_b$ and $X_a \rightarrow X_c$, capturing some of the information but not as accurately as with an *OR* relation.

Nevertheless, because partial orders do capture to a large extent the constraints on possible knowledge states and because the probabilistic nature of POKS makes it more flexible and robust to noise, the use of partial orders remains a powerful means of making knowledge assessment. Moreover, because OR relations are tertiary or higher n-ary relations, they impose larger data sets to discover and are thus more limited in their applications.

4.2 Local Independence

Another characteristic of POKS is that it makes the assumption of local independence among evidence nodes. In POKS, we essentially make the assumption that we can limit the modeling solely to binary conditional probability relations. More formally, we make the assumption that for any node X having parents $pa(X) = \{X_{p1}, \ldots, X_{pn}\}$, all parents are independent of each other:

$$P(X|X_{p1}, \ldots, X_{pn}) = \prod_{i}^{n} P(X|X_{pi}) \tag{4.1}$$

Although this assumption is obviously violated in most contexts, the question of whether it leads to significant errors is an empirical question that will be assessed and discussed further [10]. Local independency assumption implies that the acquisition of knowledge can be modeled in Directed Acyclic Graph (DAG) or "partial order". The great benefit of making this assumption is that it allows the induction of the network from a very small number of data cases. Only the analysis of binary relations data is needed to create the model. In the current experiment, less than 50 data cases were used to build the models.

[10] Note that the local independence assumption is also an issue as discussed in Section 6.9.3.

4.3 POKS Network Induction

Knowledge structures such as the example in Figure 4.4 are learned from data. The POKS graph model and induction technique is briefly reviewed here.

4.3.1 Nodes

As mentioned above, POKS structures, like other graph approaches, can include nodes that represent concepts or test items, much like other user modeling graphical models, and multiple dimensions could be represented by concepts and hierarchies of nodes. However, for the purpose of comparing IRT and POKS, the nodes are limited to representing test items. There are no other types of node, each node is a test item, and each test item is a node. All items have equal weight for this experiment.

Each node, X_i, is assigned a probability that represents an examinee's chances of mastery of that item, $P(X_i)$. Contrary to the IRT model, $P(X_i)$ is not a function of θ, the ability level. It is a direct function of the probability of other items from which it is linked with. The details of how to compute $P(X_i)$ in POKS is described in Section 4.4.

4.3.2 Relation

Relations in POKS have the same meaning as knowledge spaces' *surmise* relations: They indicate the (partial) order in which people learn to master knowledge items (see Section 4.1). Although surmise relations are different from causality relations found in Bayesian networks, they allow the same type of inferences[11]. For example, let X_a and X_b be two items in an item bank, a relation $X_a \rightarrow X_b$, means that observing an examinee succeed item X_a will increases the estimated probability of success to item

[11] In fact, causality also has the property of ordering events in time, and it is a non trivial philosophical endeavor to determine that it has any other property.

X_b by a certain amount. Conversely, a failure to item X_b will decrease the estimated probability of success to item X_a.

4.3.3 Networks structure

In accordance with the assumption of local independence, the network construction process consists in comparing items pairwise to look for a relation. To determine if there is a directed link, $X_a \rightarrow X_b$, the three following conditions must hold:

$$P([P(X_b|X_a){\geq}p_c] \mid D) > (1 - \alpha_c) \tag{4.2}$$

$$P([P(\neg X_a|\neg X_b){\geq}p_c] \mid D > (1 - \alpha_c) \tag{4.3}$$

$$P(X_b|X_a){\neq}P(X_b) \tag{4.4}$$

where:

$P(X_b|X_a) = P(X_b = 1|X_a = 1)$ and $P(\neg X_a|\neg X_b) = P(X_a = 0|X_b = 0)$

p_c is the minimal conditional probability for $P(X_b|X_a)$ and $P(\neg X_a|\neg X_b)$; an single value is chosen for the test of all relations in the network, generally 0.5;

α_c is the alpha error of the conditional probability tests (Equation 4.2 and Equation 4.3); it determines the proportions of relations that can erroneously fall below p_c; common values range from 0.2 and 0.5.

α_i is the alpha error of the interaction test (Equation 4.4);

D is the joint frequency distribution of X_a and X_b in the calibration sample. This joint distribution is a 2 × 2 contingency table with four frequency numbers, $\{x_{ab}, x_{a\neg b}, x_{\neg ab}, x_{\neg a\neg b}\}$, representing the number of examinees in the sample data broken down into these four situations:

1. x_{ab}: success for X_a and X_b

2. $x_{a\neg b}$: success for X_a and failure for X_b

3. $x_{\neg ab}$: failure for X_a and success for X_b

4. $x_{\neg a\neg b}$: failure for X_a and X_b

The first condition (Equation 4.2) states that the conditional probability of a success for X_b given a success for X_a must be above a minimal value, p_c, and that we can derive such conclusion from a sample data set D, with an error rate smaller than α_c. The second condition (Equation 4.3) is analogous to the first and states that the probability of failure for X_a given a failure for X_b must be greater than p_c, with a maximal error rate of α_c given distribution D.

These first two conditions are computed from the cumulative Binomial distribution function. In inequality Equation 4.2, the value of $P([P(X_b|X_a)] \mid D)$ is obtained by the summation of the Binomial probability function for all distributions where $x_{a\neg b}$ are less than the number actually observed in D, that is:

$$\begin{aligned}
P([P(X_b|X_a)] \mid D) &= P(x \leq x_{a\neg b}|X_a) \\
&= \sum_{i=0}^{x_{a\neg b}} Bp(i, x_a, p_c) \\
&= \sum_{i=0}^{x_{a\neg b}} \binom{x_a}{i} p_c^{[x_a - i]}(1 - P_c^i)
\end{aligned}$$

where $x_a = x_{ab} + x_{a\neg b}$. The conditional probability of the second condition (inequality Equation 4.3) rests on the same function but uses $Bp(i, x_{\neg b}, p_c)$ in place of $Bp(i, x_a, p_c)$.

The third condition (inequality Equation 4.4) is an independence test and it is verified by a χ^2 distribution test on the 2×2 contingency table of distribution D:

$$P(\chi^2) < \alpha_c$$

For small samples, the independence test used is replaced by the Fisher exact test.

The choice of value for the p_c indicates the strength of the *surmise* relations we want to keep. For example, if the order in which one learns to master two items is highly constrained, in accordance with the Theory of Knowledge Spaces, then we would expect to find that $P(B|A) \approx 1$ for a strong surmise relation $X_a \rightarrow X_b$. The value of p_c represents the lower limit for which we accept a surmise relation. The choice of a value is somewhat arbitrary, but we generally use $p_c = 0.5$ in our experiments.

The two values α_c and α_i represent the alpha error we are willing to tolerate when concluding the corresponding tests. For very small samples, these values can be as high as 0.5 in order to keep as many relations as possible. In our experiments they are set between 0.2 and 0.1 (see Section 6.9).

4.4 Item Probability Update

When an item's probability of mastery in the network changes, either through observation or through a change in the probability of a neighboring node, evidence is propagated through the connected items in the network. If the probability increases, the update will follow links forward, whereas if the probability decreases, the update will follow links backward. We use the algorithm for evidence propagation from Giarratano and Riley (1998). This algorithm is consistent with the Bayesian posterior probability computation in single layered networks and corresponds to the posterior probability update. However, for multilayered networks, in which indirect evidence gets propagated (transitive evidence from non directly connected nodes), an interpolation scheme is used. This is explained in the numerical example of Section 4.6.

For computational convenience, the algorithm relies on two odds ratios: the *likelihood of sufficiency* and the *likelihood of necessity* respectively defined as:

$$LS_{a \to b} = \frac{O(X_b|X_a)}{O(X_b)} \tag{4.5}$$

$$LN_{a \to b} = \frac{O(X_a|\neg X_b)}{O(X_a)} \tag{4.6}$$

where $O(X)$ is the odds function, $P(X)/Q(X)$ (where $Q(X) = P(\neg X) = 1 - P(X)$), and $O(X|Y)$ is the conditional form, $P(X|Y)/Q(X|Y)$.

It follows that if we know X_a to be true (i.e. $P(X_a) = 1$), then the probability of X_b can be updated using this form of equation Equation 4.5:

$$O(X_b|X_a) = LS_{a \to b} O(X_b) \tag{4.7}$$

and conversely, if X_a is known false, then:

$$O(X_a|\neg X_b) = LN_{a \to b} O(X_a) \tag{4.8}$$

The update process recursively propagates forward using Equation 4.7 when a node's probability increases, and backward using Equation 4.8 when it decreases.

In accordance with the local independence assumption in equation Equation 4.1, it follows that the odds ratios are combined as the product of the LS of each parent that is observed:

$$O(X_j|pa(X_j)) = O(X_j) \sum_{X_i \in pa(X_j)} LS_{i \rightarrow} j \tag{4.9}$$

where $pa(X_j)$ are the observed parents of node X_j and $O(X_j)$ is the initial odds ratio. Conversely, the LN odds ratios are also combined for the children nodes:

$$O(X_k|ch(X_k)) = O(X_k) \sum_{X_i \in ch(X_k)} LN_{k \rightarrow} i \tag{4.10}$$

where $ch(X_k)$ are the observed children of node X_k. We emphasize again that this strong assumption is surely violated in most contexts, but it greatly simplifies node updates by relying on functional computations (as opposed to the computations required for optimizing a system of equations) and on the network's Markovian property: only the network's current state is sufficient to make future predictions. The impact of this assumption's violation will be assessed in the experimental evaluation.

4.5 Evidence Propagation Directionality

The evidence propagation scheme is unidirectional in the sense that if a node's probability increases, no backward propagation is performed, and, conversely, no forward propagation is performed when a node's probability decreases. This may look as a breach into standard Bayesian theory since posterior updates can occur in both directions. In fact, it is not. It follows from POKS principle of pruning non-significant posterior updates relations with the statistical tests Equation 4.2, 4.3, and 4.4. Let us illustrate this with a simple example. Assume the following two question items:

a : Examinee is able to solve for x : $\frac{3}{2x} \times \frac{7}{4} = \frac{3}{8}$

b : Examinee is able to find the answer to $\frac{3}{7} \times \frac{7}{4}$ =?.

