ABSTRACT

The current studies represent an effort to advance the feasibility of cognitive diagnostic computerized adaptive testing (CD-CAT), an intelligent educational measurement tool that was envisioned as enhancing individualized learning over twenty years ago. Several new selection algorithms are proposed for addressing two important issues in CD-CAT: measurement efficiency and item exposure control. The posterior-weighted CDM discrimination index (PWCDI) and posterior-weighted attribute-level CDM discrimination index (PWACDI) are computationally affordable and highly efficient alternatives to other information index-based algorithms. The binary stratification algorithm offers an elegant solution to item exposure control in both fixed-length and variable-length CD-CAT, compared with the restrictive stochastic methods for fixed-length CD-CAT and SHTVOR for variable-length CD-CAT.
To Xuhui and Charlotte
ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. Hua-hua Chang, for his unwavering support and guidance over the years. And I thank members of my committee, Professor Carolyn J. Anderson, Professor Jinming Zhang, Professor Jeffrey A. Douglas, and Dr. Steven A. Culpepper, for the stimulating discussions and suggestions that led to a much improved research project.

I thank my friends and colleagues at the Department of Educational Psychology, Psychology and Statistics at the University of Illinois at Urbana-Champaign and Beijing Normal University. In particular, I would like to thank Drs. Chun Wang, Shiyu Wang and Haiyan Lin for their generous help with my research, Justin Kern and Edison Choe for numerous discussions and “chitchat”, and Poh Hua Tay for careful proofreading of my dissertation.

I thank my internship mentors, Dr. Chunyan Liu at the ACT and Dr. Oliver Zhang at AICPA for the enriching internship experience.

I thank Mr. Feng Xiao who kindled an enthusiasm for psychology in me many years ago.

Finally, I thank my extended family for the support and love, especially my parents who have toiled to support my education.
# Table of Contents

Chapter 1 Introduction ........................................................................................................... 1
  1.1 CD .................................................................................................................................. 2
  1.2 Computer-based testing techniques .................................................................................. 4
  1.3 Educational policy .......................................................................................................... 6
  1.4 Online Education (Moocs) ............................................................................................ 8
Chapter 2 A Synthesis of Literature on the Item Selection Algorithms in CD-CAT....................... 11
  2.1 Basic Algorithms Concerned with Measurement Precision and Efficiency ...................... 12
  2.2 Algorithms for specific applications .............................................................................. 19
Chapter 3 High-efficiency KL-based Item Selection Algorithms for CD-CAT............................ 37
  3.1 PWCDI and PWACDI .................................................................................................... 37
  3.2 Simulation studies .......................................................................................................... 42
Chapter 4 Linear/Binary Stratification Methods for Fixed-length CD-CAT .................................. 51
  4.1 Linear and binary stratification strategies ....................................................................... 52
  4.2 Simulation studies .......................................................................................................... 60
Chapter 5 Item Exposure Control in Variable-length CD-CAT .................................................. 67
  5.1 Item selection algorithms for variable-length CD-CAT .................................................. 67
  5.2 Simulation studies .......................................................................................................... 69
Chapter 6 Discussions and Future Directions .......................................................................... 74
REFERENCES ........................................................................................................................... 76
Chapter 1 Introduction

The subject of the current study is known as cognitive diagnostic computerized adaptive testing (CD-CAT), a marriage between cognitive diagnosis (CD) and computerized adaptive testing (CAT). Both parties in this marriage enjoy a much longer history than this union itself. The maturity of these two branches within the educational measurement and psychometrics led to the birth of CD-CAT as early as 2003. Not until after 2010 did the research on this particular topic witness progress due to the challenges and opportunities brought about by recent developments in education policies. This new phenomenal trend in online learning also offers an opportunity to demonstrate the potential of CD-CAT in providing a solution to instructional quality improvement. However, this has only taken place in the educational technology and the computer science community and related industries, and has been neglected by the testing community. For the same reason, the expertise of CD, and even testing in general has been ignored by the former.

The major goals of the current research are to (a) conduct a comprehensive review of the published literature on the item selection algorithms for CD-CAT, which is the most important element of a CD-CAT system, and (b) develop several practical item selection algorithms in order to address some important issues, such as measurement efficiency, non-statistical constraints and so on, and make CD-CAT ready for real-life operational use. The emphasis will be on how to tap the potential of CD-CAT for instructional improvement. A brief review follows regarding the development of CD and CAT, the relevant educational policies and the
growing concern concerning the issue of quality of education delivered by online learning system, which motivated the current study.

1.1 CD

CD is the union of cognitive psychology and psychometrics (Leighton & Gierl, 2007; Steinberg, 1984). Traditional testing theory pursued a behavioral interpretation of test scores. The development of the notion of test validity highlighted the need for a substantive approach to depicting the mental processes involved in obtaining test scores and this led to the birth of CD (Leighton & Gierl, 2007; Messick, 1989). As a result, the concept of construct validity became popular.

The research on CD has flourished. More than 60 cognitive diagnostic models (CDMs) have been proposed. They can generally be identified as fitting into one of the following four categories (Bolt, 2007): the Rule Space Method (Tatsuoka, 1983) and the Attribute Hierarchy Method (Gierl, 2007), discrete-skill models, continuous-skill models, and Bayesian Networking Modeling (Almond, DiBello, Moulder, & Zapata - Rivera, 2007). A comprehensive review can be found in DiBello, Roussos, and Stout (2006) and Rupp, Templin, and Henson (2010).

One common element shared by most CDMs is the Q matrix (Tatsuoka, 1995; Tatsuoka, 1990), an incidence matrix which specifies the relationship between attributes/skills and items. The construction of the Q matrix is quite labor-intensive and is usually accomplished by a panel discussion among subject experts and psychometricians. This represents the most critical endeavor made by the testing community to unify testing theories and cognitive psychology.

CD is new modeling philosophy which represents a drastic departure from traditional testing in many aspects. It also enjoys some advantages over traditional testing, so the term diagnosis is preferred over testing by the testing community. Large-scale standardized
assessments typically provide a single summary score which reflects the overall performance level of students in a certain content area. By contrast, cognitive diagnostic assessments generates a multidimensional cognitive profile (a skill vector) of students which provides information about their cognitive strengths and weaknesses. The shift from a single summary score to a skill vector has important implications. In traditional testing, students may obtain the same scores for different reasons (Tatsuoka & Tatsuoka, 1989). In diagnostic assessments, the skill vector can facilitate the identification of these individual differences with respect to the specific content in the domain of interest (Nichols, Chipman, & Brennan, 1995) and can provide useful information for students and teachers. The more comprehensive profiles for students’ skills can both rank or classify students as does traditional testing, and can also help integrate instruction and assessments (Campione & Brown, 1990).

More specifically, diagnostic testing offers a means of measuring students who have the capacity to learn. This involves identifying individuals who are likely to experience difficulties (Embretson, 1990) or who are ready to move on to higher levels within a given content domain (Gott, 1990). An adaptive personalized learning plan, of either a remedial or gifted-learning nature, can be made, and instructional material can be selected accordingly based on the instructionally useful diagnostic information (Embretson, 1990).

In summary, CD is a huge step made by the testing community that embraces a substantive approach to psychometric theories. Although the cognitive models in CD are simplistic compared with those in cognitive psychology, it does make cognitive assessments possible.
1.2 Computer-based testing techniques

Bunderson, Inouye, and Olsen (1988) proposed a useful taxonomy framework for computer-based testing techniques. They distinguished four generations of computer-based techniques: computerized testing, computerized adaptive testing (CAT), continuous measurement (CM), and intelligent measurement (IM). In the first generation, computerized testing, also known as “linear computer-based testing”, was the very first development that used computer technology for testing. It consisted of administering conventional tests using computers. The second generation, CAT, involved an optimal and individualized test for every examinee being constructed in real time. The third generation, CM, focused on assessment of learning progress. It involved using calibrated measures which were embedded in a curriculum to continuously and unobtrusively estimate dynamic changes in a student's achievement trajectory and profile as a learner. The fourth generation, MI, involved knowledge bases and inferential procedures. IM sought to provide intelligent scoring, interpretation of individual profiles, and advice for learners and teachers.

In retrospect, this taxonomy is a vision for, rather than a tentative summary of, the newest developments in the computer technologies which are used in testing. Many important research topics on CAT have begun to emerge. Several important large-scale applications of CAT such as the ASVAB and the GRE, demonstrated the feasibility and advantages of CAT, and presented interesting research questions regarding CAT. During a period of two decades, studies on many important aspects of CAT were conducted. These studies included item selection algorithms dealing with item exposure control (Leung, Chang, & Hau, 2002; Chang, Qian, & Ying, 2001; Chang & Ying, 1999; Sympson & Hetter, 1985), content balancing (Cheng, Chang, Douglas, & Guo, 2009; Cheng, Chang, & Yi, 2007; van der Linden & Chang, 2003), multidimensional IRT
CAT (Wang & Chang, 2011; Wang & Chang, 2009; Veldkamp & van der Linden, 2002; van der Linden, 1999), and the mathematical foundation of CAT (Chang, 2014; Chang & Ying, 2009), etc. These studies advanced the field’s knowledge of CAT in a significant manner. The state of research on CAT has made it a well-established field, and this research has contributed to the advancement of research on high-stakes testing.

CM can be regarded as a naïve version of CD-CAT as currently understood. CM takes advantage of CAT’s flexibility in terms of continuous administration and administers multiple assessments of milestones of student learning. It differs from CD-CAT in two aspects. First, it puts more emphasis on the monitoring role rather than instructional advice. Second, it lacks the support of a psychometric theory which should accomplish cognitive diagnosis. Bunderson et al. (1988) noted that the CM system had not been fully realized yet because it lacked certain critical features such as multidimensional scaling and a new psychometric procedure. Given recent developments in CD and CD-CAT, these gaps can now be filled. Various CDMs now make the development of cognitive diagnostic assessments possible, and large item banks for CD-CAT can be developed. Test researchers have also begun to research the item selection algorithms used for the CD-CAT, and begun to address important issues such as measurement accuracy, non-statistical constraints (item exposure control and content balancing) and simultaneously obtaining both cognitive information and general ability. A comprehensive review will be provided in Chapter 2.

The fourth generation is the prototype of smart learning that aims to fully integrate diagnosis and instruction/learning, and offers instructional advice for individualized learning, remediation programs and tailored advanced programs. Bunderson et al. (1988) used expert tutor systems as examples of the fourth generation, but also expressed the concern that there was a
disconnection between the computer science community and the educational measurement community. Unfortunately, tutor system design mostly conducted by the computer science community, and did not consider psychometric issues. Similarly, measurement research has not provided a great deal of research regarding tutoring. Snow and Mandinach (1991) made the same observation. The intersection of these two is critical if "intelligence" is to be built into computerized instructional systems. CD-CAT as conceived in the present study can represent a possible solution because it can serve as the driving engine or navigator which is responsible for charting personalized learning routes in an adaptive manner in a smart learning system.

In summary, the testing community has achieved the full realization of the second generation and can state confidently that the current state-of-the-art of CD-CAT can meet the challenges of the third generation. More studies are needed to advance our knowledge of CD-CAT and make it ready (or easier) for the convergence of different scientific endeavors in testing, cognitive psychology and computer science for the purpose of building a true smart learning system.

1.3 Educational policy

A brief review of The Element and Secondary Education Act (ESEA) presents a clear picture of changes in emphasis of educational assessments in the United States. When ESEA was first signed into law by President Lyndon Johnson in 1965, the emphasis was on allocating funds to students in need, while assessments were not considered as a priority.

After two decades, the emphasis shifted to student achievement. This led to significant change represented by the No Child Left Behind (NCLB) Act. NCLB relied on the Adequate Yearly Progress (AYP) in standardized testing to hold public-funded schools accountable.
However, this accountability system was based solely on standardized testing, and was criticized as “test and punish,” and some suggested that the emphasis should be changed to “support and improve” (American Federation of Teachers, 2014).

Duncan, the Secretary of the Department of Education, responded by calling for replacing NCLB with a new ESEA that “would ensure that all young people are prepared to succeed in college and careers” (US Department of Education, 2015a). This new viewpoint has not yet become a law, but an important initiative which reflects the same spirit has been implemented. This is the Common Core State Standards Initiative. It seeks to prepare America’s students for college and careers. As an incentive to adopt the Common Core Standards and motivate for educational reform, the competitive federal Race to the Top (RTTT, RTT or R2T) grants were announced in 2009. In addition to the prior requirements regarding teacher effectiveness, school effectiveness and the aid to the lowest-performing schools, these RTTT grants also ask states to advance reforms in two areas (US Department of Education, 2015c):

- Adopting standards and assessments that prepare students to succeed in college and the workplace and to compete in the global economy; and
- Building data systems that measure student growth and success, and inform teachers and principals about how they can improve instruction.

One significant change is the new emphasis on a data system for improving instruction that can lead to a new era of K-12 assessments in which both accountability and instructional improvement are emphasized (Chang, 2012). This reflects the growing importance of student assessment by emphasizing on development of state-wide longitudinal data warehouses for monitoring student growth and learning. This is intended to help teachers provide highly targeted and effective instruction in order to prepare the next generation of students for success in college and the
workforce (US Department of Education, 2015b). Thus, in addition to providing a summary score for accountability purposes, providing diagnostic information is intended to promote instructional improvement, and has become an important goal of next-generation assessment.