The POKS induction algorithm would readily derive $a \rightarrow b$ from a data sample taken from the general population on these two items, indicating that it is worth updating b's posterior probability if we observe a. However, the converse relation $b \rightarrow a$ would probably fail the statistical tests for inequalities Equation 4.2 and Equation 4.3, indicating that the inverse relation is not strong enough. Indeed, it is fairly obvious that a success for item b does not significantly increase the chances of success of a because the latter involves algebra and is significantly more advance than the former. However, if we replace a with an item of closer difficulty to b, such as:

a : Examinee is able to find the answer to $\frac{4+8}{11} \times \frac{11}{12}$ =?

then we would probably also derive $b \rightarrow a$. The result would be a symmetric relation (probably with different LN and LS values for $a \rightarrow b$ and $b \rightarrow a$). In that case, a probability increase or decrease in any node would affect the other node's probability in accordance with Bayes posterior probability update, and propagation would be bidirectional.

When relations are symmetrical, $X_b \rightarrow X_a$ and $X_a \rightarrow X_b$, cycles involving two nodes are created. There are two solutions to this problem, the first solution consists in grouping symmetrical nodes into a single one. A second solution, adopted for this study, is simply to keep symmetrical relations but to stop the propagation of evidence once a node has already been visited during a single propagation run. This is a standard procedure in message propagation and constraint programming systems.

4.6 Numerical Example

Let us illustrate numerically the evidence propagation with an example. Assume the following relations hold:

$$a \longrightarrow b, \quad b \longrightarrow c$$

and that in our sample we find:

$$P(X_c) = 0.3, \quad P(X_c|X_b) = 0.6, \quad P(X_c|X_a) = 0.9$$

It follows from the above equations that observing X_a first (i.e. $P'(X_a) = 1$) [12] would bring $P'(X_c) = 0.9$, which corresponds to the value of the sample's observed conditional probability $P(X_c|X_a)$. Further observing X_b would bring $P''(X_c) = 0.969$, which corresponds to $P(X_c|X_b, X_a)$. Inversion of the order of observation would bring instead $P'(X_c) = 0.6$ after observing X_b (i.e. $P(X_c|X_b)$) and $P''(X_c) = 0.969$, as expected (i.e. $P(X_c|X_b, X_a)$).

Although odds are used in the algebra for computing the posterior probabilities, it is equivalent to using the standard Bayes formula for obtaining the posteriors given the observation $X = 1$ or $X = 0$. However, when the probability of a node increases or decreases, an interpolation scheme is used to further propagate evidence.

When the probability of one of a node's parent nodes changes by some value, without actually being observed and thus reaching the value of 1 or 0, two interpolation formulas are used to update this node's probability. Assuming a relation $a \longrightarrow b$, and an increase in $P(X_a)$ of $\triangle_a$ (i.e. $P'(X_a) = P(X_a) + \triangle_a$), where $P(X_a)$ represents the probability before the update and $P'(X_a)$ the probability after the update, then the

[12] We use the notation $P'(X)$ to represent an updated probability and drop the conditional form, $P(X|\mathit{evidence})$, to better emphasize the stages of updating. $P'(X)$ is the value of $P(X)$ after the first stage of updating, whereas $P''(X)$ is the value after the second stage.

value of $P'(X_b)$ is given by:

$$P'(X_b) = P(X_b|X_a) + [P(X_b|X_a) - P(X_b)]\,\frac{P'(X_a) - P(X_a)}{P(X_a)}$$

where $P(X_b)$ is the probability of X_b before the update.

Following $a \longrightarrow b$ in the backward propagation direction and assuming a decrease $P(X_b) - P'(X_b) = \triangle_b$, the updating formula is:

$$P'(X_a|\neg X_b) = P(X_a|\neg X_b) + [P(X_a) - P(X_a|\neg X_b)]\frac{P'(X_b)}{P(X_b)}$$

This interpolation method is a simple approximation of $P(X|E_1, E_2)$, where $E_1 \longrightarrow X$ and $E_2 \longrightarrow E_1$ are directly linked, but $E_2 \longrightarrow X$ are not. Its validity for the field of CAT is a question we investigate empirically in this study. More details about the interpolation method can be found in Giarratano and Riley (1998).

Summary

POKS is a specific Bayesian modeling approach which makes several strong assumptions to reduce the complexity in Bayesian network modeling. POKS allows the the Bayesian modeling of item-to-item knowledge structure in accordance to the Theory of Knowledge Spaces (Falmagne et al., 1990). One of the purposes of this study is to explore the validity and performance of POKS modeling framework under the strong assumptions imposed.

The network of POKS is defined solely over the test items and no concepts nodes are included. Imposing this rule relieves POKS from any knowledge engineering effort to construct the network and thus makes the approach more comparable to IRT than other Bayesian modeling approaches that would require a knowledge engineering step. The same data can be used for both POKS and IRT approaches without any manipulation or transformation, thus allowing a comparison of the two approaches on an equal basis.

The POKS network induction algorithm relies on a pairwise analysis of item-to-item relationships. Such analysis attempts to identify the order in which people master knowledge items. It is inspired from the Knowledge Spaces Theory which states that skill acquisition order can be modeled by an AND/OR graph (Falmagne et al., 1990). For our purpose, we impose a stronger assumption that the skill acquisition order can be modeled by an directed acyclic graph (DAG). This assumption allows us to limit POKS network induction algorithm only to pairwise analysis.

Given the assumption of independence in evidences, the probability update of evidence can be written in posterior odds. This allows POKS to use Bayes' Theorem in its own implementation based on odds and likelihood algebra. This implementation of updating evidence is consistent with the Bayesian posterior probability computation in single layered networks and corresponds to the posterior probability update.

So far, we have explained IRT in Section 2.3.4 and POKS in this section. The next section will explore the simulation experiments and their results of comparison between IRT and POKS approaches.

CHAPTER 5

ITEM SELECTION

As mentioned in Section 1.4, there is a common concern to IRT and POKS in adaptive testing, i.e. item selection. In the CAT context, the goal of item selection is, to use the least number of items to identify the examinee's ability level with maximum precision, or in other words, choosing the most informative item.

This section describes two folds of item selection methods used in this study, namely Fisher information approach and information gain approach.

5.1 Fisher Information

In Section 3.6, the IRT test information (Equation 3.29) is defined as

$$I(\theta) = \sum_{i=1}^{n} \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)}$$

Since the right-hand side of above equation is a sum, it can be decomposed into the contribution of each item to the amount of test information.

The item information function is the amount of information contributed by an individual item. It is given by

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \tag{5.1}$$

The item information function is also called Fisher information function due to their equivalence in IRT context (see Section 7.2).

The equations of Fisher information for logistic ICC models are

$$\begin{aligned} I_i(\theta) &= P_i(\theta)Q_i(\theta) && \text{1-PL or Rasch} && (5.2) \\ I_i(\theta) &= (Da_i)^2 P_i(\theta)Q_i(\theta) && \text{2-PL} && (5.3) \\ I_i(\theta) &= (Da_i)^2 \frac{Q_i(\theta)}{P_i(\theta)} \left[\frac{P_i(\theta) - c_i}{1 - c_i}\right] && \text{3-PL} && (5.4) \end{aligned}$$

From Equation 5.4, it is relatively easy to infer the role of the b_i, a_i, and c_i parameter in the Fisher information function: (a) information is higher when the b_i value is close to θ than when the b_i value is far from θ; (b) information is generally higher when the a_i parameter is high; and (c) information is increases as the c_i parameter goes to zero.

Under all three logistic models, the $I_i(\theta)$ curve is always bell-shaped, with its maximum at θ_{max}. For 1-PL and 2-PL models, $\theta_{max} = b_i$ exactly; and for 3-PL modle, $\theta_{max} = b_i + \frac{1}{a_i} \log \left[\frac{1+\sqrt{1+8c_i}}{2}\right]$. In all cases, the amount of information can be quite small for the ability levels that deviate considerably from θ_{max}. This indicates that the estimation of ability is better when the difficulty of the item θ_{max} is matched to the examinee's ability.

For a test consisting of n items, the test information is the sum of Fisher information of every item. Selecting the item with maximum Fisher information maximizes the contribution to the test information. The usefulness of this is readily understood if an examinee's ability estimate is wanted, especially when the maximum likelihood estimator (MLE) is used. In MLE case, the standard error of $\hat{\theta}$ is estimated as $SE(\hat{\theta}) \equiv \sqrt{Var(\hat{\theta})} = 1/\sqrt{I(\hat{\theta})}$ (see Section 7.2). Therefore, by selecting items having maximum information, the contribution to the decrease of the standard error is the greatest. Furthermore, from the definition of Equation A.6, it can be seen that maximizing the information is the same as maximizing the contribution of an item to the expected relative rate of change in the likelihood function. The greater this change rate at given θ, the better it can be distinguished from points near to this value, and the better this value can be estimated (see Chang & Ying, 1996).

In Section 6.7, we will discuss the two variants of Fisher information based item selection methods. One is to select the item having the largest item information at current ability estimate. The other is to select the item having the largest item information at predefined cut point on ability scale which is tailored to the two-category classification problem (e.g. pass/fail).

5.2 Information Gain

The second approach to item selection we investigate is the information gain approach. The principle of this approach is to choose the item that will maximize the expected reduction of entropy of the test. This is explained below.

The entropy of a single item X_i is defined as:

$$H(X_i) = -[P(X_i)\log(P(X_i)) + Q(X_i)\log(Q(X_i))] \tag{5.5}$$

where $Q(X) = 1 - P(X)$. The entropy of the whole test is the sum of all individual item's entropy:

$$H_T = \sum_{i}^{k} H(X_i) \tag{5.6}$$

where the subscript T in H_T indicate it is the entropy of the test which contains k items.

If all item probabilities are close 0 or 1, the value of H_T will be small and there will be little uncertainty about the examinee's ability. It follows that we minimize this uncertainty by choosing the item that maximizes the difference between the current test entropy (i.e. H_T) and the entropy after the response of that item is observed (i.e. H'_T). Since H'_T is unknown in the case where the response to item i is not determined. Thus, the expected value of the whole test entropy after a response to item X_i is given instead:

$$E_i(H'_T) = P(X_i)H'_T(X_i = 1) + Q(X_i)H'_T(X_i = 0) \tag{5.7}$$

where $H'_T(X_i = 1)$ is the entropy after the examinee answers correctly to item i, and $H'_T(X_i = 0)$ is the entropy after a wrong answer. We then look for the item that will have the maximum difference:

$$\max_i \left[H_T - E_i(H'_T)\right] \tag{5.8}$$

5.3 Computing the Fisher Information with POKS

The Fisher information defined in Equation 5.1 is defined with respect to the value of θ in IRT ability level scale. In order to apply Fisher information methods to POKS, POKS must provide an assessment of examinee in IRT's ability level scale, i.e. θ.

In POKS, the equivalent θ is computed from a measure of the estimated mastery level. That measure, m, corresponds to average probability over all k items:

$$m = \frac{\sum_i^k P(X_i)}{k} \tag{5.9}$$

Note that we could also have used the expected item success rate as an alternative[13], but the current measure is more sensitive as it discriminates between an item with a probability of success .49 and another with probability .01.