The Race to the Top Assessment Program, is an integral part of the RTTT program. Two comprehensive Assessment Systems, the Partnership for the Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessments Consortium, have been developed. This new mission is reflected in the design of PARCC and the Smarter Balanced assessments. These two assessment systems include both summative and formative assessment components (PARCC, 2012; Smarter Balanced, 2012). Both assessment systems emphasize the important role of technology in delivering the assessment: PARRC adopts computer-based testing, the equivalent of the first generation in Bunderson et al. (1988)’s taxonomy, while Smarter Balanced takes full advantage of CAT (the second generation).

1.4 Online Education (Moocs)

Recent developments in online education call for a method of delivering quality individualized education. The greatly hyped phenomena in distance learning, known as Massive Open Online Courses (Moocs), is a platform initiated by MIT to offer online courses to the public which aims for unlimited participation and open access. Like many tutoring systems, Moocs are another large-scale educational technology experiment implemented by the computer science community. It has attracted a great deal of attention and acclaim, but serious criticism has also been raised. The most poignant and succinct criticism probably is the one made by Vardi (2012) who noted "absence of serious pedagogy in MOOCs" in conjunction with a format of "short, unsophisticated video chunks, interleaved with online quizzes, and accompanied by social networking.” This major concern regarding the MOOCs pedagogical
issue was echoed by several other researchers (Hew & Cheung, 2014; Cooper & Sahami, 2013; Yuan & Powell, 2013). These criticisms have led to the development of an alternative view which involves relegating online learning to a supplementary category of the formal classroom teaching as a form of hybrid education (Cooper & Sahami, 2013).

It is worth noting that the statement about the disconnection between the computer science and the testing communities remains true in Moocs studies, although some progress has been made. The importance of assessments has been recognized in Moocs research (Hew, 2015; Hew & Cheung, 2014; Yuan, Powell, & Olivier, 2014; Cooper & Sahami, 2013; Piech et al., 2013; Sandeen, 2013; Yuan & Powell, 2013; Vardi, 2012). The major assessment topics in Moocs are peer grading and automatic grading for complex responses (Hew, 2015; Hew & Cheung, 2014; Yuan et al., 2014; Piech et al., 2013; Yuan & Powell, 2013). Automatic grading has been an important research topic in psychometrics, and has usually been referred as automatic scoring. A prominent example is the E-rater engine (Burstein, 2003). However, there is no explicit documentation of how the Moocs applies the Classic Test theory (CTT) or item response theory (IRT) to assessments. The notion of integrating assessments into instruction has not yet been introduced to the Moocs research community. This unfortunate state of affairs has persisted after more than two decades.

CD offers great potential in the improvement of the soundness of the online learning pedagogy if it can be seamlessly integrated into online instruction design and charter individualized learning pathways for each learner. In addition, hybrid education is the equivalent of the online version of a remedial program in which CD plays an important role. Furthermore, CAT enjoys the same advantages as online instruction because it breaks time constraints and
physical constraints of traditional classroom teaching. CD-CAT can augment online education using measurement techniques in both formal online teaching and hybrid education.

In summary, Chapter 1 briefly summarized recent developments in two important subfields in educational measurement and psychometrics, CD and CAT. This indicates the testing community’s readiness to develop CD-CAT and the challenges and opportunities presented by educational policies and online education. The union of CD and CAT, an important endeavor within the testing community, can finally come to fruition.
Chapter 2 A Synthesis of Literature on the Item Selection Algorithms in CD-CAT

Research on item selection algorithms in CD-CAT has exhibited some progress since its inception in the early 2000s. Many item selection algorithms have been developed to address various needs in the application of CD-CAT. This review attempts to identify important similarities and differences in existing algorithms and summarize their relationship using a new taxonomy for these algorithms. This process led to the conceptualization and proposition of some possible future studies.

The item selection algorithms for CD-CAT in this review will be generally organized into two broad categories: (a) basic algorithms concerned with measurement precision and efficiency; and (b) applications of basic algorithms which address specific issues such as item exposure control, content management, dual-purpose CD-CAT which seek to assess both cognitive diagnostic information and general ability in one test administration and so on. Unlike the well-established topic of item selection algorithms in traditional CAT, the one in CD-CAT is in its infancy and for certain important subtopics, there may only be one or two papers. A more feasible and fruitful strategy is to provide a brief summary of a particular subtopic in the traditional CAT as needed.

The review has two parts: (a) Part I concerns basic algorithms involved with measurement precision and efficiency and (b) Part II concerns three important applications of CD-CAT which deal with item exposure control, content management and dual-purpose CD-CAT. Comments will be made with regard to each subtopic and a general conclusion will be provided.
Before proceeding to the review, the basic setup for item selection in CD-CAT is described briefly in order to clarify the mathematical notation which will be used throughout this dissertation. For a CD-CAT with $K$ independent attributes, the general population can be assigned to $2^k$ distinct cognitive patterns. The cognitive pattern vector for examinee $i$ will be denoted as $\mathbf{a}_i$ and the specific element in the vector for the $k^{th}$ attribute is $a_{ik}$. An item bank consisting of $J$ items is calibrated according to certain CDM. For the $j^{th}$ item with Q matrix $q_{jk}$ where $k = 1, 2, ..., K$ in the bank, the generic item response function is denoted as $P(Y_{ij} = y | \mathbf{a}_i)$ where $Y_{ij}$ is the random variable for examinee $i$’s response to item $j$ and $y$ is the realization value of $Y_{ij}$.

In order to construct an optimal test for each examinee, items are selected using an item selection algorithm. The posterior distribution of the cognitive pattern for examinee $i$ is updated after each item is administered. The estimated cognitive pattern for examinee $i$, $\hat{\mathbf{a}}_i$, can be obtained from the posterior. Assuming that $t$ items have been administered, examinee’s response vector to the $t$ items is denoted as $\mathbf{y}_i$ and then the posterior is $\pi(\mathbf{a}_i | \mathbf{y}_i)$. Most CAT item selection algorithms are information-based although there are some other alternatives. Most item selection algorithms in CD-CAT are also information-based. They have been constructed using the Kullback–Leibler (KL) index for the distribution of item responses or the Shannon Entropy (SHE) of the posterior distribution of the cognitive patterns.

2.1 Basic Algorithms Concerned with Measurement Precision and Efficiency

The primary concern of an item selection algorithm is to achieve high measurement precision in an efficient manner. This is also the theme of the development of item selection
algorithms in CD-CAT during the early years and today. A new taxonomy is proposed to facilitate the evaluation of the basic algorithms in terms of measurement precision and efficiency (with respect to the computational burden).

Two general approaches can be found among the basic algorithms: the KL approach and the SHE approach. Research on item selection in CD-CAT originated from the SHE algorithm for the sequential classification experiment for CD by Tatsuoka (2002) and Tatsuoka and Ferguson (2003) and the KL algorithm by Xu, Chang, and Douglas (2003). The key difference between these two algorithms, and the justification for the new taxonomy, is that the distributions involved in the calculation have important implications for measurement and computational efficiency. The KL approach attempts to develop a global summative measure for the difference between the distributions of the response to the candidate item conditional on all of the possible true and estimated cognitive patterns. By contrast, the SHE approach involves the current (and/or previous) posterior distribution(s) of cognitive patterns conditional upon all of the previous responses and the possible response to the candidate item.

Both approaches have certain strengths and weaknesses. Compared with the indirect KL approach, the SHE approach is a more direct and effective measure in the context of CD-CAT in which an accurate estimate of the cognitive pattern obtained via the updated posterior distribution is the ultimate goal. As regards the distributions involved, we can see that the KL approach attempts to measure the difference a candidate item can make in the distributions of the current response to it while the SHE approach attempts to measure the difference a candidate item can make in the current posterior distribution of the cognitive pattern. Therefore, the SHE approach enjoys certain advantages in terms of measurement efficiency. The KL approach, however, has an edge over the SHE approach with respect to computational efficiency. All of the
components in the KL approach can be exhaustively numerated. Thus, all of the possible values of the KL index can be calculated beforehand. During the execution of CD-CAT, item selection is reduced to picking out the maximum values in the pre-stored matrix. The SHE approach, by contrast, has to be calculated on-the-fly during the execution of CD-CAT because the current posterior probability of the cognitive patterns requires real-time updating after an examinee has answered the administered item. These advantages and disadvantages are shared by other algorithms in the two approaches described below.

The original KL and SHE algorithms and their important developments are presented as follows based on the new taxonomy.

**The KL Approach.** In order to obtain a measure of global discrimination power of item $j$ between the distribution of the response conditional on the estimated cognitive pattern $P(Y_{ij} \mid \hat{\alpha}_i)$ and the distributions which are conditional on all possible cognitive patterns $P(Y_{ij} \mid \alpha_c)$ for $c = 1, 2, \ldots, 2^K$, Xu et al. (2003) proposed using the KL index, which is the sum of KL distance between $P(Y_{ij} \mid \hat{\alpha}_i)$ and all $P(Y_{ij} \mid \alpha_c)$. This is formulated as:

$$KL_j(\hat{\alpha}_i) = \sum_{c=1}^{2^K} \left[ \sum_{y=0}^{1} \log \left( \frac{P(Y_{ij} = y \mid \hat{\alpha}_i)}{P(Y_{ij} = y \mid \alpha_c)} \right) P(Y_{ij} = y \mid \hat{\alpha}_i) \right].$$

The item with the maximum value for KL, given the cognitive pattern $\hat{\alpha}_i$ for examinee $i$, will be administered.

From the equation, we can see that all the possible KL indices for each item in the entire bank can be calculated without any information about the estimated cognitive patterns. There are only two possible values for the random variable $Y_{ij}$ (0 and 1) and there are $2^K$ possible values.
for cognitive patterns. The empirical simulation studies indicated that the KL index cannot achieve a pattern recovery rate similar to the SHE index for the DINA (Cheng, 2009) and the Fusion model with a fix-length CD-CAT (Xu et al., 2003).

The low efficiency issue was remedied by a Bayesian KL index, namely the posterior-weighted KL (PWKL) index (Cheng, 2009). In order to reflect the varying importance of different patterns, the addend in the KL index is weighted by the corresponding posterior probability, and this modification leads to the PWKL (Cheng, 2009):

$$PWKL_j(\tilde{a}_i) = \sum_{c=1}^{2^k} \left\{ \sum_{y=0}^{1} \log \left( \frac{P(Y_{ij} = y | \tilde{a}_j)}{P(Y_{ij} = y | a_c)} \right) P(Y_{ij} = y | \tilde{a}_j) \right\} g(a_c | y_{t-1})$$

where $g(a_c | y_{t-1}) = p(a_c) \prod_{j=1}^{t-1} P(Y_{ij} = 1 | a_c)^{y_j} \left[1 - P(Y_{ij} = 1 | a_c)\right]^{1-y_j}$, $p(a_c)$ is the prior probability of the cognitive patterns, and $y_{t-1}$ is the vector of responses on $t-1$ items for examinee $i$. Inspired by Henson and Douglas (2005)’s discussion on the relationship between the item discrimination power and the cognitive pattern distance, Cheng (2009) further assigned additional weights to the cognitive patterns that are closer to the current cognitive pattern estimate and defined the hybrid KL (HKL) whose formulation is presented below:

$$HKL = \sum_{c=1}^{2^k} \left\{ \sum_{y=0}^{1} \log \left( \frac{P(Y_{ij} = y | \tilde{a}_j)}{P(Y_{ij} = y | a_c)} \right) P(Y_{ij} = y | \tilde{a}_j) \right\} g(a_c | y_{t-1}) \frac{1}{d(a_c, \tilde{a}_j)}$$

where $d(a_c, \tilde{a}_j)$ is the Euclidean distance $\sqrt{\sum_{k=1}^{K}(a_{ck} - \tilde{a}_{ik})^2}$ or, equivalently, the Hamming distance $\sum_{k=1}^{K}|a_{ck} - \tilde{a}_{ik}|$ for the possible cognitive pattern $a_i$ and the current cognitive pattern estimate $\tilde{a}_j$.

Cheng (2010) also proposed the modified maximum global discrimination index (MMGDI) to remedy the problem with the KL index by considering the balance of attribute
coverage. One clarification of the terminologies here is that so-called global discrimination index (GDI) is another term coined for the KL index. The attribute-balancing index is defined as follows:

\[
\prod_{k=1}^{K} \left( \frac{B_k - b_k}{B_k} \right)^{q_{jk}}
\]

where \( B_k \) is the minimum number of items required to measure the \( k^{th} \) attribute and \( b_k \) is the number of items measuring the \( k^{th} \) attribute that have already been selected. The MMGDI can be reformulated as follows:

\[
MMGDI_j(\tilde{a}_j) = \prod_{k=1}^{K} \left( \frac{B_k - b_k}{B_k} \right)^{q_{jk}} \times GDI_j
\]

\[
= \prod_{k=1}^{K} \left( \frac{B_k - b_k}{B_k} \right)^{q_{jk}} \times KL_j
\]

\[
= \prod_{k=1}^{K} \left( \frac{B_k - b_k}{B_k} \right)^{q_{jk}} \sum_{\gamma=0}^{1} \sum_{c=1}^{2^{k}} \log \left( \frac{P(Y_{ij} = y | \tilde{a}_j)}{P(Y_{ij} = y | a_{c})} \right) P(Y_{ij} = y | \tilde{a}_j).
\]

The MMGDI can be interpreted as the KL index weighted by the attribute-balancing index similar to the manner in which the PWKL resolves the issue with the KL index while taking advantage of the attribute balancing instead of the information in the posterior of the cognitive patterns.

PWKL is an important development for the KL approach and it becomes the “gold standard” for CD-CAT item algorithms. Due to the existence of additional information on the posterior of the cognitive patterns, the PWKL and HKL are much more efficient than the original KL index and even slightly better than the SHE index. Furthermore, they also enjoy the
advantage of having KL part and distance weights (in HKL) calculated beforehand, and only the posterior weights need to be updated.

The SHE Approach. Tatsuoka (2002) and Tatsuoka and Ferguson (2003) proposed the SHE item selection algorithm. Shannon entropy quantifies the uncertainty inherent in a distribution. Shannon entropy is maximized if the distribution is uniform, and is minimized if the probability mass concentrates on a single point. In CD-CAT, an ideal item would be one that minimizes the expected Shannon entropy of the posterior distribution of $\hat{\alpha}_i$ conditional on previous responses. Thus, the SHE index is defined as

$$SHE = \sum_{y=0}^{1} \left[ \sum_{c=1}^{2^x} \pi(\alpha_c | y_{t-1}, Y_t = y) \log \left( \frac{1}{\pi(\alpha_c | y_{t-1}, Y_t = y)} \right) \right] \left[ \sum_{c=1}^{2^x} P(Y_t = y | \alpha_c) \pi(\alpha_c | y_{t-1}) \right],$$

where $y_{t-1}$ denotes the response vector of $t-1$ items for examinee $i$. SHE can be considered as the KL distance between the uniform distribution and the current posterior distribution of $\hat{\alpha}_i$ (so, the KL in this review is defined in the narrow sense in the same way as in the KL approach above, if not indicated otherwise).

It is easy to observe from the equation that the value of the posterior of the cognitive patterns needs online updating during CAT administration. Thus, no calculation can be made in advance. Fortunately, this does not appear to be an issue in the SHE. As mentioned above, in terms of measurement efficiency, it is superior to the KL index and is comparable to the PWKL and HKL.

A recent development for the SHE approach is the mutual information (MI) for CD-CAT (Wang, 2013). MI is equivalent to the KL distance between two subsequent posterior distributions. Thus, the SHE is a special case of MI. The expected MI is calculated as
\[ MI = \frac{1}{h_1} \sum_{y=0}^{h_2} \sum_{c=1}^{x} P(y_{t-1}, Y_t = y | \alpha_c) P(\alpha_c) \log \frac{P(Y_t = y | \alpha)}{\frac{P(y_{t-1}, Y_t = y)}{h_2}}. \]

MI has been shown to be more efficient than competing item selection methods, such as those based on KL-information and SHE, particularly for short tests. However, the computational efficiency issue of the SHE approach poses a serious practical challenge, since the online updating of the posteriors and a triple summation are involved in MI. By some algebraic manipulations, Wang (2013) presented a simplified version of MI:

\[ MI = \frac{1}{h_1} \sum_{y=0}^{h_2} \sum_{c=1}^{x} P(y_{t-1}, Y_t = y | \alpha_c) P(\alpha_c) \log \frac{P(y_{t-1}, Y_t = y)}{h_2}. \]

where \( h_1 = \sum_{c=1}^{x} P(y_{t-1} | \alpha_c) P(\alpha_c) \) and \( h_2 = \sum_{c=1}^{x} P(y_{t-1}, Y_t = y | \alpha_c) P(\alpha_c) \). Wang (2013) reduced the calculation burden by dropping some terms only related to \( h_1 \) because it is a constant term over different items. One problem with such a simplification scheme is that it only preserves the rank of the original index and there is a change in scale and/or sign. Therefore, if weighted by an item exposure control and constraint management index via multiplication, a simplified MI would produce an incorrect ordering of items.

In summary, there are several feasible algorithms developed to achieve high measurement precision in an efficient manner. A new classification framework is proposed here to facilitate the evaluation of these methods. The analysis of the strengths and weaknesses of the existing algorithms points to the possibility of developing a KL-based computationally feasible algorithm whose measurement performance is comparable to, or even better than, the MI.

In concluding this section, we state that a method different than information-based indices is a rate function approach to CD-CAT (Liu, Ying, & Zhang, 2013) which attempts to
tackle the item selection issue from the angle of misclassification probabilities. Simulation studies have shown that its efficiency is similar to SHE and some advantages might be demonstrated when the test length exceeds 20 items for a test with 5 and more attributes.

2.2 Algorithms for specific applications

The basic algorithms focus exclusively on measurement precision and disregard all other important practical constraints in CD-CAT. In real-world applications of CD-CAT, many practical issues, such as item exposure control and content management, must be solved. One unique application of CD-CAT, labeled as dual-purpose CD-CAT, is using it to obtain diagnostic information and traditional single summative information in a single test administration. Most currently used algorithms are built upon basic algorithms (the measurement precision part) by incorporating other objectives. The most common strategies involve adding constraint weighting and/or constructing a linear combination of the multiple objectives. There are also some other separate methods that deal with additional requirements such as the Monte Carlo method. They are used together with one of the basic item selection algorithms. A detailed analysis of the three applications for CD-CAT, with a reference to the corresponding subtopics in the traditional CAT if possible, will be presented.

**Item exposure control.** In recent decades, there have been a number of different exposure control approaches proposed for traditional CAT in the literature. Some focus on preventing overexposure of items, others focus on increasing the use of under exposed items, and some aim to combine both objectives. Georgiadou, Triantafillou, and Economides (2007) identified that there are at least five different types of exposure control strategies: (1) randomization (Kingsbury & Zara, 1989; McBride & Martin, 1983); (2) conditional selection (Chen & Lei, 2005; van der Linden & Veldkamp, 2004; Chang & Ansley, 2003; Stocking, 1993; Sympson & Hetter, 1985);
(3) stratification strategies (Yi & Chang, 2003; Chang et al., 2001; Chang & Ying, 1999); (4) combined strategies (van der Linden & Chang, 2003; Leung et al., 2002; Eggen, 2001; Revuelta & Ponsoda, 1998); and (5) multiple-stage adaptive test designs (Luecht & Nungester, 1998).

Two papers are devoted to the item exposure control issue in CD-CAT, but present two distinctively different approaches: Wang, Chang, and Huebner (2011) proposed two restrictive stochastic algorithms that add some randomness to the PWKL, while Hsu, Wang, and Chen (2013) developed a complicated multi-purpose item exposure control method based on the Sympson-Hetter (SH) method (Sympson & Hetter, 1985). Another important difference is that the former is only suitable for fixed-length CD-CAT, while the latter is suitable for variable-length CD-CAT.

**Restrictive stochastic item selection in CD-CAT.** The two restrictive stochastic methods for item selection in CD-CAT, restrictive progressive method (RP) and restrictive threshold method (RT), fall into the first and fourth categories introduced above. As their names indicate, the basic idea behind the methods is to change the original deterministic approach, based purely on item information, to a stochastic approach. This is accomplished by imposing a random component in item selection or selecting an item from a candidate set rather than strictly selecting the item with the maximum information.

RP consists of two controls, progressive control and restrictive control. The primary idea of progressive control is to add a stochastic component to the item selection criterion (Revuelta & Ponsoda, 1998), such that it will not always choose the items with the greatest amount of information. The restrictive control seeks to suppress overexposure by adding a restriction on the maximum exposure rate. Combining the two ideas leads to the restrictive progressive method item selection index being denoted as
\[
RP_{PWKL_j} = \left(1 - \frac{\exp_j}{r}\right)[(1 - x/L)R_j + PWKL_j \times \beta x / L],
\]

where \(x\) is the number of items administered, \(L\) is the test length, \(\beta\) is the importance parameter, and \(R_j \sim \text{uniform}(0, \max(PWKL(X_j)))\). The restriction component is the term \(1 - \frac{\exp_j}{r}\), which ensures that the maximum item exposure rate will be kept below a certain value, \(r\). The progressive component is the changing weight \((1 - x/L)\) of the random component. More specifically, the importance of the information increases as the test progresses whereas that of the stochastic component decreases. It is reasonable to assume that interim ability estimates are more likely to be distant from the true values at the early stage of the test and thus the information component should contribute less to the item selection. However, as the test progresses and the provisional ability estimates approach the true ability of the examinee, the information component should gain in importance. In so doing, the stochastic component can achieve a decent item exposure rate during the early stage while the measurement precision can still be maintained or only slightly decreased due to the increasing importance of the information during the later stage.

RT also consist of two parts, a restrictive component and a threshold component. The threshold component is designed to construct sets of items with information values that are close to the maximum so that selecting items from this set may equalize item exposure rates. More specifically, using a restrictive threshold method defines an information interval

\[
\left[\max(PWKL_j) - \delta, \max(PWKL_j)\right],
\]
where $\delta$ is the threshold parameter and is defined as $\left[\max(PWKL_j) - \min(PWKL_j)\right] \times f(x)$, in which $x$ is the number of items already administered and $f(x) = \left(1 - \frac{x}{L}\right)^\beta$ is a monotone decreasing function. The importance parameter $\beta$ controls the relative importance of the exposure balance versus the estimation precision. Following the same rationale as is the case for the restrictive progressive method, the information interval should be large during the early stage of the test and gradually shrink as the test progresses. In addition, the maximum exposure rate can be set. The items with exposure rates which exceed the maximum exposure rate will be excluded from item selection. For a constant $\delta$, the items whose information lie in this interval form a candidate set, $S_{(c)}$, from which one item is selected randomly as the next one to be administered in CD-CAT.

It is clear that RP and RT share the same idea that turning the deterministic nature of the purely information-based method into a stochastic approach by incorporating some randomness. The simulation studies show that RP and RT perform very well in terms of maintaining the accuracy and achieving decent item exposure balance. There are two issues with the RP and RT: the algorithm is highly complex and is only applicable to fixed-length CD-CAT. The second issue can be remedied by using the SH-based method described below.

The SH-based method. Hsu et al. (2013) proposed a separate item control mechanism based on the SH method (Sympson & Hetter, 1985) for variable-length CD-CAT that can be used together with the basic algorithms. This method is called the Sympson–Hetter method, which comprises test overlap control, variable length, online update, and restricted maximum information (SHTVOR). This procedure is based on the SH method and is capable of controlling test overlap.
for variable-length termination, online updating the exposure-control parameters, and using restricted maximum information to freeze items with an exposure rate greater than the pre-specified maximum until their exposure rate decreases. As the name suggests, SHTVOR falls into the fourth category (the combined strategy) and attempts to achieve multiple purposes. Its implementation is more complicated than the restrictive stochastic methods. SHTVOR consists of the following 7 steps (Hsu et al., 2013).

1. Initialize/set the parameters, such as the number of items in the bank $J$, the target maximum item exposure rate $r_{max}$, target test overlap rate $\bar{T}_{max}$ and the exposure control parameter of item $p_k$ which is set to be 1, etc.

2. Administer CAT to an examinee by comparing $p_k$ to a randomly generated number from $U(0,1)$. Exclude the administered item for this examinee from the item pool.

3. Update the examinee’s cognitive pattern estimate and select another item as described in Step 2 until the examinee has reached the pre-specified fixed precision or until the maximum test length is reached.

4. Update $\bar{T}$ and $p_k$ as follows:

$$\bar{T} = \frac{N \times \sum_{j=1}^{J} P(A_j)^2}{\bar{L} \times (N-1)} - \frac{1}{(N-1)}$$

where $P(A_j)$ and $P(S_j)$ are the percentages an item has been administered and selected, respectively, for item $j$ ($j = 1, 2, ..., J$); $\bar{L}$ is the mean test length across all examinees; and $N$ is the total number of examinees who have undergone CAT thus far.
\[ p_k = \begin{cases} 
0, & \text{if } P(A) \geq r_{\text{max}} \\
\frac{r_{\text{max}}}{P(S)}, & \text{if } P(A) \leq r_{\text{max}} \text{ and } P(S) > r_{\text{max}} \\
1, & \text{if } P(A) \leq r_{\text{max}} \text{ and } P(S) \leq r_{\text{max}}. 
\end{cases} \]

5. If \( \bar{T} > \bar{T}_{\text{max}} \), then.

1) Calculate the target variance of the item exposure rate across items \( S^2_0 \) while \( \bar{T} = \bar{T}_{\text{max}} \).

2) Set \( P(A)' = S_0 \left[ \frac{P(A) - \bar{r}}{S} \right] + \bar{r} \) and \( P^*_k = \frac{P(A)'}{P(S)} \) where \( P(A)' \) is the adjusted percentage of times that an item has been administered based on \( S_0 \) and \( P^*_k \) is the adjusted exposure-control parameter.

3) If \( P^*_k > 1 \), then \( P_k' = 1 \); if \( P_k > P_k' \), then \( P_k = P_k' \).

6. Set the \( L_{\text{max}} \) largest \( P_k \)'s as 1 to guarantee that all examinees will complete the CAT before exhausting the entire bank.

7. With the updated \( P_k \)'s, repeat Step 2-6 to administer the CAT again until all of the examinees have finished the CAT.