The value of m varies on a scale $[0, 1]$, whereas θ in IRT is on the scale $[-\infty, +\infty]$. To bring m onto the θ scale, we apply the *logit transformation*, with item parameters a and b that are commonly shared with IRT:

$$\theta_m = logit(m)/a + b = \left[\log\left(\frac{m}{1-m}\right)\right]/a + b \tag{5.10}$$

Therefore, the above θ_m obtained from POKS can be viewed as a transformed "equivalent" ability level value in genuine IRT. By taking this conceptual equivalence in ability level, POKS is entitled to use IRT's exclusive formula of item information function, e.g. for 2-PL model, the Fisher information for θ_m is

$$I_i(\theta_m) = (Da_i)^2 P(\theta_m)(1 - P(\theta_m))$$

where a_i is the discrimination parameter of item i and D is constant 1.7 (see original item information function, Equation 5.3).

Summary

Fisher information function, or item information function, is related to test information function in IRT. Choosing the item with maximum Fisher information maximizes the contribution to the test information, which in turn leads to more precise ability estimate in the MLE case.

[13] For k items, the expected success rate r is defined as

$$r = \frac{\sum_i^k d_i}{k}$$

where $d_i = 1$ for $P(X_i) > 0.5$, and $d_i = 0$ otherwise.

Information gain item selection aims at bringing down the test entropy which eliminates the uncertainty of ability estimate in test process. The item favored by the information gain approach is the one that maximizes the expected reduction of entropy of the test.

Since POKS has no internal index comparable to IRT's ability level θ, it cannot apply the Fisher information approach directly. A special treatment is introduced to create equivalent θ_m for POKS to ease the use of Fisher information approach with POKS.

CHAPTER 6

SIMULATION AND RESULTS

In the previous sections, we explain the IRT and POKS student modeling approaches. This section compares the POKS approach (see Section 3.6) with the IRT-2PL approach (see Section 2.3.4) and reports the corresponding results of the two approaches. For the purpose of this study, the POKS network is defined solely over the test items and no concepts nodes are included. Imposing this requirement not only relieves us from any knowledge engineering effort to construct the network but makes this POKS approach more comparable to IRT than other Bayesian modeling approaches that would require a knowledge engineering step. The same data can be used for both approaches, thus allowing a comparison on an equal basis. A small simulation study was completed. The analysis was conducted on two empirical data sets. All data set consisted of dichotomous item. 2-PL ICC model was used in each set.

6.1 Experimental Evaluation of The Approaches

For the two category classification problem (e.g. pass/fail determination), the general goals of assessing any approach are,

- Effectiveness - Does the approach always yield the results that is better than number-right interpolation? and how much is the margin? The number-right interpolation is the common way of inferring the score, for example, an examinee succeed 2 out of 4 questions administered at the moment, suppose the total number of items in the test is 20, so the number-right interpolation of score is

$2/4 \times 20 = 10$ [14]. The results from number-right interpolation usually serves as the baseline in this comparison study, we generally expect that any inference strategy (POKS or IRT-2PL) would yield better results than crude number-right interpolation approaches.

- Efficiency - How fast the approach yields the pass/fail decision with some confidence? e.g. how many items (or what is the proportion of total number of items) are required to determine an examinee's pass/fail score with the confidence that such decision will be correct 90% of times.

- Generality and robustness - Can the approach achieve effectiveness and efficiency goals under different data sets, e.g. large vs. small items number, large vs. small examinees number, or even ill-designed test?

Moreover, we designed the comparison schemes with the following concerns,

- $N - 1$ principle. Given the N examinees' responses data, $N - 1$ examinees' responses are used for building the model (c.f. training and validation); the responses of the remaining one examinee are used for simulation and analysis. This rule prevents the creation of a bias by using the same data for validation and training (see Section 6.2).

- Various knowledge or ability domains of interests in the test. For example, two sets of empirical test data are explored, UNIX knowledge test and French language proficiency test.

- Distinct passing score (e.g. 50%, 60%, 70%) of determination criteria in pass/fail classification.

[14] The test score used in this comparison study is defined as the number of correctly answered items.

- Item selection strategies. Several item selection procedures are investigated, namely the information gain approach and the Fisher information approach (see Section 6.7). A random item selection procedure is also reported for benchmark comparison.

6.2 Methodology

The performance comparison rests on the simulation of the question answering process. For each examinee, we simulate the adaptive questioning process. The answers given by the examinee during the simulation are based on the actual answers collected in the real test. After each item is administered, an overall estimated score S is computed from POKS and IRT-2PL approaches. This estimated score S changes during the test process and finally approaches the examinee's true score S_t. We monitor this S after every item administered in the test simulation, and compared it with the true score S_t.

As mentioned before, any type of score used in this comparison study is defined based on the concept of number-right score. Therefore, the true score S_t of an examinee is computed from his actual responses in the test data. It is defined as

$$S_t = \frac{\sum_{i}^{n} x_i}{n} \tag{6.1}$$

where $x_i = 1, 0$ is the response to item i, and n is the total number of items in the test. The gap between estimated score S and true score S_t is recorded for .

The estimated score S consists of two parts: the observed (responded) items and the estimated items. The items already responded are assigned the observed values (x_i), whereas the remaining unobserved items take the expected score ($\hat{x_j}$).

The overall estimated score is thus a weighted sum of the observed scores from *responded* items and expected scores from *unobserved* items. That is, if I_r is the set of items responded and I_e is the set of items unobserved, the examinee's estimated score, S, is:

$$S = \frac{\sum\limits_{x_i \in I_e} x_i + \sum\limits_{x_j \in I_r} \hat{x_j}}{n} \tag{6.2}$$

where x_i is 1 if the corresponding response to item i is a success and 0 otherwise, and $\hat{x_j}$ is 1 if the estimated probability of success $P(x_j = 1)$ (with the respective method used, POKS or IRT-2PL) is above 0.5 and 0 otherwise. Recall that in the IRT 2-PL model, the probability of success to an item is given by Equation 3.8 and depends on the current ability estimate $\hat{\theta}$, whereas in POKS, it is computed through the propagation of evidence as explained in section Section 3.6.

During the simulation process, an examinee is classified as master if his estimated score S is above a given cut score S_c. This cut score S_c is often expressed as a percentage value, e.g. 60%. The classification decisions are recorded after every item administered.

This simulation procedure results in a 100% correctly classified examinees after all test items are observed. It rests on the fact that we do not know the actual ability state of an examinee apart from the test results. Indeed, contrary to a frequently used approach that consists in generating test response data cases from Monte Carlo simulations, we use real data to validate the models. This procedure has the obvious advantage of having good ecological external validity. However, it leaves us with the epistemological position of having the test data as the sole indicator of examinee ability. Performance results, then, should be interpreted as the ability of the models to predict examinee score for the given test. If we assume that a test is a true reflection of ability, then we can extend the interpretation of the models' performance as a measure of their accuracy to predict examinee ability.

6.3 Test Data

The simulations are performed on two sets of data:

- UNIX test - a 34 items test of the knowledge of UNIX shell commands administered to 48 examinee,
- FLC test - a 160 items test of French language administered to 41 examinees.

The first test is taken from Desmarais et al. (1996) and it assesses a wide range of knowledge of the UNIX commands, from the simple knowledge of 'cd' to change directory, to the knowledge of specialized maintenance commands and data processing (e.g. 'awk', 'sed'). The second one is a test from *Formation Linguistique Canada (FLC)*. It is designed by linguistic professionals and covers a wide range of language skills.

Mean scores for the UNIX and French language tests are respectively 53% and 57%, and mean standard deviation per examinee for both test is about 0.5. Figure 6.5 illustrates the dispersion of scores for each test.

A wide distribution of scores is necessary for the proper calibration of both POKS and the IRT-2PL model. To avoid sampling bias error, all calibrations of the models's parameters are done on $N-1$ data cases: we remove from the data set the examinee for which we conduct the simulation. As a result, simulations are conducted with parameters calibrated from 47 data cases for the UNIX test and 40 data cases for the French language test.

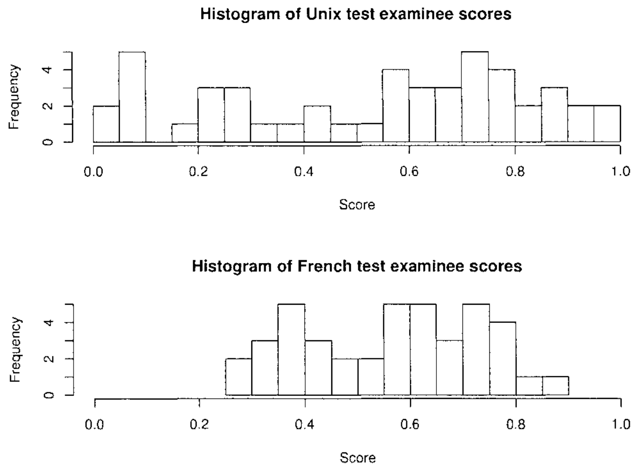

Figure 6.5: Histogram of examinee scores for each test

6.4 Parameters Estimation

The discrimination and difficulty parameters a and b were estimated with a maximum log-likelihood estimator package of the R application (Venables, Smith, & the R Development Core Team, 2004) over the two data sets. These same parameters are shared by the IRT-2PL and POKS approaches. They are used for computing $P(u_i = 1|\theta)$ in IRT 2-PL model, and for computing the Fisher information for the choice of the next item with both the POKS and the IRT-2PL approaches.

The program of IRT ability estimation $\hat{\theta}$ is written in C++ and takes advantage of the availability of C libraries of numeric routines for equation solving (Press, Teukolsky, Vetterling, & Flannery, 1992).

It is noteworthy to mention our attempt of using 3-PL model in item parameter estimation. We met serious divergence problem for 3-PL model (see Section 3.5.2) in numeric computing process. Therefore, we use 2-PL model in this study. We also used some means to bound ability and item parameters within a certain range, so that it limits the chance of occurence for Heywood case as mentioned in Section 3.5.2.

According to Baker (1992), MLE yields better results when the size of sample data set (both the number of examinees and questions) is relatively large (see Section 3.5.2). Due to the lack of big sample size in this study, the quality of item parameter estimation may be compromised. However, it cannot be verified at present. The replication of sample data, e.g. duplicating 48 examinees' responses in the UNIX test and making a forged data set of 96 examinees, makes no contribution to the improvement of MLE results. Because the values of parameters that fit into the MLE equations established from 48 examinees also fit those equations from 96 examinees.

6.5 Graph Structures and Statistics

Statistics on the graph structures inferred are given in Table 6.2. The number of relations reported represent the average over the 48 and 41 networks that were used for the simulations (one per simulation to avoid the over-sampling bias). Note that symmetric relations are in fact two directed relations between a pair of nodes (dividing the numbers by two gives the actual individual symmetric relations). Note also that, when counting transitive relations, groups of nodes linked through symmetric relations are merged into one single node[15], to avoid cycling, and that the numbers represent the transitive relations actually induced by the algorithm (not the relations that can be derived through transitivity).

Table 6.2: Graph statistics averaged over all N structures

	UNIX graph	French language graph
Total number of relations	587	1160
Symmetric relations	252	131
Transitive relations	229	668
α_c	0.25	0.10
p_c	0.50	0.50

[15] Merging nodes of symmetric relations into one is only for the purpose of counting transitive relations and not for performing inferences.

The minimal conditional probability, p_c, for both tests networks is the same, 0.5. The values for α_c and α_i are 0.25 for the UNIX data set and 0.10 for the French one. The choice of $\alpha_c = 0.10$ for the French language test proved to be more reliable during the preliminary testing. However, values of α_c ranging from 0.2 to 0.5 showed little effect on the results for the UNIX data set, but performance degradation started to appear around $\alpha_c = 0.1$.

6.6 Computational Resources

Computational resources for building the graph structure and performing inferences is often an issue for operational systems and thus we report some indicators here. For our test data, time for constructing a graph structure with the Unix and French language data set is very fast: less than $10ms$ on a standard $1.5Ghz$ PC. Inferences for CAT is also fast. We find that a full cycle involving (1) the update of item probabilities and (2) determining the next question to ask, varies from $0.03ms$ for the UNIX test with the Fisher information condition, to a much longer $106ms$ for the French language test under the information gain condition. The information gain condition is much slower because it involves simulating correct and wrong responses to every other test item to find the expected entropy. Moreover, the high number of connections in the French language network significantly affects the time to compute the entropies for the information gain technique.

We implemented the IRT-2PL framework based on Birnbaum's paradigm of examinee ability and item parameters estimation (Baker, 1992). The numeric computing program was written in C/C++ and the R language. We used a number of numerical programming techniques to improve the quality of model fit. The item parameter estimation for 2-PL IRT is a relative heavy computing. It is implemented in R language (Venables et al., 2004). Since the program emits many debugging information in execution, the performance of item parameter estimation part has not been measured solely. The speed of item parameter estimation is not a big issue in this study because it is outside the CAT loop where fast responsiveness is one of the critical constraints. The ability estimation for 2-PL IRT is implemented in C/C++ to explore its speed limit since the ability estimation process is inside the CAT loop. The time of computing ability increases as more item responses are observed. A rough estimate of ability estimation time on $866Mhz$ PC is, $1{\sim}2s$ for 34 items in UNIX test data, and $3{\sim}6s$ for 160 items in French language test data.

6.7 Item Selection

It is essential to select the next item presented to an examinee with regard to his already observed performance. Many methods have been proposed for this task. IRT generally uses item information/Fisher information as a criterion in item selection. Eggen (1998) investigated the impact of different item selection strategies on termination condition in IRT context. POKS uses an entropy based item selection method described in Section 5.2. Both Fisher information based and entropy based item selection methods are experimented in our simulation.

Item selection methods based on IRT Fisher information have strong relation to optimal estimate (see Section 7.2 and Section 7.2). The most popular item selection method based on Fisher information in IRT is to select the item that has maximum item information function/Fisher information $I_i(\theta)$ at current ability estimate $\hat{\theta}$.

As estimating current ability level is a nontrivial task in CAT, for those circumstances that only fail/pass decision is involved, it is preferable to have a speedy item selection method that does not rely on the ability estimate every time. Spray and

Reckase (1994) have shown that in a classification problem with two categories (e.g. pass/fail) where SPRT procedure is usually being used (Wald, 1947), it is more efficient to select the items which have maximum information at a fixed cut point θ_c rather than at current ability estimate.

In the comparison of IRT 2-PL and POKS models, we experimented two Fisher information based item selection methods, i.e. maximum Fisher information at current ability estimate (F1), and maximum Fisher information at fixed cut point (F2).

The two Fisher information based item selection methods (F1 and F2) performed roughly equally well for the IRT 2-PL model in the simulation. Therefore, the prevalent one, Fisher information at current ability estimate (F1) is reported in this study. For the fairness of comparison, POKS also uses the same one (F1). The technique of applying Fisher information to POKS is described in Section 5.2.

An alternative to IRT Fisher information method is to base item selection process on the information gain. The principle of this methods is to choose the item that will maximize the expected reduction of entropy of the test. The procedure is described in Section 5.2.

Contrary to the success of adapting IRT's Fisher information based item selection methods into POKS, the idea of migrating entropy based item selection methods into IRT met some difficulties in our study. The major problem is the the computational resources required for computing every expected test entropy $E_i(H'_T)$ among all the item candidates. The process of computing expected test entropy $E_i(H'_T)$ requires two new estimates of ability level (see Section 5.2). One is the new ability estimate when the response to this item candidate is 1 (success), and the other for 0 (fail). Since the ability estimation process for IRT (root finding process) is relatively expensive (see Section 6.6), so does the case of computing $E_i(H'_T)$.

Due to the fact that the number of item candidates that need to be evaluated decreases in the test, the workload of selecting an item in the beginning of a test is heavier than that near the end of a test. The latency due to the intensive workload of item selection in the beginning of a test may become an obstacle to the use of interactive application in practice. As a result, we did not report entropy based item selection method for IRT-2PL approach in this study.

A Random item selection method is also investigated in our study. Any measurement for random item selection method, e.g. classification accuracy, is the average one over a bunch of simulation runs.

6.8 Performance Metrics

Measuring the performance of each approach is based on a simple metric: the proportion of correctly classified examinees after each number of responses to test items. Classification of a single examinee is determined by comparing the examinee's estimated score S (see Equation 6.2) with the passing score S_c.

The percentage of correctly classified examinees is reported as a function of the number of test item responses given. Figure 6.6 illustrates an example of a performance graph. The curve starts at 0 item, i.e. before any items are given, at which point we use the sample average to initialize the probabilities of each test item. Each examinee will thus start with an estimated $\hat{\theta} = \bar{x}$, the sample average score in percentage points. If the sample average is above θ_c, all examinees will be considered master, otherwise they are considered non-master. As a consequence, the performance at $0th$ item generally starts around 50% when the cut score is around the sample average, and gradually reaches 100% at the end of the test when all items are observed. As the cut score departs from the average, the $0th$ item initial performance (or "blind score") increases and eventually reaches 100% if everyone is above or below this score in the sample. For example, at a cut score of 80% this initial score is 40/42 for the French language test because only two examinees score above this level and we start with the estimate that everyone is a non master.

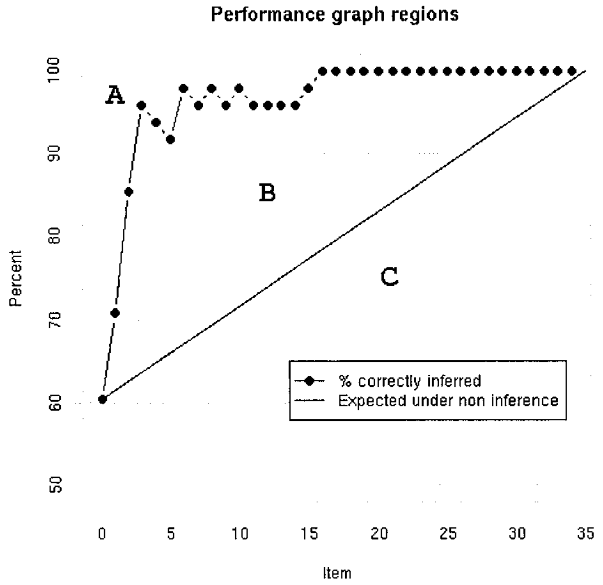

Figure 6.6: Example of metrics G

The diagonal line in Figure 6.6 represents a baseline performance used in measuring a global score, G (see below). Thus, region C of Figure 6.6 represents a linear approximation of the "given facts" (i.e. the proportion of examinees that are are correctly classified due to gradual observation of responses of test items), region B represents the "correct inferences" (i.e. the proportion of examinees correctly classified by the inference method), and region A represents "wrong inferences" (i.e. the proportion that are still incorrectly classified).

Besides graphically reporting the classification results, a single scalar metric, G, is defined for characterizing the performance over a complete simulation run. It corresponds to the ratio of surfaces $B/(A+B)$ in Figure 6.6 and is computed by:

$$G = \sum_{i=1}^{k} \frac{C_i - C_{ei}}{n - C_{ei}} \tag{6.3}$$

where n is the number of examinees, k the number of items, C_i is the number of correctly classified examinees after i number of item responses (the line with black circles), and C_{ei} the expected number of examinees correctly classified by sole observation of test items (i.e. the diagonal line in the performance figures). G values can range from $\frac{-k}{2n/[nC_{e0} - 1]}$ to 1, where C_{e0} is the number of correctly classified examinees before any response is given). A value of 1 represents a perfect classification throughout the simulation, a value of 0 indicates no gain over observation only, and a negative value indicates a worst classification than that obtained by combining the 0 item initial classification with the given responses.

6.9 Results

6.9.1 Simulations at $\theta_c = 60\%$

The simulation results for the cut score $\theta_c = 60\%$ are summarized in Figure 6.7 and Figure 6.8 for the UNIX and French Language tests. They show the number of correctly classified examinees as a function of the number of items asked. For better visibility, the French language test data points are plotted every 4 items. Both the information gain and the Fisher Information item selection methods are reported for the POKS model. However, for IRT-2PL approach, only the Fisher Information function is given because of limitations with the IRT simulation program we are using (see Section 6.7).

The simulation shows that both POKS and IRT-2PL approaches yield relatively good classification after only a few item responses, especially considering the low number of data cases used for calibration. In the UNIX test, all approaches reach more than 90% correctly classified between 5 and 10 item responses. However, for the French language test, only the POKS information gain and POKS Fisher Information approaches stays above 90% correct classification after about 20 items, whereas the IRT approach requires about half of the 160 test items to reach and stay above the 90% score. At this 60% passing score, we can conclude that the POKS Information gain approach performs better in general than the two others but, as we see later, this advantage is not maintained for all different cut scores.