To summarize, the existing algorithms for controlling item exposure rates in CD-CAT can be identified from the first and fourth categories. The common difficulty shared by the two methods is the conceptual understanding of them and their implementation, especially the SHTVOR, even though it is more flexible than the restrictive stochastic methods to be applicable in variable-length CD-CAT. Some of the easy alternatives, ideally applicable to both fixed-
length and variable-length tests, would be welcomed and the second category (the stratification strategy) offers great potential.

**Content management.** Three major approaches for the item selection algorithms with content constraints have been implemented in traditional CAT (Mao & Xin, 2013). The first is a heuristic algorithm that selects items in a restrictive manner for different item subsets. Examples include the constrained CAT (Kingsbury & Zara, 1991), the modified multinomial model (Chen & Ankenman, 2004), and the modified constrained CAT (Leung, Chang, & Hau, 2003). These algorithms enjoy the advantage of easy implementation, but are limited in the number of constraints involved and also impose some requirements on the proportions among the constraints. The selection of items will become cumbersome if the number of the constraints is large or the proportions are inappropriate. Moreover, some of the constraints may be violated due to sampling error during the real testing process. The second also applies a heuristic algorithm that weights the item information index using the content constraint requirement. Typical algorithms in this family are the weighted deviation model (Stocking & Swanson, 1993), the maximum priority index method (Cheng & Chang, 2009), and the constraint-weighted a-stratification method (Cheng et al., 2009). This method type avoids the computational complexity and infeasibility issues, but at the price of ignoring some requirements and/or some measurement precision loss (Cheng & Chang, 2009). The third approach is the mathematical programming method. The shadow-test method (van der Linden & Reese, 1998), the multiple shadow-test method (van der Linden, 2005), and the Monte Carlo approach (Belov, Armstrong, & Weissman, 2008) are instances of this type. Mathematical programming methods single out test items based on a number of shadow tests assembled before the test. As a result, these mathematical programming methods select the approximately optimal item with which all the
constraints can be satisfied by the point when the test is over. However, the computations will be very complex and might possibly have no solution if there are too many constraints.

Mao and Xin (2013) proposed the first and only algorithm to address the content constraints in CD-CAT by extending the Monte Carlo method (Belov et al., 2008) to CD-CAT. The main idea of the Monte Carlo method is to construct many shadow tests in a uniform manner prior to the selection of each item. Each shadow test must include all of the items already presented and meet all of the constraints. In order to determine which item to administer next, the most informative item from all of the free items within these shadow tests is found. The algorithm can be broken into two elements: input/out parameters and steps. Input parameters include the set of administered items, the current cognitive pattern \( \bar{\alpha} \), the set of the shadow tests \( S \), the number of shadow tests required to be assembled beforehand \( m \) and the number of shadow tests already in set \( S \); the output parameters include the selected item \( \varphi \) and the updated set \( S' \). The 5 steps taken to choose the \((t+1)^{th}\) item can be described as follows.

1. Assemble \( m-r \) shadow tests uniformly to guarantee that all of the shadow tests have an equal chance of being chosen.
2. Add these \( m-r \) shadow test to the set \( S \) and then there are a total of \( m \) shadow tests.
3. Draw all of the items that have not yet been administered from every test in set \( S \).
4. Assign the most informative one at cognitive pattern \( \bar{\alpha} \) from all of the items that have not been administered to \( \varphi \).
5. Renew the set \( S \) by keeping all of the shadow tests that contain item \( \varphi \).

The Monte Carlo algorithm falls into the third approach and suffers from problems such as intensive computations and infeasible solutions. It will be interesting to develop some
computationally feasible algorithms using the second approach for CD-CAT. Possible ideas include the maximum priority index method (ideally enhanced by the restrictive progressive method for addressing the item exposure issue) and the counterpart of the constraint-weighted a-stratification method in CD-CAT if the stratified method in CD-CAT is available.

**Combing CAT and CD-CAT.** The new federal grant program known as “Race to the Top” (RTTT) has led into a new era of K-12 assessments which emphasized both accountability and instructional improvement (Chang, 2012). In addition to providing a summary score for accountability purposes, providing diagnostic information to promote instructional improvements becomes an important goal of next-generation assessment. This new mission is reflected in the design of the Assessment of Readiness for College and Careers and the Smarter Balanced Assessments Consortium. In these two assessment systems include both a summative assessment component and a formative assessment component.

In light of this, researchers began to consider the problem of obtaining the estimation of general ability (denoted by $\theta$) and diagnostic information regarding more specific skills (denoted by $\alpha$) in a single administration of CAT. Designing a test that targets both general ability estimation and specific cognitive feedback is not entirely new (Gibbons & Hedeker, 1992). Tatsuoka (1991) first proposed a rule-space methodology that provides a sound framework for showing how to extract a useful attribute mastery profile (i.e., cognitive information) from a test originally designed to produce a total score.

The majority of this line of research uses model based approaches that replace the rule-space framework (Tatsuoka, 1991) with more structured diagnostic classification models. In addition, the test are delivered via adaptive testing modes such that the latent traits can be more effectively estimated. At present, most research concerning item selection rules in CAT are based upon either
IRT or CDMs separately. This line of research attempts to develop a CAT in which the test is tailored interactively to both an examinee’s overall ability level and attribute mastery level, so that information carried by both $\theta$ and $\alpha$ are maximized.

The four studies that addressed the dual-purpose item selections in adaptive testing were McGlohen and Chang (2008), Cheng and Chang (2007) which was further developed by Wang, Zheng, and Chang (2014), and Wang, Chang, and Douglas (2012). They solved the dual-objective optimization problem using three different strategies. McGlohen and Chang (2008) proposed a two-stage method, in which the “shadow” test functions as a bridge to connect information gathered at $\theta$ for IRT and information accumulated at $\alpha$ for CDM. More specifically, a shadow test that is optimized according to the ability estimate $\hat{\theta}$ is constructed before the administration of each item. The best item for measuring the cognitive pattern $\hat{\alpha}$ on the basis of the current $\hat{\alpha}$ is then selected from the shadow estimate, using the SHE or KL algorithms. Wang et al. (2012) proposed a constraint weighted item selection algorithm that treats information at $\theta$ as the objective function and information at $\alpha$ as the statistical constraints. Cheng and Chang (2007) proposed a dual information method (DIM), which is a weighted sum of information gathered at $\theta$ and $\alpha$. Based on the DIM, Wang et al. (2014) proposed an aggregate ranked information method (ARI), an aggregate standardized information method (ASI), and three different weighting schemes based on theoretical findings, empirical needs and attribute-level information. A more detailed review of the second and third strategies are presented below but the first strategy is omitted due to the inherent problem with the shadow test approach and the mathematic programming methods mentioned above.

Wang et al. (2012) borrowed the idea of the maximum priority index from CAT and proposed three variant priority indexes for CD-CAT: the $Q$ (matrix) -control priority index, the $Q$ discrimination-control priority index and the KL information-control priority index. The $Q$-control
priority index for item $j$ is formulated as follows:

$$P_j = \prod_{k=1}^{K} \left[ \frac{u_k - x_k - q_{jk}}{u_k} \right] \left[ \frac{(L - l_k) - (t - x_k - q_{jk})}{L - l_k} \right],$$

where $u_k / l_k$ is the upper/lower bound for attribute $k$; $x_k$ is the number of items that have been administered for attribute $k$; $q_{jk}$ denotes the Q-matrix entry for attribute $k$ of item $j$; $L$ is the test length; and $t$ is the number of administered items. This index can be combined with the $\alpha$-stratification method (StrQ) or the maximum Fisher information (MIQ). The Q discrimination-control index can be developed with additional item quality information. The formula can be different depending on which specific CDMs are used. To take an example, the Q discrimination-control index for the DINA model is presented as

$$P_j = (1 - s_j) (1 - g_j) \prod_{k=1}^{K} \left[ \frac{u_k - x_k - q_{jk}}{u_k} \right] \left[ \frac{(L - l_k) - (t - x_k - q_{jk})}{L - l_k} \right],$$

where $s_j$ and $g_j$ are the slipping and guessing parameter respectively for item $j$. Correspondingly, the StrQD and MIQD can be derived.

The third priority index is based on the attribute-level CDM discrimination index (Henson, Roussos, Douglas, & He, 2008). Henson and Douglas (2005) proposed a CDM discrimination index (CDI) which facilitates test construction for cognitive diagnosis purposes. The building block for CDI is the $D$ matrix for the $j^{th}$ item, $D_j$, and this is a $2^K \times 2^K$ matrix whose entries are the KL distance between the response distributions for all of the distinct cognitive patterns. Each $u, v$ element in $D_j$ is

$$D_{juv} = E \left[ \log \left( \frac{P(X_j \mid a_u)}{P(X_j \mid a_v)} \right) \right] = \sum_{x=0}^{1} P(X_j \mid a_u)^x \log \left( \frac{P_{u,x} (X_j \mid a_u)}{P_{u,x} (X_j \mid a_v)} \right).$$
A weighted mean can be computed as the CDI of the discriminating power among cognitive patterns for the $j^{th}$ item.

$$CDI_j = \frac{1}{\sum_{uv} d(a_u, a_v)^{-1}} \sum_{uv} d(a_u, a_v)^{-1} D_{juv},$$

where $d(a_u, a_v) = \sum_{k=1}^{K} |a_{uk} - a_{vk}|$ is the Hamming distance between two cognitive patterns.

Henson et al. (2008) and Rupp et al. (2010) further developed the attribute-level CDI (ACDI) which is defined as

$$ACDI_{jk} = \frac{1}{2^K} \sum_{\text{all relevant cells}} D_{juv}.$$

$ACDI_{jk}$ is partial sum of the matrix D in which only the entries for two cognitive patterns with the Hamming distance of 1 are involved. The ACDI indicates the contribution of an item to the correct classification for each attribute. The corresponding priority index is

$$P_j = \sum_{k=1}^{K} (u_k - x_k) ACDI_{jk}.$$

Using the same line of thinking allow us to obtain the StraInfor and MIinfor.

It is not obvious that there is a progression among the three priority indexes from the description in Wang et al. (2012) paper, but a new perspective can help uncover the connection among them: the key element for the priority indices is the item discrimination index and crucial difference lies in the refinement of the item discrimination index. The Q-control index does not use any item discrimination index and only the qualitative information regarding the attributes is involved. $(1-s_j)(1-g_j)$ can be interpreted as a modified version of the item discrimination index in
the Classic Test Theory (CTT) for DINA, \( (1 - s_j - g_j) \) (Rupp et al., 2010). In addition, \( (1 - s_j)(1 - g_j) \) is highly correlated with \( (1 - s_j - g_j) \), but the current multiplicative formula is more suitable for serving as the weighting factor. The CDM discrimination index, by definition, is the discrimination index for CDMs constructed based on the KL distance and is a function of the slipping parameter, \( s_j \), the guessing parameter, \( g_j \), and the Q matrix, which are specified for the specific CDM. It requires a more complicated form than the multiplicative version of the CTT item discrimination index, and thus is much more refined.

From this point of view, the three priority indices can be regarded as a progression from no item discrimination to the coarse CTT item discrimination index to the more refined KL-based item discrimination index. It can expected that the KL-based algorithms (StraInfor and MIinfor) will outperform their counterparts which were derived from the first and second priority indices (StraQ, StraQD, MIQ and MIQD) with respect to the measurement precision. In addition, this perspective can afford to offer the possibility that some new item algorithms can be readily proposed by further extending this progression. PWKL and the Bayesian version of the CDI (BCDI) can also identified as an item discrimination index which is more complicated and refined than the ACDI, after which StraPWKL, StraBCDI, MiPWKL and MiBCDI can be proposed. However, one conceptual difficulty remains for the new methods. All of these priority index-based algorithms are a constraint weighted item selection algorithm that treats information at \( \theta \) as the objective function and information at \( \alpha \) as the statistical constraint. The new extensions, PWKL and BCDI, are also item information indices and thus they should be perceived as equal counterparts of the information indices in IRT. Thus a new term such as information product algorithm which encompasses the previous constraint weighted algorithms, as well as the new ones, is more appropriate. It would be interesting to explore their performance compared with the previous algorithms.
The third strategy is to combine both objective functions into a single functional form. A straightforward combination is a weighted linear sum of the objectives, and is defined as follows:

\[
Objective = wPWKL(\hat{\alpha}) + (1 - w)KL(\hat{\theta})
\]

where \( w \) is the weight, and \( \hat{\alpha} \) and \( \hat{\theta} \) are the intermediate estimates after each item is administered. This dual information method (DIM) was first proposed by Cheng and Chang (2007). Wang et al. (2014) addressed two issues intrinsic to DIM: non-comparability of the two information addends and the arbitrary selection of weight. The non-comparability appears as a result of integration.

It is clear that \( KL(\hat{\alpha}) \) consist of \( 2^k \) addends, whereas the number of addends in \( KL(\hat{\theta}) \) depends on how the integration domain is sliced. In addition, the size of the terms in \( KL(\hat{\theta}) \) differs greatly because every addend has \( \Delta\theta \) as a multiplier, which means \( KL(\hat{\theta}) \) will always be smaller than \( KL(\hat{\alpha}) \). Therefore \( KL(\hat{\alpha}) \) will play a dominant role in item selection when the above linear combination of the two information pieces is considered.