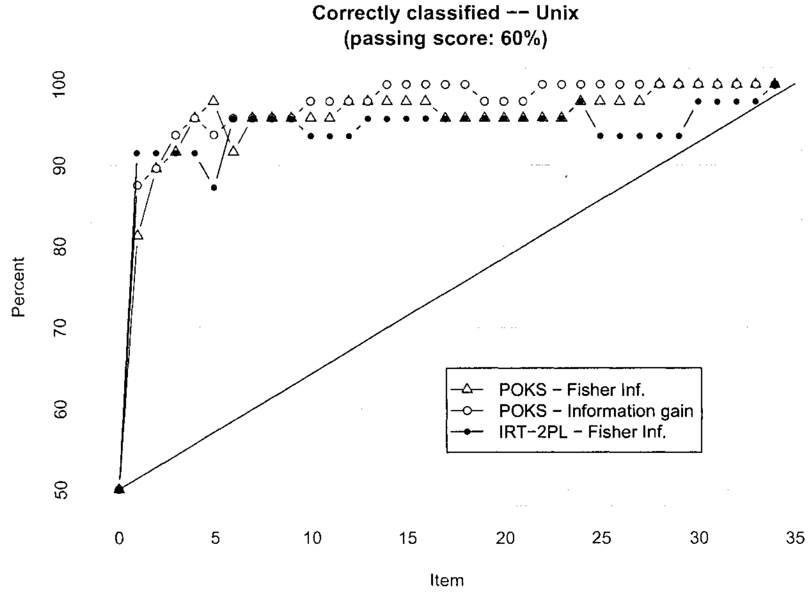

Figure 6.7: Classification accuracy of UNIX test

6.9.2 Performance under different θ_c and item selection strategies

To gain an idea of the general performance of POKS under different conditions, we investigate the following variations:

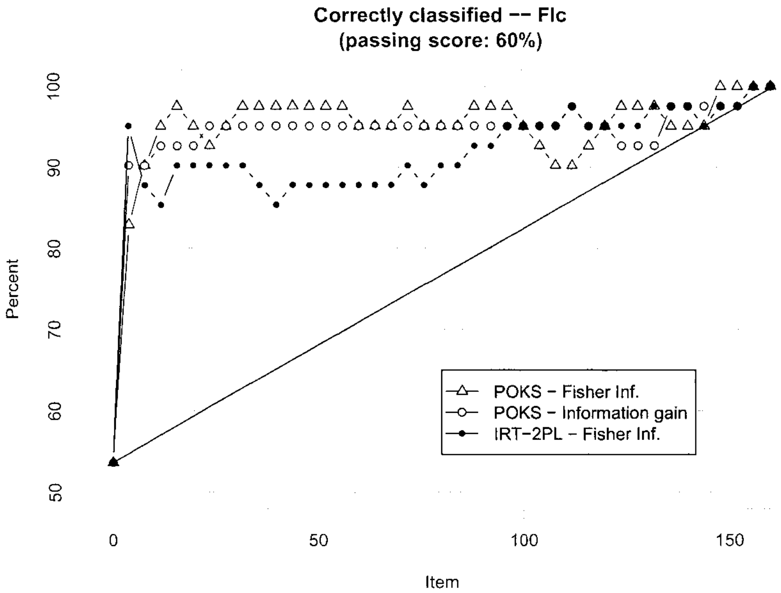

Figure 6.8: Classification accuracy of French language test

- Different cut score, from 50% to 70%[16],
- Item selection strategies, including a random selection of items,
- Two different values of the α_c and α_i parameters for inequalities Equation 4.2 and Equation 4.3 (we set $\alpha = \alpha_c = \alpha_i$). One set at $\alpha = 0.15$ for all conditions, and another one tailored for each test.

Table 6.3 summarizes the results of the simulations under these different conditions. The random selection represents the average of 9 simulation runs for each cut score. We use the G metric for reporting the performance of a whole simulation, from the first to the last test item, into a single scalar value.

[16] Scores above 70% and below 50% are not reported because the large majority of examinees are correctly classified initially (as one can tell from Figure 6.5) and little performance gain is possible (at least for the French test). Reporting scalar values within these ranges becomes misleading.

Table 6.3: Performance comparison of the G metric under different conditions of cut cores θ_c and POKS α values.

	IRT	POKS				
θ_c	Fisher	Fisher		Information		Random 95%
	information	information		Gain		confidence interval
UNIX test		α=0.25	α=0.15	α=0.25	α=0.15	α=0.25
50%	0.93	0.85	0.89	0.86	0.84	0.75(±4.4)
60%	0.81	0.92	0.89	0.86	0.84	0.77(±4.8)
70%	0.75	0.80	0.81	0.68	0.67	0.50(±7.1)
average	0.83	0.86	0.86	0.80	0.78	0.67(±7.1)
French test		α=0.10	α=0.15	α=0.10	α=0.15	α=0.10
50%	0.81	0.72	0.80	0.80	0.85	0.64(±2.0)
60%	0.68	0.78	0.79	0.74	0.79	0.57(±5.7)
70%	0.69	0.60	0.60	0.83	0.74	0.48(±4.9)
average	0.73	0.70	0.73	0.79	0.79	0.54(±7.1)

Three item selection techniques are reported for the POKS approach (information gain, Fisher information, and random item selection with a 95% confidence interval), whereas only the Fisher information technique is reported for the IRT framework, which is the most commonly used.

The G metric at the 60% level reflects that POKS has a slight advantage over IRT and that all approaches perform better for the UNIX test than the French language test. However, the POKS advantage is not systematic for all cut scores and across the two item selection techniques. The averages of cut scores across the tests suggest a relatively similar performance between POKS and the IRT model. The average score advantage is inverted between POKS Fisher information and POKS information gain, but exploration with different statistical parameters for the statistical tests of Equation 4.2 and Equation 4.2 (not reported here) indicates that this inversion is not systematic. All methods perform better than a random selection of items, as expected.

There is a noticeable decrease of performance for POKS at the 70% cut score where the Fisher information method score drops to 60% for the French test, and also drops for the UNIX test, but this time over the information gain method. This suggests that POKS may suffer weaknesses at boundary conditions.We link these results to a known problem with POKS that is further discussed in Section 7.1.3

6.9.3 Question predictive accuracy

The comparison of IRT and POKS is also conducted at the question level. In the previous sections, we assessed how accurate each method is at classifying examinees as master or non master according to some passing score. The question predictive evaluation is a more fine grained assessment of the ability of each method to predict the outcome of each individual question item. In principle, the ability of an approach to predict individual item outcome offers a means for detailed student assessment, provided that individual skills and concepts can be related to specific items.

The measure for the question accuracy score is relatively simple. It consists in the ratio of correctly predicted item outcome and it is reported as a function of the number of items administered. For both methods, the probability of success of an item, $P(X_i)$, is continuously reassessed after each item posed. If that probability is greater than 0.5, then the predicted outcome is for that item is a correct response, otherwise it is an incorrect response. Predicted responses are then compared with real responses for measuring their accuracy. Once an item is administered, the predictive accuracy score is considered 1 for that item and, as a consequence, the question predictive ratio always reaches 1.0 after all items are administered. All items are treated with equal weights.

Figure 6.9 and Figure 6.10 report the question predictive accuracy score for both tests. Only POKS information gain approach was investigated for this experiment. In addition to the two approaches, IRT-2PL and POKS, a third line is also displayed, "Fixed". It represents the score for the simple method of choosing the most uncertain

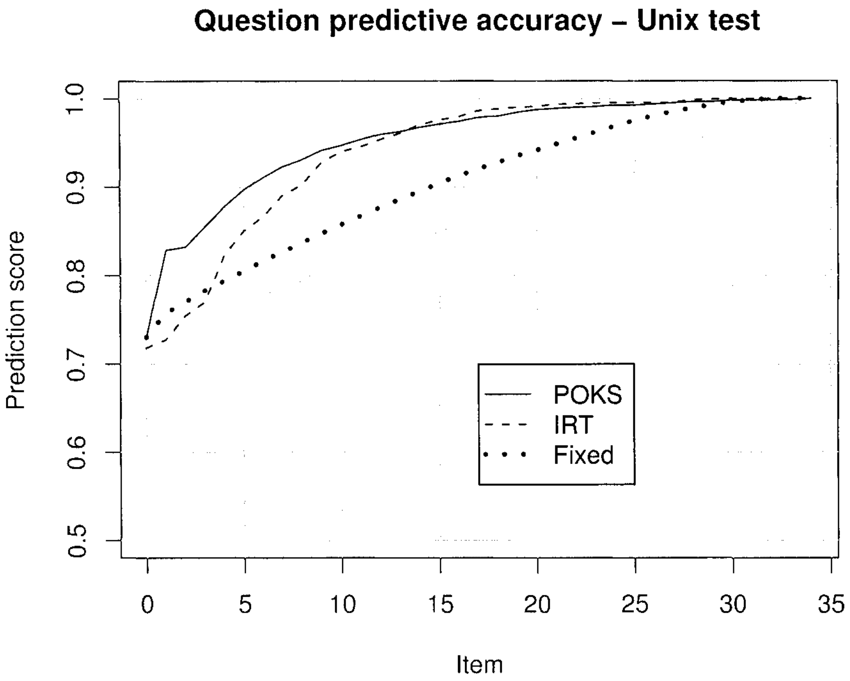

Figure 6.9: Question predictive accuracy of UNIX test

item remaining, i.e. the item whose sampled success ratio (the percentage of examinees that succeeded on this item) closest to 0.5. This method is non adaptive: The sequence of items is fixed for all examinees. It serves as a baseline comparison. We note that the IRT approach starts at a lower score than the other two. This is due to the fact that the items probabilities, $P(X_i)$, is computed from the initial θ and that value turns out to be less accurate than taking the initial probabilities calibrated from the sample.

The standard deviations of the question predictive accuracy ratio is given in figure Figure 6.11. They are also reported as a function of the number of items administered and for the three corresponding curves of Figure 6.9 and Figure 6.10.

The obvious finding is that POKS clearly outperforms IRT in the French language test, whose performance does not even match that of the fixed sequence method.