In order to solve this non-comparability issue, two modifications were made to: ARI and ASI. ARI was modified to transform both pieces of information to an ordinal scale in such a manner that each item would have two ranks for \( KL(\hat{\theta}) \) and \( PWKL(\hat{\alpha}) \) separately. ARI is therefore computed as

\[
ARI = wpe(PWKL(\hat{\alpha})) + (1 - w)pe(KL(\hat{\theta}))
\]

where \( pe(\bullet) \) represents “rank”. The rationale behind this method is that by using the ordinal scale, the information captured by \( \theta \) and \( \alpha \) can be aligned along the same scale and the weight \( W \) \((0 \leq w \leq 1)\) will reflect the true relative importance of the two pieces.

On the other hand, ASI standardized both of them to remove the scale difference.
\[
KL^*(\hat{\theta}) = \frac{(KL(\hat{\theta}) - \text{mean}(KL(\hat{\theta})))}{SD(KL(\hat{\theta}))},
\]
\[
PWL^*(\hat{\alpha}) = \frac{PWKL(\hat{\alpha}) - \text{mean}(PWKL(\hat{\alpha}))}{SD(PWKL(\hat{\alpha}))},
\]
\[
ASI = wPWKL^*(\hat{\alpha}) + (1-w)KL^*(\hat{\theta}).
\]

Wang et al. (2014) also proposed three different weighting schemes to determine the value for \( w \) more effectively than the arbitrary weights in Cheng and Chang (2007): theory-based weights, empirical weights and attribute-level weights. Theory-based weights are based on the differential rate of \( \hat{\theta} \) converging to \( \theta \) (Chang & Stout, 1993) and \( \hat{\alpha} \) to \( \alpha \) (Tatsuoka, 2002), and it is reasonable to give more weight to the faster convergent \( KL(\hat{\alpha}) \) at the beginning of the test to accelerate its convergence. During the later stage of the test when \( \hat{\alpha} \) is estimated more accurately, more weight can be put on \( KL(\hat{\theta}) \). A simple way to define weight to reflect this transition is \( w = 1 - n/L \) where \( n \) is the number of items that have been chosen so far, and \( L \) is the test length. This approach assumes that the test length \( L \) is determined in advance, which is the case for fixed-length CAT.

Empirical weighting empirically chooses the weights so as to balance the contribution of both information pieces--whenever one information piece lags behind, more weight is assigned to it. Such an idea is reflected by the following definition of weight \( w \):

\[
w_1 = (u_{\theta} - x_{\theta}^{(k)}) / u_{\theta} \]
\[
w_2 = (u_{\alpha} - x_{\alpha}^{(k)}) / u_{\alpha} \]
\[
w = \frac{w_2}{w_1 + w_2}
\]

where \( u_{\theta} \) and \( u_{\alpha} \) are the pre-chosen upper bounds of the total information at \( \theta \) and \( \alpha \), respectively, and \( x_{\theta}^{(k)} \) and \( x_{\alpha}^{(k)} \) are the accumulated information at \( \theta \) and \( \alpha \) after \( k \) items have been
administered. The weight defined here has a built-in “minimax mechanism”—it tends to pick the items that maximize the information of the estimator (either $\tilde{\theta}$ or $\tilde{a}$) which is lagging behind.

Attribute-level weights attempts to exploit the relative importance of different attributes which can be reflected in the construction of $KL_j(\tilde{a})$. Assuming the current intermediate estimate is $a_u$, the attribute-weighted KL information index can be computed as

$$KL_j(\alpha_u) = [w_1, w_2, \ldots, w_K]$$

where $D_{juv}$ is defined as in CDI; $w_1, w_2, \ldots, w_K$ are the user-defined weights that reflect the relative importance of each attribute and $\sum_{i=1}^{K} w_i = 1$; $d_i(u,v)$, $i = 1, 2, \ldots, K$ is defined as being the number of different attributes between $a_u$ and $a_v$.

$$d_i(u,v) = \begin{cases} 1 & \text{if } \alpha_{ui} \neq \alpha_{vi} \\ d & \text{otherwise} \end{cases}$$

where $d$ is the Hamming distance, the total number of different attributes between $a_u$ and $a_v$.

It is interesting to discuss the connection between these two strategies. They offer different solutions to the question of how to obtain accurate estimations of both the cognitive pattern and summative ability at the same time. The second strategy suggests that an item with the maximum of the product of two pieces of information is the best candidate while the third strategy advocates in favor of the one with the largest value of the linear combination of the two
pieces with appropriate weights. However, both solutions share something in common: achieving the maximum composite index by selecting the one with large IRT information and CDM information indices. In fact, a simple trick can reveal that the information product algorithms can be deemed to be a special case of the linear combination methods. We apply logarithmic operations to the product, and then transform the information product algorithms into a linear combination of the logarithm of the two information pieces with equal weights. However, the incomparability issue of the two information indices has not been discussed in the context of the information product algorithms. One possible explanation is that the logarithm operation mitigates the incomparable scale issue, but additional empirical simulation studies are warranted to compare the two strategies with respect to both measure precision and item exposure control.

To conclude this review, I made a brief survey of most existing item selection algorithms in CD-CAT. A new taxonomy for the basic algorithms was proposed as an aid for evaluating these algorithms. Three important applications of the basic algorithms which address the practical needs in CD-CAT were identified. These include item exposure control, content management and dual-purpose algorithms for general ability and cognitive diagnostic information. A summary of algorithm development strategies for these three subtopics is provided with a reference to the traditional CAT algorithms whenever possible.

The new taxonomy and the summary of the strategies entail some interesting observations regarding the algorithms in CD-CAT in general. First, the information-based methods enjoy a dominant role, in comparison with other competing approaches such as the mathematical programming methods (the shadow test approach and the Monte Carlo method) and the rate function approach. Second, the most prevalent strategy for developing algorithms in CD-CAT is weighting, which involves taking any one of the basic algorithms as the kernel and
then weighting it with the required constraints (attribute coverage, item exposure, content, etc.). Even among the basic algorithms, the PWKL and MMGDI follow the same line of thought.

The new taxonomy and the summary of the strategies also points to several possibilities for future studies. First, the computational advantage of the KL-based algorithms can be further exploited. Second, flexible stratification algorithms can be proposed to control item exposure rates in CD-CAT. Third, constraint-weighted and/or stratification algorithms are possible for the content management. Last, but not least, information product algorithms can be proposed for the dual-purpose CD-CAT.
Chapter 3 High-efficiency KL-based Item Selection Algorithms for CD-CAT

The goal of this study is to introduce two high-efficiency KL-based item selection algorithms by modifying the two item discrimination indexes for test assembly in cognitive diagnosis proposed by Henson and Douglas (2005), Henson et al. (2008) and Rupp et al. (2010). In CD-CAT, high-efficient item selection can achieve higher precision in a fixed-length test or satisfy the pre-specified precision criterion using fewer items in variable-length test. The two newly-proposed item selection algorithms are the counterpart of MI in KL-based item selection methods, and can be as efficient as the MI in a short fixed-length test and more efficient than the MI and PWKL in a long fixed-length test. It requires fewer items to satisfy the pre-specified termination rule than the MI and PWKL algorithms in a variable-length test.

3.1 PWCDI and PWACDI

CDM discrimination index (CDI). Henson and Douglas (2005) proposed a CDM discrimination index (CDI) for facilitating test construction for cognitive diagnosis purposes. The building block for CDI is the D matrix for the \( j^{th} \) item, \( D_j \), and it is a \( 2^K \times 2^K \) matrix whose entries are the KL distance between the response distributions for all of the distinct cognitive patterns. Each \( u,v \) element in \( D_j \) is

\[
D_{juv} = E \left[ \log \left( \frac{P(X_j | \alpha_u)}{P(X_j | \alpha_v)} \right) \right] = \sum_{x=0}^{1} P(X_j | \alpha_u) * \log \left( \frac{P_{\alpha_u}(X_j | \alpha_u)}{P_{\alpha_v}(X_j | \alpha_v)} \right).
\]

A weighted mean can be computed as the CDI of the discriminating power among cognitive patterns for the \( j^{th} \) item.
\[ CDI_j = \frac{1}{\sum_{u \neq v} d(a_u, a_v)^{-1}} \sum_{u \neq v} d(a_u, a_v)^{-1} D_{juv}, \]

where \( d(a_u, a_v) = \sum_{k=1}^{K} |a_{uk} - a_{vk}| \) is the Hamming distance between two cognitive patterns. Henson et al. (2008) and Rupp et al. (2010) further developed the attribute-level CDI (ACDI) for attribute \( k \). This is defined as

\[ ACDI_{jk} = \frac{1}{2^k} \sum_{\text{all relevant cells}} D_{j,uv}. \]

All of the relevant cells are defined as the entries in the D matrix where only the \( k^{th} \) attribute is different for cognitive patterns \( a_u \) and \( a_v \). The ACDI for item \( j \), \( ACDI_j \), is simply the sum of \( ACDI_{jk} \) for \( k \) from 1 to \( K \):

\[ ACDI_j = \sum_{k=1}^{K} ACDI_{jk} = \sum_{k=1}^{K} \frac{1}{2^k} \sum_{\text{all relevant cells}} D_{j,uv}. \]

To simply the notation,

\[ ACDI_j = \frac{1}{2^K} \sum_{\text{all relevant cells}} D_{j,uv} \]

where all of the relevant cells refers to all of the entries for any pair of cognitive patterns with the Hamming distance of 1 being included. \( ACDI_j \) for item \( j \) is a partial sum of the matrix D in which only the entries for two cognitive patterns with the Hamming distance of 1 are included.

**Comparison of the KL item selection method and CDI/ACDI.** It is easy to observe that the KL item selection method can be constructed from the D matrix and it is the summation of the
column corresponding to the interim cognitive pattern estimate. The KL item selection is thus part of CDI without the cognitive pattern distance weighting. The first major difference lies in the fact that the CDI contains the information for all of the possible pairwise comparisons for the cognitive patterns, while the KL item selection method contains comparisons between all of the possible cognitive patterns with the estimated cognitive pattern. In this sense, CDI is a generalized KL item selection method. This difference makes CDI a superior item selection method, particularly during the early stage of CD-CAT where the cognitive pattern estimate is neither accurate nor reliable. The KL item selection method always assumes that the interim estimate of the cognitive pattern is accurate, whereas CDI does not have to invoke this assumption.

The second major difference is related to the first difference. The CDI of a candidate item remains unchanged while the KL item selection index of a candidate item does not since it might choose a different column of the D matrix based on the cognitive pattern estimate if CDI is used as a CD-CAT selection algorithm. The static nature of CDI can be anticipated because the original CDI were developed to assemble an optimal linear paper-and-pencil cognitive test for the general population. This is not desirable for CD-CAT, which is supposed to be adaptive. Some modifications are needed to make the D matrix “dynamic/adaptive” in order to construct a more effective dynamic version CDI. One possible benefit of the static nature of CDI concerns with the computation cost of CAT administration. Although the calculation of CDI is intensive, only one calculation of CDI values is needed before the CAT administration. During the CAT administration, however, one only needs to pick the candidate item with the largest CDI value.

ACDI is also a partial sum of the D matrix, but the former is constructed in a way that is different than the KL item selection method. The KL item selection method is an index for the
comparison between one particular cognitive pattern and all of the other possible cognitive patterns. ACDI, however, is an index for all of the possible pairwise comparisons among the cognitive patterns with a Hamming distance of 1. Other information with regard to those with a Hamming distance of $2, \ldots, K$ in CDI is omitted. In a sense, ACDI is another method of simplifying the complicated CDI as is the KL item selection method. It would be interesting to compare these two different ways of constructing a simplified index. It is worth noting that the two differences between CDI and KL are also applicable to ACDI, so it is also necessary for constructing a dynamic version of ACDI, which will be explained below.

**The PWCDI and PWACDI.** The key change is to incorporate the posterior distribution of cognitive patterns into the static D matrix. The posterior-weighted D (PWD) matrix can defined as

$$
PWD_{juv} = E_{a_u} \left[ \pi(a_u)^* \pi(a_v)^* \log \left( \frac{P(X_j | a_u)}{P(X_j | a_v)} \right) \right]
$$

$$
= \sum_{x=0}^{1} \pi(a_u)^* \pi(a_v)^* P(X_j | a_u)^* \log \left( \frac{P(X_j | a_u)}{P(X_j | a_v)} \right).
$$

PWD takes the varying importance of different cognitive patterns into account, and it follows the same reasoning as does PWKL. The only difference is that PWD calculates the pairwise comparisons and thus the weights, for both of the cognitive patterns involved which are considered. The posterior-weighted CDI (PWCDI) and posterior-weighted ACDI (PWACDI) can then be easily defined in the same manner as the original CDI and ACDI:

$$
PWCDI_j = \frac{1}{\sum_{u \neq v} d(a_u, a_v)^{-1} \sum_{a \neq v} d(a_u, a_v)^{-1} PWD_{juv}}
$$
\( PWACDI_j = \frac{1}{2^k} \sum_{all~relevant~cells} PWD_{j,uv} \).

We may construct a new item selection method which is similar to the PWKL and HKL from the PWD matrix in the way KL is obtained from the D matrix. The only difference between the HKL and its counterpart constructed from PWD is that there is an extra weighting factor for the estimated cognitive pattern. This weight remains constant over all of the candidate items within a particular item selection iteration, so this new method maintains the same ranking as the HKL, and this new item selection method can essentially be considered to be HKL. The corresponding new PWKL method can be obtained by removing the Hamming distance weighting. It can produce the same ranking as the PWKL and is thus equivalent to the PWKL.

**Computational Simplification.** The dynamic nature of the PWD matrix poses some computational challenges, particularly for CAT administration where real-time delivery is the key. Just like MI, PWCDI and PWACDI require a triple summation over \(2^k\) possible mastery profiles. This problem can be remedied easily using the same reasoning for the construction of the PWD matrix. We may partition the PWD matrix into the “dynamic” posterior weighting and the “static” D matrix. The “dynamic” posterior weighting requires updating using the cognitive pattern estimate in each iteration of CAT administration while the “static” D matrix remains constant over different iterations of CAT and examinees. The computational demands for these two parts are drastically different. Only one multiplication is needed for the calculation of weighting, while that for the static part is much more complicated due to the KL information involved. Translate this into mathematical language and the PWD can be reformulated as follows:
\[ PWD_{j,a,v} = \pi(a_u)^* \pi(a_v)^* \sum_{x=0}^1 P(X_j | a_u)^* \log \left( \frac{P_{a_u}(X_j | a_u)}{P_{a_v}(X_j | a_v)} \right), \]

and the matrix form is

\[ PWD = \pi^* \pi^T \cdot D, \]

where \( \pi \) is the vector for the posterior probability of cognitive patterns and \( \pi^T \) is its transpose. The symbol “\( \cdot \)” indicates the element-wise matrix multiplication. In practice, D matrix can be calculated beforehand and stored for the use in CD-CAT administration. In a matrix-oriented programming language such as Matlab, this simplification can improve calculation speed significantly. Compared with the computation simplification made for MI, algebraic manipulation is easier in PWD and the issues of negativity and scale change are also conveniently avoided. Therefore, PWCDI and PWACDI are a superior computational alternative to MI.

3.2 Simulation studies

Cognitive diagnostic models. Two common cognitive diagnosis models will be used in the simulation studies: the Deterministic Input; Noisy And gate (DINA) model (Junker & Sijtsma, 2001; Haertel, 1989) and the fusion model (Hartz, 2002). A Q-matrix is an essential element of most of the CDMs. For an item bank consisting of \( J \) items, the Q-matrix is a \( J \times K \) matrix of 1s and 0s that specifies the association between items and \( K \) attributes (Tatsuoka, 1983). The entry corresponding to the \( k^{th} \) attributes for the \( j^{th} \) item, \( q_{jk} \), is equal to 1 if item \( j \) requires the mastery of attribute \( k \), and \( q_{jk} = 0 \) otherwise.

The DINA model assumes that, in principle, an examinee must have mastered every
attribute associated with a particular item in order to respond correctly to that item (“And Gate”) while recognizing that examinees might respond in a contrary manner to predictions (“Noisy”). Certain examinees may answer an item incorrectly even though they have mastered all of the required attributes, whereas other examinees may answer an item correctly when they have not mastered at least one of the required attributes. Given these properties, the DINA model-predicted probability that examinee $i$ will respond correctly to item $j$ is defined by

$$P(Y_{ij} = 1 | \alpha_i) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}},$$

where $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik})$ is an indicator vector for examinee $i$’s cognitive pattern (i.e., $\alpha_{ik}$ equals 1 if examinee $i$ has mastered attribute $k$, and is 0 otherwise), $s_j$ is the probability that an examinee with all of the required attributes will “slip” and answer item $j$ incorrectly, $g_j$ is the probability that an examinee with at least one missing attribute has successfully “guessed” the correct answer, and $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{ij}} = 1$ if examinee $i$ has mastered all of the attributes measured by item $j$, otherwise, $\eta_{ij} = 0$.

The fusion model needs two types of item level parameters: (a) the baseline parameter $\pi_j^*$ represents the probability of a correct response to item $j$ if all of the measured attributes have been mastered, and (b) the penalty parameter $r_{jk}^*$ represents the probability of a correct response to item $j$ for not having mastered attribute $k$. The probability of a correct response conditional on the cognitive pattern and item parameters in the fusion model is defined as

$$P(Y_{ij} = 1 | \alpha_i, \pi_j^*, r_{jk}^*, c_j) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{q_{ij}(1-\alpha_{ik})} P_{c_j}(\theta_i),$$

where $P_{c_j}(\theta_i)$ follows the Rasch model with item difficulty $c_j$, and $\theta_i$ is the latent trait for
examinee \( i \) to account for the attributes which are not specified in the \( Q \)-matrix. In general, \( P_{c_j}(\theta) \) is set to 1 in order to make the specification of the \( Q \)-matrix complete (Wang et al., 2011; McGlohen & Chang, 2008; Henson & Douglas, 2005). In this case, the fusion model is called the noncompensatory reparameterized unified model (NC_RUM), which will be used in the current study.

### 3.2.1 Study I: fixed-length test

**Design.** A fixed-length CD-CAT simulation study was carried out in order to evaluate the efficiency of the new algorithms. Three factors were manipulated in the simulation study: test length (short versus long), item bank quality (high versus low) and item selection algorithms. The details were as follows:

**Examinees generation.** 3000 examinees were generated assuming that every examinee has a 50% chance of mastering each attribute. In a 5-attribute test, there were 32 distinct types of cognitive patterns which were assumed to be equally likely in the population.

**Item bank generation.** The item bank consisting of 500 items for 5-attribute DINA model is generated in the same manner as Cheng (2010). The Q-matrix used in this study is generated item by item and attribute by attribute. Each item has a 30% chance of measuring each attribute. This mechanism was employed to ensure that every attribute is adequately and equally represented in the item pool. The item parameters \( s_j \) and \( g_j \) were both generated from \( U(0.05, 0.25) \) for the high quality item bank and from \( U(0.10, 0.30) \) for the low quality bank.

**Test length.** The length of the short test was set as 5 items and the length of the long test was set to be 10 items.
**Item selection algorithms.** Six selection algorithms were compared in this study: KL, PWKL, ACDI, CDI, PWACDI and PWCDI. The comparisons of KL, ACDI and CDI can reveal the efficiency of ACDI and CDI if they were used as item selection algorithms against KL. The original ACI and CDI are also compared to PWACDI and PWCDI to demonstrate the effectiveness of the static-to-dynamic change of D matrix. The performance of PWACDI and PWCDI against PWKL is of the greatest interest for the current study.

**Evaluation criteria:** the efficiency of the algorithms can be demonstrated using by the high attribute correct classification rate (ACCR) and mastery pattern correct classification rate (PCCR). ACCR is defined as

\[ ACCR_I = \frac{1}{3000} \sum_{i=1}^{3000} I(\alpha_{ik} = \hat{\alpha}_{ik}) / 3000 , \]

where \( I \) is the indicator function and the PCCR is defined as

\[ PCCR = \frac{1}{3000} \sum_{i=1}^{3000} I(\alpha_i = \hat{\alpha}_i) / 3000 . \]

**Results.** The ACCR and PCCR for the six algorithms in various item banks and of various test lengths are presented in Table 3.1. In the short test under the high quality item bank, the PCCRs for ACDI and CDI, 0.187 and 0.245 respectively, were higher than that of KL, even though ACDI and CDI were not proposed as an item selection algorithm for CD-CAT. The table shows that PWACDI and PWCDI have a 0.773 PCCR, and outperform ACDI and CDI, which indicates that the modification proposed in this study is quite effective. More interesting results concern the PCCRs for PWKL, MI, PWACDI and PWCDI. The performances of PWACDI and PWCDI are indistinguishable from those of MI. PWACDI in particular achieves the same measurement precision as PWCDI. The lost information on other entries in the matrix D does not exert a
negative effect on item selection. As expected, there was a substantial difference of 0.154 between the PCCRs for these three algorithms and PWKL. Similar observations can be made easily for the low quality item bank.

Table 3.1 The ACCR and PCCR for Six Algorithms in Various Item Banks and of Test Lengths

<table>
<thead>
<tr>
<th>Test</th>
<th>Item</th>
<th>Selection</th>
<th>ACCR</th>
<th>PCCR</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>high</td>
<td>KL</td>
<td>0.581</td>
<td>0.964</td>
<td>0.548</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ACDI</td>
<td>0.720</td>
<td>0.989</td>
<td>0.492</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CDI</td>
<td>0.503</td>
<td>0.989</td>
<td>0.939</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PWKL</td>
<td>0.805</td>
<td>0.942</td>
<td>0.881</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MI</td>
<td>0.942</td>
<td>0.945</td>
<td>0.925</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PWACDI</td>
<td>0.911</td>
<td>0.943</td>
<td>0.915</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PWCDI</td>
<td>0.923</td>
<td>0.943</td>
<td>0.917</td>
</tr>
<tr>
<td></td>
<td>low</td>
<td>KL</td>
<td>0.571</td>
<td>0.943</td>
<td>0.587</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ACDI</td>
<td>0.690</td>
<td>0.967</td>
<td>0.492</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CDI</td>
<td>0.504</td>
<td>0.966</td>
<td>0.888</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PWKL</td>
<td>0.762</td>
<td>0.875</td>
<td>0.835</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MI</td>
<td>0.887</td>
<td>0.888</td>
<td>0.846</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PWACDI</td>
<td>0.838</td>
<td>0.892</td>
<td>0.856</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PWCDI</td>
<td>0.852</td>
<td>0.892</td>
<td>0.857</td>
</tr>
<tr>
<td>10</td>
<td>high</td>
<td>KL</td>
<td>0.591</td>
<td>0.989</td>
<td>0.753</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ACDI</td>
<td>0.719</td>
<td>0.991</td>
<td>0.937</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CDI</td>
<td>0.939</td>
<td>0.992</td>
<td>0.939</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PWKL</td>
<td>0.997</td>
<td>0.981</td>
<td>0.977</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MI</td>
<td>0.981</td>
<td>0.973</td>
<td>0.967</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PWACDI</td>
<td>0.976</td>
<td>0.980</td>
<td>0.978</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PWCDI</td>
<td>0.983</td>
<td>0.983</td>
<td>0.984</td>
</tr>
<tr>
<td></td>
<td>low</td>
<td>KL</td>
<td>0.578</td>
<td>0.981</td>
<td>0.825</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ACDI</td>
<td>0.688</td>
<td>0.972</td>
<td>0.883</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CDI</td>
<td>0.877</td>
<td>0.966</td>
<td>0.885</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PWKL</td>
<td>0.935</td>
<td>0.946</td>
<td>0.927</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MI</td>
<td>0.933</td>
<td>0.941</td>
<td>0.936</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PWACDI</td>
<td>0.925</td>
<td>0.939</td>
<td>0.933</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PWCDI</td>
<td>0.938</td>
<td>0.947</td>
<td>0.944</td>
</tr>
</tbody>
</table>

Note: Difference refers to the difference in PCCR compared to PWKL; KL = Kullback–Leibler Index method; ACDI = attribute-level CDM discrimination index; CDI = CDM discrimination index; PWKL = posterior weighted Kullback–Leibler information method; MI = mutual information method; PWACDI = posterior weighted attribute-level CDM discrimination index; PWCDI = posterior weighted CDM discrimination index.
In the long test, regardless of item bank quality, the PCCRs for ACDI and CDI are higher than is the case for KL. The difference between the ACDI/CDI and PWKL is still noticeable. The difference between MI and PWKL almost disappears and the difference between the PWACDI/PWCDI and the PWKL shrinks to about 0.01 to 0.03.

3.2.2 Study II: variable-length test

**Design.** Study II seeks to investigate the efficiency of the two proposed algorithms against PWKL in a variable-length test. A more efficient algorithm can terminate the test with fewer items than a less efficient algorithm in a variable-length test.

Three factors were manipulated in the simulation study: item bank quality (high versus low), the termination rule and three item selection algorithms (PWKL, PWACDI and PWCDI). Examinees and item banks were simulated in the same manner as in Study I. The termination rule for the variable-length test was proposed by (Tatsuoka & Ferguson, 2003) and stops the test when the probability of the cognitive pattern with the largest probability reaches a pre-specified value, such as 0.7, 0.8 and 0.9, in the current study.

**Evaluation criteria.** The efficiency of an algorithm in a variable-length test can be measured by the mean test length. Other descriptive statistics of the test length including the maximum, minimum and standard deviation were also reported.

**Result.** All of the descriptive statistics for three algorithms under various combinations of item banks and different criteria for the stopping rule are summarized in Table 3.2. Regardless of the item quality and stopping rule criterion, the mean test length for MI, PWACDI and PWCDI is smaller than that of the PWKL. Except in the low item quality bank, MI produces a larger mean test length when the stopping rule criterion is conservative (i.e., 0.8, 0.9). Item bank quality and
stopping rule criteria have some effect on MI and PWACDI, but under all of the conditions, PWCDI uniformly has about 0.5 items fewer than is the case for the PWKL.

**Discussion.** The PWKL is a well-established efficient Bayesian item selection algorithm in CD-CAT. The current study enhances this method by modifying the CDI which was originally developed for constructing paper-and-pencil diagnostic tests. Two simulation studies demonstrate that the new algorithms can improve the PCCR greatly in the short test and can satisfy the pre-specified stopping rule with fewer items than is the case in a variable-length test.