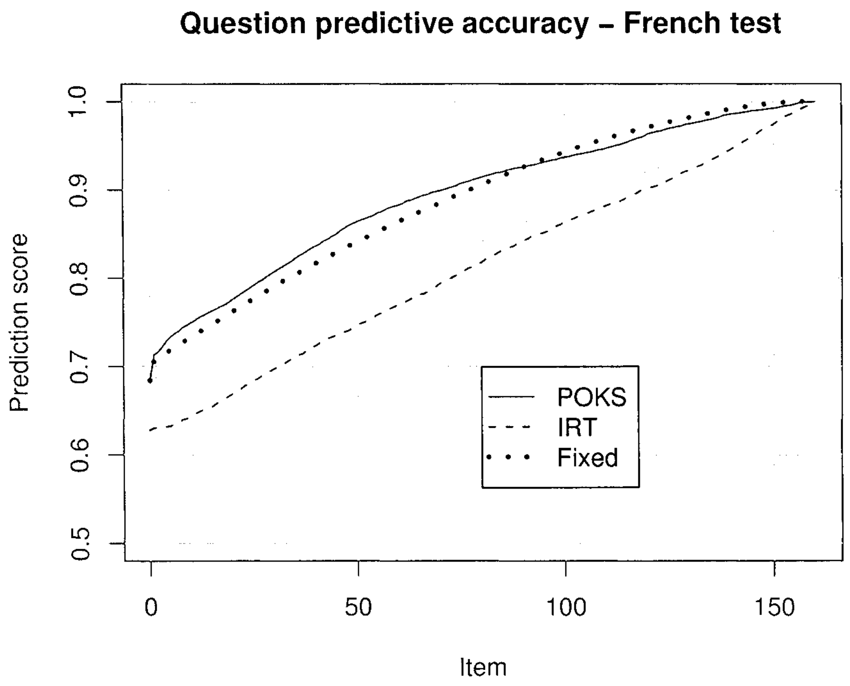

Figure 6.10: Question predictive accuracy of French language test

However, it is not significantly better than the fixed sequence one. For the UNIX test, POKS advantage over IRT is only apparent before the 10th item, but it does perform well above the fixed sequence, contrary to the French test. We also note a slight but systematic decrease across both tests of POKS performance after half (UNIX test) or two thirds (French test) of items observed. We return to this observation in Section 7.1.3.

These results confirm that the French test does not lend itself as well to adaptive knowledge assessment as does the UNIX test. This could be in part due to the smaller sample size (41 instead of 48 for the UNIX test), but it is also very likely due to the sampling distribution that is not as wide as UNIX sample (see Figure 6.5). The wider is the range of abilities, the easier it is to assess someone's knowledge from that sample[17].

The low performance of IRT for the French test is not necessarily a surprise since IRT never claims to lend itself to fine grained and multidimensional skill assessment. However, it is a surprise that it can provide a good performance for the UNIX test, possibly because that test is more unidimensional than the French test, and also because the item success ratio distribution has a wider range. Obviously, an interesting followup would be to verify if a MIRT approach could yield better results for that test.

Nevertheless, the comparison does demonstrate that POKS has the potential of providing more fine grained assessment, if we assume that question predictive accuracy is more fine grained. For example, by segmenting tests items according to different skills, then individual question item prediction could provide useful information on individual skills. More investigation is required to confirm this hypothesis but, this is an early indication that supports it.

[17] For example, a sample whose test score varies only within 5% would be very insufficient since most people are of the same ability and even the test prediction itself may not be that accurate.

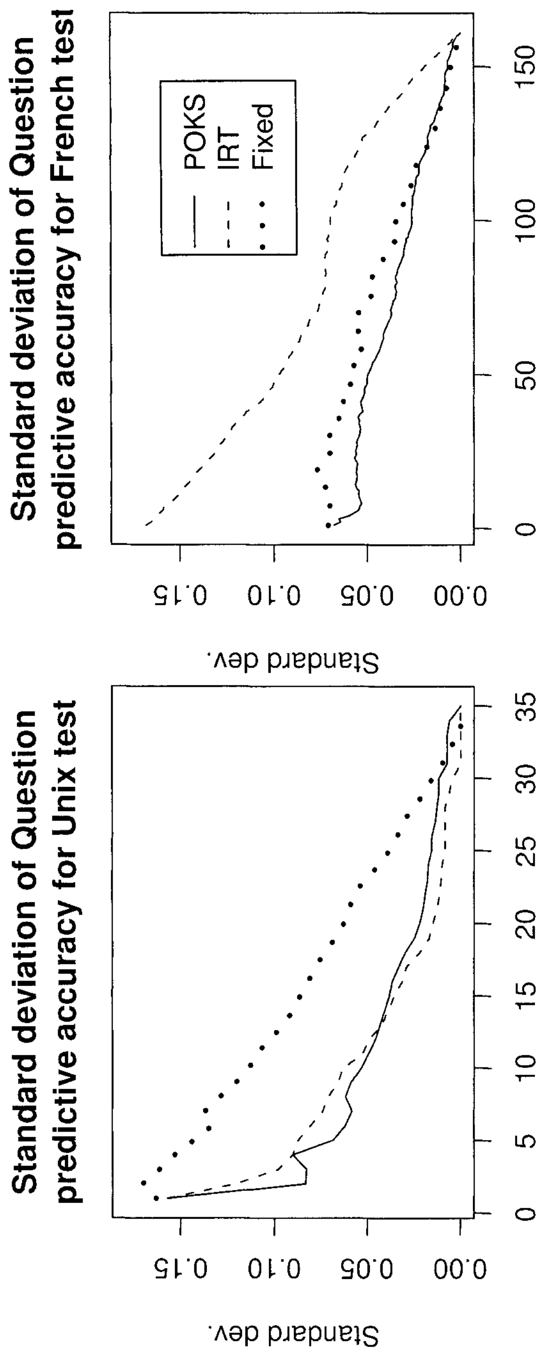

Figure 6.11: Standard deviations of each test as a function of the number of test item administered

CHAPTER 7

DISCUSSION AND CONCLUSION

7.1 Discussion

The comparison of the POKS approach with the IRT-2PL one shows that they both can perform correct classification of examinees under different tests and passing scores, although their performance differs according to that passing score. However, their ability to predict individual question item outcome varies considerably between themselves and also between tests. POKS can predict item outcome well above that of the fixed sequence performance for the UNIX test, but only at about the same level for the French test. The performance of IRT is distinctively lower for IRT over both test, but it does perform as well as POKS for the UNIX test after the tenth question.

Given that the POKS approach is computationally much simpler than the IRT approach, which relies on maximum-likelihood estimation techniques that can prove relatively complex for IRT (see Baker (1992) and Section 2.3.4), these results show the POKS approach has a good potential.

However, beyond these findings, a number of issues remain. We discuss some of the major ones below.

7.1.1 Including concepts in POKS structures

Estimating the mastery of concepts is essential to provide the fine grained knowledge assessment that many learning environments often require. The POKS approach must provide some means of assessing concepts.

Like other graph modeling techniques, we can include concepts and skills within POKS structures. However, the manner in which items should be linked to concepts, and how concepts should be linked among themselves is an open question. Numerous schemes have been proposed for linking items to concepts, such as leaky-*AND*/*OR* gates (Martin & Vanlehn, 1995; Conati et al., 2002), dynamic Bayesian networks(Mayo & Mitrovic, 2001), weighted means (Millán & Pérez-de-la-Cruz, 2002), or BN organized in a number of topologies as can be found in Vomlel (2004b).

For POKS structures, one possibility is to keep the whole structure uniform and link concepts with surmise relations as is done with items POKS the and knowledge structures framework (Falmagne et al., 1990). For example, mastery of a concept by examinees can be independently assessed and the induction of the structure can proceed in much the same process as that described in Section 4.3. Preliminary exploration of this simple scheme seems to suggest that the results are not very positive and further investigation is necessary to confirm this hypothesis and determine why.

Another approach was recently explored using the data from Vomlel (2004b) (see Section 2.2) and deriving concept mastery with a single layered network comprising items as predictors (Desmarais & Meshkinfam, 2005; Meshkinfam, 2005). Concepts had already been assessed independently. The POKS updating algorithm serves to determine the probability of mastery of each item as items are answered, and the new probabilities are, in turn, fed to logistic regression models to determine concept mastery. The approach is compared to Vomlel's own predictions. The results show that although POKS is better than the BN constructed by Vomlel for predicting answers to question items, it is less accurate to predict concept mastery.

Finally, the simplest way of including concepts into POKS is to use the traditional breakdown that teachers do. Subject matters are divided into a hierarchy of more and more specific topics. Items are the leaves of this hierarchy and a weighted mean is used to estimate mastery of the next level down. Note that the structure does not need to be a pure hierarchy and that a single item can belong to many concept/skill nodes. Exams are often structured this way. The accuracy of this method is directly linked to the accuracy of the leave nodes mastery estimates (test items) and the validity of the weighted means. This approach may not have the ability to model misconceptions and non linear combinations of items and concepts, but it has the quality of being universally used in schools and understood by everyone.

Furthermore, that approach avoids problem of estimating concepts independently for constructing a graph model and for calibrating conditional probabilities. In fact, in our view, that problem plagues graph models in general. Modeling with hidden nodes is very difficult to envision by non statisticians. The multidimentional-IRT model is also subject to this issue.

7.1.2 Automated learning constraint

POKS is an algorithmic learning/calibration approach. Structures such as Figure 4.4 are built automatically. It shares the same advantages as IRT in that respect. However, as a graphical modeling approach, it also has the expressiveness of these models, namely that items can be aggregated into concepts and further aggregated into hierarchies of concepts. Techniques such as those of VanLehn, Niu, Siler, and Gertner (1998) can be used for modeling the "concept" part of the network that stands above the test items. Alternatively, a concept can be defined as a function of the probability of mastery of a set of items. For example, it can be a weighted average of the probability of set of items which composes a concept, as (Millán & Pérez-de-la-Cruz, 2002) did.

7.1.3 POKS's sensitivity to noise

One of the critical issue with the POKS approach is the problem of correcting errors due to noise. This is a direct consequence of pruning the bi-directionality of posterior updates, and that can result in nodes having no incoming, or outgoing links. For example, a difficult item can often have many outgoing links, but no incoming links (i.e. no other item's success significantly increases its probability). It follows that this node's probability can only decrease according to POKS updating scheme. If, for some reason, an expert misses an easy item, these items with no incoming links (the more difficult ones in general) will see their probability decrease with no chance of being raised later on, until they are directly observed. Over test with a large number of items, such noisy answers are bound to happen and create these sticky errors. They will also tend to affect more significantly the performance at the end of the test when only a few items are not yet observed.

This weakness can explain the poor result found at the 70% cut score for the French language test (see Table 6.3), as it is a relatively large test of 160 items. This explanation is also consistent with the fact that, as the cut score nears the edges of the examinees's performance distributions (see Figure 6.5), small errors have more weight on the G score and their weight increases as the test nears the end.

Moreover, it can also can explain the decrease in both tests for the question predictive experiment. We can see in Figure 6.9 and Figure 6.10 that the performance of POKS decreases relative to the fixed sequence performance. The decrease is apparent after a third (UNIX test) to two thirds (French test) when the POKS performance drops below the fixed sequence performance. Again, this is consistent with the fact that sticky errors will accumulate and more significantly affect the performance at the end of a test.