The key to the improvement is the information on all of the other possible cognitive patterns besides the estimated cognitive pattern. This is particularly important during the early stage of CD-CAT. The unreliability of the theta and cognitive pattern estimate is well recognized. Thus the item selection methods are not efficient during the early stage since the cognitive pattern estimate plays an important role in the calculation. The PWKL remedied this issue by incorporating the posterior distribution of the mastery profile, which is the usual Bayesian solution. The proposed methods provide a further improvement by taking advantage of all of the pairwise comparison of all possible cognitive patterns in the CDI, together with the Bayesian solution.
Table 3.2  The Descriptive Statistics of the Test Length for Variable-length Test

<table>
<thead>
<tr>
<th>Item quality</th>
<th>stopping rule</th>
<th>PWKL</th>
<th>MI</th>
<th>PWACDI</th>
<th>PWCDI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>max</td>
<td>min</td>
<td>Mean</td>
<td>SD</td>
<td>max</td>
</tr>
<tr>
<td>High</td>
<td>0.7</td>
<td>14</td>
<td>4</td>
<td>5.75</td>
<td>1.13</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>20</td>
<td>4</td>
<td>7.19</td>
<td>1.87</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>23</td>
<td>5</td>
<td>9.21</td>
<td>2.36</td>
</tr>
<tr>
<td>Low</td>
<td>0.7</td>
<td>24</td>
<td>4</td>
<td>8.75</td>
<td>2.81</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>31</td>
<td>5</td>
<td>10.33</td>
<td>3.27</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>29</td>
<td>5</td>
<td>12.10</td>
<td>3.58</td>
</tr>
</tbody>
</table>

Note: PWKL = posterior weighted Kullback–Leibler information method; MI = mutual information method; PWACDI = posterior weighted attribute-level CDM discrimination index; PWCDI = posterior weighted CDM discrimination index.
It is worth noting that there might be an issue of Q-matrix completeness for the short test length conditions in Study I. Chiu, Douglas, and Li (2009) stated that the necessary and sufficient condition for a complete Q-matrix was that it contained all the unit vectors. More specifically, the Q-matrix for an examinee in the short test length conditions is complete if it is a 5 by 5 identity matrix after some necessary column swapping. According to this rule, we may empirically check the completeness of the Q-matrices produced by all of the algorithms. The Q-matrices produced by MI, PWCDI and PWCDI are complete while those produced by PWKL are not, which can be additional evidence of the superiority of the new algorithms.

Among the questions that deserve further studies, the most interesting one is to investigate the efficiency of the two new methods if they are combined with the item exposure control mechanism. In practice, some statistical and non-statistical constraints are important such as the item exposure rates. Wang et al. (2011) proposed two restrictive stochastic item selection methods for addressing the issue of the tradeoff between measurement precision and item security based on the PWKL, namely the restrictive progressive method (RP) and the restrictive threshold method (RT). The PWACDI and PWCDI can be easily generalized into the RT and RP methods, and replace the PWKL index in RP and RT. It would be interesting to determine whether the RP and RT based on the MI and the two proposed methods can still maintain this advantage against the original RP and RT.
Chapter 4 Linear/Binary Stratification Methods for Fixed-length CD-CAT

Most of the efforts for the development of item selection algorithms were devoted to improving the accuracy of estimated cognitive pattern (Wang, 2013; Cheng, 2010; Cheng, 2009; Tatsuoka & Ferguson, 2003; Xu et al., 2003; Tatsuoka, 2002). Significant progress has been made in this respect, notably Cheng’s (2009) PWKL and HKL, which can achieve a recovery rate of above 0.9 for the entire cognitive pattern in the simulation studies. However, these works did not address the item exposure rate issue. This may lead to test security threat and underutilization of the item bank. Xu et al. (2003) pointed out that the exposure rates under the Shannon entropy procedure can be as high as 0.97 and nearly 60% of items in the bank were unused in the entire simulation. Although cognitive diagnostic testing is often of low-stakes nature, this cannot preclude the concerns for test security exclusively. The other issue is related to the prohibitively high cost of item development for cognitive diagnostic testing. The two issues can be addressed by equalizing item exposure rates without sacrificing measurement precision. Thus, it remains a question of how to balance the item exposure rates in CD-CAT. An efficient procedure that controls exposure rates must be developed before CAT is practically feasible in cognitive diagnosis.

The restrictive progressive method (RP) and the restrictive threshold method (RT) proposed by Wang et al. (2011) are the only two methods developed specifically for the purpose of addressing the item exposure control issue in fixed-length CD-CAT. They are built upon the PWKL index but also include additional stochastic components either in the item selection index or in the item selection procedure. They are information index-based methods, so they are computationally intensive and conceptually difficult to understand.
The current study proposes a stratification method for the purpose of addressing the item exposure issue, which is an extension of the $\alpha$ -stratified method in traditional IRT based CAT (Chang & Ying, 1999). Compared with the information based methods, the major advantages of the stratification method are two-fold: (i) the algorithm is simple and intuitive; and (ii) it is computationally much less intensive. However, the extension is nontrivial since cognitive diagnostic models are drastically distinct from IRT models and there is no straightforward counterpart available for the b matching step in CD-CAT as in CAT. This difficulty can be overcome by using the linear and binary search algorithms in computer science.

4.1 Linear and binary stratification strategies

**General framework of the stratification strategy**

Chang and Ying (1999) proposed the $\alpha$ -stratified selection method. The basic setup can be described as follows:

1) Partition the item bank into $M$ levels according to the item $\alpha$ values;

2) Partition the test into $M$ stages;

3) In the $M^{th}$ stage, select $n_m$ items from the $M^{th}$ level based on the similarity between $b$ and $\theta$ , then administer the items;

4) Repeat Step 3 from $k = 1, 2, ..., M$.

There are two essential elements in carrying out the $\alpha$ -stratified method: the item discriminatory index $\alpha$ used to partition of the item bank and the item difficulty parameter used to select items at every stratum via the b-matching method (Hulin, Drasgow, & Parsons, 1983; Weiss, 1974; Urry, 1971).
A general framework for the stratification method for CD-CAT can be set up in the similar manner as follows:

1) Partition the item bank into $M$ levels according to the item discrimination indices;

2) Partition the test into $M$ stages;

3) In the $M^{th}$ stage, select $n_m$ items from the $M^{th}$ level based on (cognitive) pattern-matching method;

4) Repeat Step 3 from $k = 1, 2,..., M$.

The following will explain how to identify the two essential elements of the stratification strategy in CD-CAT. Pattern-matching is the counterpart of b-matching in CD-CAT, and is not straightforward. Two versions of pattern-matching will be proposed using the linear/binary search perspective.

**Item discrimination index for CDMs**

The natural candidate for CD-CAT item bank stratification is the item discrimination indices for diagnostic classification models (DCM). Rupp, Templin and Henson (2010) provided a summary of the item discrimination indices for DCMs. They pointed out that there are two types of indices: the classical testing theory (CTT)-based global indices and the KL information-based indices. The classical theory-based global indices can be regarded as the counterpart of the $\alpha$ parameter in the IRT and thus, and can be used as the bank stratification criterion.

The underlying philosophy behind the classical theory-based global indices is the item discrimination in the CTT. It is measured as the difference in the proportion of respondents who respond correctly to an item in the upper tail of the total score distribution and the proportion of
those who respond correctly to the same item in the lower tail (e.g. the upper 25% versus the lower 25%). The item discrimination index can be denoted as \( d_j = P_u - P_l \) where \( P_u \) and \( P_l \) are the proportions of correct response to an item \( j \) for respondents in the upper and lower tail respectively. Thus the fundamental question of item discrimination for CDMs is “How well does this item help differentiate between respondents who have mastered ‘more’ attributes and those who have mastered ‘fewer’ attributes?” In alignment with the CTT index, a generic item discrimination index for item \( j \) in the context of CDMs can be defined as follows (Rupp et al., 2010):

\[
d_j = P_{u_j} - P_{l_j},
\]

where \( P_{u_j} \) is the probability of a respondent who has mastered several of attributes as measured by item \( j \) (i.e., in the upper tail) correctly answering that item and \( P_{l_j} \) is the probability of a respondent who has mastered fewer, or none, of the attributes as measured by item \( j \) (i.e., in the lower tail) correctly answering that item.

Take DINA as an example. The probabilities of a respondent in the upper tail and in the lower tail correctly responding to an item \( j \) is \( P_{a_u} = 1 - s_j \) and \( P_{a_l} = g_j \) respectively. Thus the item discrimination index for DINA model is

\[
d_j = P_{a_u} - P_{a_l} = (1 - s_j) - g_j,
\]

where \( s_j \) and \( g_j \) are the slipping and guessing parameters in the DINA model. Following the same reasoning, the index for other major CDMs can be obtained. Four of them including the deterministic input, noisy-or-gate (DINO) model and the compensatory reparameterized unified
model (C-RUM), are given by Rupp et al. (2010) and presented in Table 4.1. This index can be derived for the majority of the cognitive diagnosis models. Thus, the method proposed here can be readily extended to other models. However, in this study only the DINA and NC-RUM model will be used.

Table 4.1 Item Discriminatory Index for Major CDMs

<table>
<thead>
<tr>
<th>Model</th>
<th>Global item discrimination</th>
<th>A “good” item is one where…</th>
</tr>
</thead>
<tbody>
<tr>
<td>DINA</td>
<td>$d_{j,DINA} = (1-s_j) - g_j$</td>
<td>$s_j$ and $g_j$ are low</td>
</tr>
<tr>
<td>NC-RUM</td>
<td>$d_{j,NC-RUM} = \pi^*<em>j - \pi_j \prod</em>{a=1}^{A} r_{ja}^{*q_{ja}}$</td>
<td>$\pi^*<em>j$ is high and $r</em>{ja}$s are low</td>
</tr>
<tr>
<td>DINO</td>
<td>$d_{j,DINO} = (1-s_j) - g_j$</td>
<td>$s_j$ and $g_j$, the slipping and guessing parameters, are low</td>
</tr>
<tr>
<td>C-RUM</td>
<td>$d_{j,C-RUM} = \frac{\exp\left(\lambda_{j,0} + \sum_{a=1}^{A} \lambda_{j,1(a)} q_{ia}\right)}{1 + \exp\left(\lambda_{j,0}</td>
<td></td>
</tr>
</tbody>
</table><p>ight)} - \frac{\exp\left(\lambda_{j,0}\right)}{1 + \exp\left(\lambda_{j,0}\right)}$ | $\lambda_{j,0}$, the common intercept, is low and $\lambda_{j,1(a)}$, the attribute-specific slopes, are high |</p>

The linear and binary searching

This brief introduction to the linear and binary search in computer science heavily borrows from Rosen (2011) and Knuth (1973), but in a more accessible manner. The problem of locating an element in an ordered list occurs in many contexts. For instance, a program that checks the spelling of words searches for them in a dictionary which is an ordered list of words. Problems of
this nature are known as searching problems. The general setup for searching problems can be described as follows: Locate an element $x$ in a list of distinct elements $a_1, a_2, \ldots, a_n$, or determine that it is not in the list. The solution to this search problem is the location of the element in the list that equals $x$ (that is, $i$ is the solution if $x = a_i$) and is 0 if $x$ is not in the list. A searching algorithm is one that finds an item with specified properties among a collection of items. The linear and binary searching are two fundamental searching algorithms.

Linear searching or sequential searching is the simplest search algorithm and it is a special case of brute-force search. It is a method for finding a particular value in a list, which consists of checking each element, one at a time and in sequence, until the desired element is found. More specifically, the linear searching algorithm begins by comparing $x$ and $a_1$. When $x = a_i$, the solution is the location of $a_i$, namely, 1. When $x \neq a_1$, compare $x$ with $a_2$. If $x = a_2$, the solution is the location of $a_2$, namely, 2. When $x \neq a_2$, compare $x$ with $a_3$. Continue this process, and compare $x$ successively with each term of the list until a match is found where the solution is the location of that term, unless no match occurs. If the entire list has been searched without locating $x$, the solution is 0. There are two common cases for linear searching. The first one is the one where all of the ordered elements in the list are equally likely to be searched (uniformly distributed), and this is denoted as linear searching with equal probabilities. The second case is the one in which some elements on the list are more likely to be searched than others, and this is denoted as linear searching with unequal probabilities.

Binary searching is a dichotomic divide-and-conquer searching algorithm. Binary searching or half-interval searching algorithm proceeds by comparing the element located in the
middle of the ordered list, namely \( a_{n/2} \) if \( n \) is even or \( a_{(n+1)/2} \) if \( n \) is odd. The list is then split into two smaller sub-lists of the same size or two smaller sub-lists with one sub-list having one fewer element than the other. The search continues by restricting the search to the appropriate sub-list based on the previous comparison until the solution is obtained.