This is not an insurmountable problem, but it does involve developing some means to avoid the accumulation of noise over items that are either very difficult or very easy.

7.2 Conclusion

POKS offers a fully algorithmic means of building the model and updating item probabilities among themselves without requiring any knowledge engineering step. Indeed, the specific POKS approach uses the same data as the IRT-2PL approach to provide similar accuracy. It shows that a graphical modeling modeling approach such as POKS can be induced from a small amount of test data to perform relatively accurate examinee classification. This is an important feature from a practical perspective since the technique can benefit to a large number of application contexts.

The graphical modeling approaches such as POKS or as Bayesian networks are still in their infancy compared to the IRT techniques developed since the 1960s, and their potential benefit remains relatively unexplored. However, applications of CAT techniques to tutoring systems and to different learning environments are emerging. The availability of simple and automated techniques that are both effective and efficient, relying on little data and allowing seamless updates of test content, will be critical to their success in commercial applications.

Unlike the conventional use of IRT in CAT industry which usually have thousands of examinees in sample data set, the IRT model (2-PL) used in this study is limited to small sample data set. In the occasion where only limited sample data is available, such as UNIX and French language test in this study, IRT is still a workable model. However quality of item parameter estimation may be compromised.

REFERENCES

BAKER, F. B. (1992). *Item Response Theory: parameter estimation techniques.* New York: M. Dekker.

BIRNBAUM, A. (1968). *Some latent trait models and their use in inferring an examinee's ability*, Chap. 17-20, pp. 397–479. Reading, MA: Addison-Wesley.

BOCK, R. D., & AITKIN, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, (46), 443–459.

BOCK, R. D., & LIEBERMAN, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, (35), 179–197.

BRUSILOVSKY, P., EKLUND, J., & SCHWARZ, E. (1997, Sydney, australia, July 14–18). Adaptive navigation support in educational hypermedia on the world wide web. *INTERACT97, 6th IFIP World Conference on Human-Computer Interaction* (pp. 278–285).

BRUSILOVSKY, P., EKLUND, J., & SCHWARZ, E. (1998, 14–18). Web-based education for all: A tool for developing adaptive courseware. *Proceedings of Seventh International World Wide Web Conference* (pp. 291–300). Brisbane, Australia.

BRUSILOVSKY, P., SCHWARZ, E., & WEBER, G. (1996). A tool for developing adaptive electronic textbooks on WWW. *Proceedings of WebNet'96 - World Conference of the Web Society* (pp. 64-69).

CHANG, H. H., & YING, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213–229.

CHENG, J., GREINER, R., KELLY, J., BELL, D., & LIU, W. (2002). Learning

Bayesian networks from data: An information-theory based approach. *Artificial Intelligence, 137*(1–2), 43–90.

COLLINS, J. A. (1996). Adaptive testing with granularity. Master's thesis, University of Saskatchewan, Department of Computer Science.

COLLINS, J. A., GREER, J. E., & HUANG, S. X. (1996). Adaptive assessment using granularity hierarchies and Bayesian nets. *Intelligent Tutoring Systems* (pp. 569–577). Montreal, Canada.

CONATI, C., GERTNER, A., & VANLEHN, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction, 12*(4), 371–417.

CRAMER, H. (1946). *Mathematical methods of statistics.* Princeton University Press.

DEMPSTER, A. P., LAIRD, N. M., & RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*(39), 1–38.

DESMARAIS, M. C., FU, S., & PU, X. (2005). Tradeoff analysis between knowledge assessment approaches. *Proceedings of the 12th International Conference on Artificial Intelligence in Education, AEID'2005* (pp. 209–216). Amsterdam.

DESMARAIS, M. C., MALUF, A., & LIU, J. (1996). User-expertise modeling with empirically derived probabilistic implication networks. *User Modeling and User-Adapted Interaction, 5*(3-4), 283–315.

DESMARAIS, M. C., & MESHKINFAM, P. (2005). *Bayesian modeling with strong vs. weak assumptions in the domain of skills assessment* (Technical report). École Polytechnique de Montréal.

DESMARAIS, M. C., & PU, X. (2005a). A Bayesian student model without hidden nodes and its comparison with Item Response Theory. *The International Journal of Artificial Intelligence in Education (IJAIED).* accepted.

DESMARAIS, M. C., & PU, X. (2005b). Computer adaptive testing: Comparison of a probabilistic network approach with Item Response Theory. *UM 2005 User Modeling: The Proceedings of the Tenth International Conference.* Edinburg.

EGGEN, T. J. H. M. (1998). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*, 249–261.

EGGEN, T. J. H. M. (2004). *Contributions to the theory and practice of computerized adaptive testing.* University of Twente.

FALMAGNE, J. C., KOPPEN, M., VILLANO, M., DOIGNON, J. P., & JOHANNESEN, L. (1990). Introduction to knowledge spaces: How to build test and search them. *Psychological Review, 97*, 201–224.

FERGUSON, G. A. (1942). Item selection by the constant process. *Psychometrika, 7*, 19–29.

FINNEY, D. J. (1944). The application of probit analysis to the results on mental tests. *Psychometrika, 19*, 31–39.

GIARRATANO, J. C., & RILEY, G. (1998). *Expert systems: Principles and programming (3rd edition).* Boston, MA: PWS-KENT Publishing.

HALEY, D. C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error* (Stanford, Calif.: Contract No. ONR-25140 15). Applied Mathematics and Statistics Laboratory, Stanford University.

HAMBLETON, R. K., SWAMINATHAN, H., & ROGERS, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, Calif.: Sage Publications.

HARRIS, J. W., & STOCKER, H. (1998). *Handbook of mathematics and computational science*, p. 824. New York: Springer-Verlag.

HECKERMAN, D. (1995). *A tutorial on learning with Bayesian networks* (Technical Report MSR-TR-95-06). Redmond, WA: Microsoft Research (MSR).

JAMESON, A. (1995). Numerical uncertainty management in user and student modeling: An overview of systems and issues. *User Modeling and User-Adapted Interaction*, *5*(3-4), 193–251.

JENSEN, F., KJÆUL, U. B., LANG, M., & MADSEN, A. L. (2002, 6–8). Hugin - the tool for Bayesian networks and influence diagrams. In J. A. Gámez, & A. Salmeron (Eds.), *Proceedings of the First European Workshop on Probabilistic Graphical Models, PGM 2002* (pp. 211–221).

JIAO, H., & KAMATA, A. (2003). Model comparison in the presence of local item dependence. *Annual Meeting of the American Educational Research Association.*

KAMBOURI, M., KOPPEN, M., VILLANO, M., & FALMAGNE, J.-C. (1994). Knowledge assessment: tapping human expertise by the query routine. *International Journal of Human-Computer Studies*, *40*(1), 119–151.

KHUWAJA, R., DESMARAIS, M., & CHENG, R. (1996a). Intelligent guide: Combining user knowledge assessment with pedagogical guidance. *ITS '96: Proceedings of the Third International Conference on Intelligent Tutoring Systems* (pp. 225–233). London, UK: Springer-Verlag.

KHUWAJA, R., DESMARAIS, M., & CHENG, R. (1996b). Intelligent guide: Combining user knowledge assessment with pedagogical guidance. *Lecture Notes in Computer Science*, *1086*, 225–232.

LAZARSFELD, P. F. (1959). *Psychology: A study of a science, vol. 3*, Chap. Latent Structure Analysis. New York: McGraw-Hill.

LEWIS, C., & SHEEHAN, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*(2), 367–386.

LIU, J., & DESMARAIS, M. (1997). A method of learning implication networks from empirical data: Algorithm and Monte-Carlo simulation-based validation. *IEEE Transactions on Knowledge and Data Engineering, 9*(6), 990–1004.

LORD, F. M. (1980). *Application of Item Response Theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

LORD, F. M., & NOVICK, M. R. (1968). *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley Publishing Company. with contributions by Birnbaum, A.

MARTIN, J., & VANLEHN, K. (1995). Student assessment using Bayesian nets. *International Journal of Human-Computer Studies, 42*(6), 575–591.

MAYO, M., & MITROVIC, A. (2001). Optimising ITS behaviour with Bayesian networks and decision theory. *International Journal of Artificial Intelligence in Education, 12*, 124–153.

MCDONALD, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern Item Response Theory* (pp. 257–286). Springer-Verlag, New York.

MESHKINFAM, P. (2005). Bayesian modeling with strong vs. weak assumptions in the domain of skills assessment. Master's thesis, École polytechnique de Montréal.

MILLÁN, E., & PÉREZ-DE-LA-CRUZ, J.-L. (2002). A Bayesian diagnostic algorithm for student modeling and its evaluation. *User Modeling and User-Adapted Interaction*, *12*(2–3), 281–330.

MILLÁN, E., TRELLA, M., PÉREZ-DE-LA-CRUZ, J.-L., & CONEJO, R. (2000). Using Bayesian networks in computerized adaptive tests. In M. Ortega, & J. Bravo (Eds.), *Computers and education in the 21st century* (pp. 217–228). Kluwer.

MISLEVY, R. J., & CHANG, H. H. (2000). Does adaptive testing violate local independence? *Psychometrika*, *65*, 149–156.

MISLEVY, R. J., & GITOMER, D. (1995). The role of probability-based inference in an intelligent tutoring system. *User Modeling and User-Adapted Interaction*, *42*(5), 253–282.

NEAPOLITAN, R. E. (1998). *Probabilistic reasoning in expert systems: Theory and algorithms.* New York, NY: John Wiley & Sons, Inc.

NEYMAN, J., & SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika*, *16(1)*, 1–32.

NIST (2003). *NIST/SEMATECH e-handbook of statistical methods* (Technical report). NIST.

PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T., & FLANNERY, B. P. (1992). *Numerical recipes in c: The art of scientific computing.* New York, NY, USA: Cambridge University Press.

RECKASE, M. D. (1997). A linear logistic multidimensional model. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern Item Response Theory* (pp. 271–286). New York: Springer-Verlag.

REYE, J. (2004). Student modelling based on belief networks. *International Journal of Artificial Intelligence in Education*, *14*, 63–96.

RICHARDSON, M. W. (1936). The relationship between difficulty and the differential validity of a test. *Psychometrika*, *1*, 33–49.

RUDNER, L. M. (2002, 1–5). An examination of decision-theory adaptive testing procedures. *Proceedings of American Educational Research Association* (pp. 437–446). New Orleans.

SAMEJIMA, F. (1977). Weakly paralle tests in latent trait theory with some critcisms of classical test theory. *Psychometrika*, *42*, 193–198.

SCHWARZ, E., BRUSILOVSKY, P., & WEBER, G. (1996, Boston, ma, june 1–22). World-wide intelligent textbooks. *Proceedings of ED-TELECOM'96 - World Conference on Educational Telecommunications* (pp. 302–307).

SPRAY, J. A., & RECKASE, M. D. (1994, April). The selection of test items for decision making with computer adaptive test. *Annual meeting of the National Council on Measurement in Education.* New Orleans, LA.

TERMAN, L. M. (1916). *The measurement of intelligence: An explanation of and a complete guide for the use of the stanford revision and extension of the binet-simon intelligence scale.* Boston: Houghton Mifflin.

VAN DER LINDEN, W. J., & HAMBLETON, R. K. (1997). *Handbook of modern Item Response Theory.* Springer-Verlag.

VANLEHN, K., LYNCH, C., SCHULZE, K., SHAPIRO, J. A., SHELBY, R., TAYLOR, L., TREACY, D., WEINSTEIN, A., & WINTERSGILL, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence and Education*, *15*(3).

VANLEHN, K., & MARTIN, J. (1997). Evaluation of an assessment system based on Bayesian student modeling. *International Journal of Artificial Intelligence in Education*, *8*, 179–221.

VANLEHN, K., & NIU, Z. (2001). Bayesian student modeling, user interfaces and feedback: A sensitivity analysis. *International Journal of Artificial Intelligence in Education*, *12*, 154–184.

VANLEHN, K., NIU, Z., SILER, S., & GERTNER, A. S. (1998). Student modeling from conversational test data: A Bayesian approach without priors. *ITS'98: Proceedings of the 4th International Conference on Intelligent Tutoring Systems* (pp. 434–443). London, UK: Springer-Verlag.

VENABLES, W. N., SMITH, D. M., & THE R DEVELOPMENT CORE TEAM (2004). *An introduction to R, notes on R: A programming environment for data analysis and graphics* (Technical report). R Project.

VOMLEL, J. (2002). Evidence propagation in Bayesian networks for computerized adaptive testing. *12*.

VOMLEL, J. (2004a). Bayesian networks in educational testing. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, *12*(Supplementary Issue 1), 83–100.

VOMLEL, J. (2004b). Building adaptive tests using Bayesian networks. *Kybernetika*, *40*(3), 333–348.

WALD, A. (1947). *Sequential analysis.* New York : J. Wiley & sons, inc.

ZAPATA-RIVERA, J. D., & GREER, J. E. (2004). Interacting with Bayesian student models. *International Journal of Artificial Intelligence in Education*, *14*(2), 127–163.

APPENDIX

A.1 Test Information Function and MLE Sample Variance

The test information function $I(\theta)$ and MLE sample variance $Var(\hat{\theta}|\theta)$ share some common ground which is

$$I(\theta) = \frac{1}{Var(\hat{\theta}|\theta)} \quad \text{or} \quad Var(\hat{\theta}|\theta) = \frac{1}{I(\theta)}$$

There is a general theorem (see Lord, 1980, p. 70), under regularity condition, satisfied here whenever the item parameters are known from previous testing: A maximum likelihood estimator $\hat{\theta}$ of a parameter θ is asymptotically normally distributed with mean θ_o (the unknown true parameter value) and variance

$$Var(\hat{\theta}|\theta_o) = \frac{1}{E\left[\left(\frac{d\log L}{d\theta}\right)^2_{\theta_o}\right]} \tag{A.1}$$

where $L = L(\theta) = \prod_i^n P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i}$ is the likelihood function.

Where the item parameters are known, we have from Equation A.1 and Equation 3.25 that

$$\begin{aligned}
\frac{1}{Var(\hat{\theta}|\theta_o)} &= E\left\{\left[\sum_{i=1}^{n}(u_i - P_i)P_i'/P_iQ_i\right]^2 |\theta_o\right\} \\
&= E\left\{\left[\sum_{i=1}^{n}(u_i - P_i)P_i'/P_iQ_i\right] \cdot \left[\sum_{i=1}^{n}(u_i - P_i)P_i'/P_iQ_i\right] |\theta_o\right\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{P_{io}'P_{jo}'}{P_{io}P_{jo}Q_{io}Q_{jo}} E\left[(u_i - P_i)(u_j - P_j)|\theta_o\right]
\end{aligned}$$

since $E(u_i|\theta_o) = P_{io}$, the expectation under the summation sign is covariance. Because of local independence, u_i is distributed independently of u_j for fixed θ. Consequently the covariance is zero except when $i = j$, in which case it is covariance. Thus

$$\frac{1}{Var(\hat{\theta}|\theta_o)} = \sum_{i=1}^{n} \frac{P_i'^2}{P_i^2 Q_i^2} Var(u_{io|\theta_o}) = \sum_{i=1}^{n} \frac{P_i'^2}{P_i^2 Q_i^2} P_{io} Q_{io} \tag{A.2}$$

Dropping the subscript o, the formula for the asymptotic sampling variance of the maximum likelihood estimator is thus

$$Var(\hat{\theta}|\theta) = \frac{1}{\sum_{i=1}^{n} \frac{P_i'^2}{P_i Q_i}} \tag{A.3}$$

Thus the (asymptotic) test information function Equation 3.29 is the reciprocal of the asymptotic sample variance of maximum likelihood estimator of ability Equation A.3:

$$I(\theta) \equiv \frac{1}{Var(\hat{\theta}|\theta)} = \sum_{i=1}^{n} \frac{P_i'^2}{P_i Q_i} \tag{A.4}$$

Theorem A.1 *The information function for an unbiased (consistent) estimator of ability is the reciprocal of the (asymptotic) sampling variance of the estimator. (see Equation A.4)*

Theorem A.2 *The test information function $I(\theta)$ given by Equation A.4 is an upper bound to the information that can be obtained by any method of scoring the test (see proof at Lord, 1980, p. 71).*

The importance of the test information function comes partly from the fact that it provides an (attainable) upper limit to the information that can be obtained from the test, no matter what method of scoring is used.

A.2 Item Information and Fisher Information

The item information in IRT is roughly the same as Fisher information, although the originality of definitions of two information functions are different.

Fisher information is thought of as the amount of information that an observable random variable X carries about an unobservable parameter t upon which the probability distribution of X depends.

Fisher information can be written as

$$I_{Fisher}(t) = E\left[\left[\frac{\partial}{\partial t}\mathrm{log}f(X|t)\right]^2\right] \tag{A.5}$$

where $f(X|t)$ is the probability density function of random variable X, and $E[\,]$ is the expectation operation.

In IRT, Fisher information for an item i is defined as

$$I_{Fisher}(\theta) \;\; = \;\; E\left[\left[\frac{\partial}{\partial\theta}\mathrm{log}L(u_i|\theta)\right]^2\right] \tag{A.6}$$

where $L(u_i|\theta) = P_i(\theta)^{u_i}Q_i(\theta)^{1-u_i}$ is likelihood function for single item, Substitute $\frac{\partial}{\partial\theta}\text{log}L(u_i|\theta)$ with Equation 3.25, and let $P_i = P_i(\theta)$, $Q_i = 1 - P_i$:

$$\begin{aligned} I_{Fisher}(\theta) &= E\left[\left[\frac{\partial}{\partial\theta}\text{log}L(u_i|\theta)\right]^2\right] \\ &= E\left[\frac{(u_i - P_i)P_i'}{P_iQ_i}\right]^2 \\ &= \left[\frac{(1 - P_i)P_i'}{P_iQ_i}\right]^2 P_i + \left[\frac{(0 - P_i)P_i'}{P_iQ_i}\right]^2 Q_i \\ &= \frac{(P_i')^2}{(P_iQ_i)^2}\left[(1 - P_i)^2P_i + (0 - P_i)^2Q_i\right] \\ &= \frac{(P_i')^2}{(P_iQ_i)^2}P_iQ_i \\ &= \frac{(P_i')^2}{P_iQ_i} \end{aligned} \tag{A.7}$$

In the meantime, Birnbaum's definition Equation 5.1 in IRT is,

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)}$$

Therefore, $I_{Fisher}(\theta) = I_i(\theta)$, and Fisher information function is equivalent to IRT Birnbaum's information function.

For 2-PL model, the $I_{Fisher}(\theta) = I_i(\theta) = (Da_i)^2P_i(\theta)Q_i(\theta)$.

A.3 Frequently Used Formulas for 3-PL and 2-PL models

A few useful formula involving the three-parameter logistic function Equation 3.6 are recorded for convenience reference. Note that these formulas do not apply to the any normal ogive function.

$$P_i = c_i + \frac{1 - c_i}{1 + e^{(-DL_i)}} = \frac{c_i + e^{-DL_i}}{1 + e^{-DL_i}} \tag{A.8}$$

where $D \equiv 1.7, \quad L_i \equiv a_i(\theta - b_i)$

$$Q_i \equiv 1 - P_i = \frac{1-c_i}{1+e^{DL_i}} \tag{A.9}$$

$$\frac{P_i}{Q_i} = \frac{c_i + e^{DL_i}}{1-c_i} \tag{A.10}$$

$$P_i' \equiv \frac{dP_i}{d\theta} = \frac{Da_i}{1-c_i} Q_i(P_i - c_i) = \frac{Da_i(1-c_i)}{e^{DL_i} + 2 + e^{-DL_i}} \tag{A.11}$$

$$\frac{P_i'}{Q_i} = \frac{Da_i}{1+e^{-DL_i}} \tag{A.12}$$

$$I_i \equiv \frac{P_i'^2}{P_iQ_i} = \frac{D^2 {a_i}^2 (1-c_i)}{(c_i + e^{DL_i})(1+e^{-DL_i})^2} \tag{A.13}$$

$$\frac{d^2P_i}{d\theta^2} = \frac{D^2 {a_i}^2}{(1-c_i)^2} Q_i(P_i - c_i)(Q_i - P_i + c_i) \tag{A.14}$$

$$\frac{P_i - c_i}{P_i} = \frac{1-c_i}{1+c_i e^{-DL_i}} \tag{A.15}$$

For two-parameter ($c = 0$) logistic function, here are frequently referred formulas:

$$P_i = \frac{1}{1+e^{-DL_i}} = \frac{e^{DL_i}}{1+e^{DL_i}} \tag{A.16}$$

$$Q_i \equiv \frac{e^{-DL_i}}{1+e^{-DL_i}} = \frac{1}{1+e^{DL_i}} \tag{A.17}$$

$$P_i' \equiv \frac{dP_i}{d\theta} = Da_iP_iQ_i \tag{A.18}$$

$$I_i \equiv \frac{P_i'^2}{P_iQ_i} = D^2 {a_i}^2 P_iQ_i \tag{A.19}$$