**Table 4.2 The Efficiency Analysis of the Linear and Binary Searching for \( n \) Objects**

<table>
<thead>
<tr>
<th></th>
<th>Average-case</th>
<th>Best-case</th>
<th>Worst-case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear searching</td>
<td>( n +2 )</td>
<td>1</td>
<td>2n</td>
</tr>
<tr>
<td>Binary searching</td>
<td>2\log n</td>
<td>1</td>
<td>2\log n</td>
</tr>
</tbody>
</table>

The efficiency of the two searching algorithms is evaluated by three types of complexity analysis: a worse-case, an average-case and a best-case analysis. A worst-case analysis refers to the largest number of operations needed to solve the given problem using this algorithm on an input of a specified size. A worst-case analysis tells us how many operations an algorithm requires to guarantee that it will produce a solution. Similar definitions can be given to the best case analysis and the average analysis. Assuming that the number of objects \( n \) is the power of 2 and the target object is in the list, the worst-case, the average-case and the best-case complexity for linear and binary searching are presented in Table 4.2. The average and worse-case analyses for the linear searching with equal probabilities show that the largest/expected number of operations required to complete the linear searching are proportional to \( n \) while those for the binary searching are proportional to \( \log n \). Therefore, when the list has a large number of elements, the binary searching algorithm is much more efficient than the linear searching.
algorithm. The performance of linear searching improves when the desired value is more likely to be closer to the beginning of the list than to its end in the case of linear searching with unequal probabilities. In particular, when the list of items is arranged in order of decreasing probability, and these probabilities are geometrically distributed, the required number of operations of linear search is only 1.

**Dynamic linear and binary search stratification strategy in CD-CAT**

From the linear/binary searching algorithm framework, the b-matching in CAT can be regarded as a form of dynamic linear searching, but this dynamic linear searching in CAT is different from that which is defined above. More specifically, only one search is performed in each iteration and then the posterior distribution of examinees is updated. This is equivalent to reordering the target list and putting the element (discretized $\theta$s in this case) with the largest probability in the first position. The next round of searching is carried out based on the reordered list. With the improving measurement precision, the b-matching can achieve the efficiency in the best-case analysis.

The idea of dynamic searching can easily be extended to CD-CAT. The linear version of the pattern-matching is straightforward. Assuming that the current posterior distribution of the cognitive profiles is accurate, the ideal candidate item must match with the cognitive patterns with the greatest probability (in fact, it is the current cognitive pattern estimate). The linear pattern-matching is very similar to the b-matching in which the interim estimate of alpha is the pinnacle of the posterior probability distribution and thus it can be regarded as the counterpart of b-matching in pattern-matching. Although the linear pattern-matching is simple and straightforward, there are several inherent defects associated with it. One prominent problem
with the linear pattern-matching is that the items with the corresponding cognitive pattern might not be available in the item bank. This is not unusual in CD-CAT and in fact, the Q matrix might not contain all of the possible cognitive patterns in practice. More commonly, it involves only one, two or three attributes. Thus, it creates some practical difficulties for the linear pattern-matching step. Another concern is its efficiency, particularly during the early stage of CD-CAT. Since little information on the posterior is obtained and the posterior is close to the uniform distribution, the linear searching strategy is essentially a linear searching with equal probabilities.

The binary version of the pattern-matching is that the ideal candidate item must have a Q matrix such that it can split the current posterior distribution in half with respect to the probability mass. Such splitting of two mutually exclusive groups is known as separation in the partially ordered set theory for CDMs. A similar algorithm, a halving algorithm, has been proposed from this theory (Tatsuoka & Ferguson, 2003; Tatsuoka, 2002). The splitting rule is very similar to the calculation of the η in the DINA model. The separation \( S_{jm} \) for item \( j \) and pattern \( m \) is defined as

\[
S_{jm} = \prod_{k=1}^{K} I\{q_{jk} \leq \alpha_{mk}\} = \begin{cases} 1 & \text{if pattern } m \text{ possesses all the required skills required for item } j \\ 0 & \text{if pattern } m \text{ lack of at least one of the required skills for item } j \end{cases}.
\]

The binary searching index \( B_j \) for \( t^{th} \) administration can then be formulated as

\[
B_j^t = \left| \sum_{S_{jm}=1} g(a_m \mid y_{t-1}) - 0.5 \right|
\]

where \( g(a_m \mid y_{t-1}) \) is the posterior probability for pattern \( m \) after \( t-1 \) items has been administered.

The binary searching index \( B_j \) can be interpreted as the distance between the splitting point of a
candidate item designated by its Q-matrix and the middle point. Therefore, the smaller the index for an item is, the better the item itself is. Particularly so in an ideal case (i.e., the posterior is just half split by the Q-matrix), it is 0.

Compared with the linear pattern-matching, the binary pattern-matching enjoys several advantages. First, it is free from the practical constraint that it might not find the pattern in the item bank. Second, it takes advantage of all of the information of the posterior distribution, unlike the linear pattern-matching which is only concerned with the single point which has the greatest probability. It can be expected that it can achieve higher efficiency in selecting items particularly during the early stage of CD-CAT. A note of caution is that during the late stage, the linear pattern-matching might be more efficient than the binary pattern-matching, provided that the precision of the cognitive pattern estimation is high and the linear search can obtain the best-case efficiency.

4.2 Simulation studies

4.2.1 Study I

Item bank and examinees generation. Study I is a simulation for a fixed-length CD-CAT of 15 items that aims to investigate the assertions regarding the linear and binary pattern-matching. In order to make the execution of linear pattern-matching possible, the Q matrix was generated by randomly selecting from the distinct patterns, so the number of items with each pattern was approximately equal. More specifically, an item bank consisting 480 items of the DINA model with 4 attributes was used in the study. Thus there were $2^4 - 1 = 15$ distinct types of cognitive patterns which were used to construct the Q matrix. The item parameters $s_j$ and $g_j$ were both generated from $U(0.05, 0.25)$. Since the test length was 10, the item bank was partitioned into 5
strata containing equal numbers of items according to the item discrimination index described above. The cognitive patterns for 2000 examinees were generated in the same manner, i.e., 16 distinct cognitive patterns were assumed to be equally likely in the population.

**Item selection algorithms.** Six item selection methods were used in this study including the random selection, PWKL, RT, RP, stratification with linear pattern-matching and stratification with binary pattern-matching. As regards RT and RP, $\beta = 2$ and the maximum item exposure rate was $r = 0.2$.

**Evaluation criteria.** These item selection algorithms were evaluated in terms of three aspects: estimation accuracy, item exposure balance and item bank usage. The evaluation criteria for estimation accuracy included recovery rates of attributes and cognitive patterns, the criterion for item exposure balance were the chi-square index that quantifies the equalization of exposure rates, and the ones for item bank usage were the number of items with less than 2% exposure rate and the number of items with exposure rate greater than 20%.

**Results.** The estimation accuracy, the measure of exposure balance and the item bank usage of each method can be found in Table 4.3. It is apparent that the purely information-based method (denoted as PWKL in the table) without any exposure control generates the highest precision, with the pattern recovery rate as high as .989. When exposure control is added, the pattern recovery rate decreases. Most significantly, the recovery rate for the stratification strategy with the linear pattern-matching can be as low as .315, which is even lower than the random selection method. As regards other methods, the recovery rates are .977, .953 and .962 for the RT, RP and the stratification strategy with the binary pattern-matching respectively. These are considered to be satisfactory for most diagnostic assessments.
As regards item exposure control, it is not surprising that the pure KL information-based method generates the largest chi-square value, which indicates that the exposure rate is quite skewed. When an item exposure control is adopted, the chi-square value decreases to a great extent. The result for the stratification strategy with binary pattern-matching is within the proximity of those of RT and RP, but the chi-square value of the linear search is still significantly larger.

Table 4.3 Recovery Rate and Exposure Balance Measures for DINA

<table>
<thead>
<tr>
<th>Item selection</th>
<th>Attribute Pattern</th>
<th>Exposure balance</th>
<th>Overused (&gt;0.2)</th>
<th>Underused (&lt;0.02)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>PWKL</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.995</td>
</tr>
<tr>
<td>RT</td>
<td>0.991</td>
<td>0.995</td>
<td>0.989</td>
<td>0.991</td>
</tr>
<tr>
<td>RP</td>
<td>0.986</td>
<td>0.985</td>
<td>0.979</td>
<td>0.985</td>
</tr>
<tr>
<td>linear</td>
<td>0.635</td>
<td>0.672</td>
<td>0.766</td>
<td>0.883</td>
</tr>
<tr>
<td>binary</td>
<td>0.99</td>
<td>0.992</td>
<td>0.988</td>
<td>0.986</td>
</tr>
<tr>
<td>random</td>
<td>0.867</td>
<td>0.860</td>
<td>0.866</td>
<td>0.866</td>
</tr>
</tbody>
</table>

The number of overused and underused items can provide more information about item bank usage. PWKL has the greatest number of overused and underused items. In particular, the underused items can account for as many as 358 out of 480 items, which indicates a huge waste of items. Linear searching does not perform well in this aspect either. There is only some minor reduction in the number of overused items, and the number of underused items can be as many as 259. RT, RP and binary search all did a good job in improving the item usage. The underused
items for binary, RT and RP searching are 34, 44 and 0 respectively. The overused items for the three methods are all 0.

4.2.2 Study II

Study II is a simulation for a fixed-length CD-CAT of 40 items that aims to investigate the stratification strategy’s performance in more realistic situations in which the Q matrix does not cover all the possible cognitive patterns. Thus the linear pattern-matching is not feasible and only binary pattern-matching was considered in this study.

Item bank and examinees generation. The item bank consists of 480 items from a 4-attribute fusion model and was generated in the same manner as was the case in Cheng (2010). The Q-matrix used in this study was generated item by item and attribute by attribute. Each item has a 20% chance of measuring each attribute. This mechanism was employed to ensure that every attribute was adequately and equally represented in the item pool. The item parameters $\pi^*_j$ and $r^*_jk$ were generated from $U(0.75, 0.95)$ and $U(0.2, 0.95)$, respectively. Since the test length is 40, the item bank was partitioned into 5 strata with equal number of items according to the item discrimination index described above. 2000 examinees were generated, and this assumes that every examinee has a 50% chance of mastering each attribute. For example, in a 4-attribute test, there were 16 distinct types of latent classes which were assumed to be equally likely in the population.

Item selection algorithms and evaluation criteria were the same as in Study I except for the stratification strategy with linear pattern-matching, which was excluded.

Results. The estimation accuracy, the measure of exposure balance and the item bank usage of each method are reported in Table 4.4. Although the model and the method of generating item
bank are different, Study II produced results that were comparable to Study 1. In terms of the estimation accuracy, except the random method, the binary pattern-matching stratification generated the lowest cognitive pattern recovery rate, 0.910, but that was not a huge loss of precision and is considered to be acceptable for low-stakes diagnostic purposes. In terms of exposure balance, the binary stratification strategy outperforms RT and RP with a chi-square value of 1.193. In terms of item bank usage, there were no overused items or underused items for the binary searching stratification strategy.

Table 4.4 Recovery Rate and Exposure Balance Measures for the Fusion Model

<table>
<thead>
<tr>
<th>Item selection</th>
<th>Attribute</th>
<th>Pattern</th>
<th>Exposure balance</th>
<th>Overused (&gt;0.2)</th>
<th>Underused (&lt;0.02)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>PWKL</td>
<td>0.999</td>
<td>0.999</td>
<td>1.000</td>
<td>0.999</td>
<td>0.998</td>
</tr>
<tr>
<td>RT</td>
<td>0.994</td>
<td>0.986</td>
<td>0.992</td>
<td>0.991</td>
<td>0.964</td>
</tr>
<tr>
<td>RP</td>
<td>0.979</td>
<td>0.970</td>
<td>0.981</td>
<td>0.981</td>
<td>0.921</td>
</tr>
<tr>
<td>binary</td>
<td>0.969</td>
<td>0.974</td>
<td>0.971</td>
<td>0.973</td>
<td>0.910</td>
</tr>
<tr>
<td>random</td>
<td>0.913</td>
<td>0.905</td>
<td>0.936</td>
<td>0.926</td>
<td>0.735</td>
</tr>
</tbody>
</table>

**Discussion.** This study attempts to address the test security and item bank usage issues in CD-CAT by extending the α-stratified method in CAT. This method enjoys the benefits of a simple algorithm and easy calculation as does the original method. Inspired by the linear and binary searching algorithms in computer science, the authors proposed two different methods of pattern-matching, the second step of the stratification method in CD-CAT. The results indicate that the
simpler stratification strategy with a binary pattern-matching algorithm achieves the goals as effectively as the information-based methods RT and RP.

The two simulation studies suggest that binary pattern-matching is more advantageous than linear pattern-matching. The biggest one is that binary pattern-matching can make use of all of the information from the posterior distribution while linear pattern-matching only use the point with greatest probability. Study 1 shows that pattern estimate accuracy in linear pattern-matching is even worse than when using the random method. We examined the items administered to individual examinees and found out about the first ten items were of the same Q matrix and it appeared that item selection was “stuck” in one pattern. During the early stage of CD-CAT, the pattern estimate was not accurate. As a result, items selected by the linear pattern-matching cannot provide useful information for updating the posterior adequately and can lead to the selection of uninformative items for next administration during the early stages. Given this vicious circle, both estimate accuracy and item exposure balance suffer. The binary pattern-matching can avoid this pitfall since it takes advantage of the entire posterior distribution and can update the posterior much more rapidly.

Several future studies are warranted. First, it is necessary to further develop an item discrimination index for other CDMs. This paper only provides this index for 4 major CDMs. If practitioners have to use some other models, we will have to work on the new item discrimination index following the same reasoning provided by Rupp et al. (2010). Some other alternative item discrimination indices are available, such as \((1 - s_j) \times (1 - g_j)\) for DINA (Wang et al., 2012). Further studies are needed to investigate the effect on the stratification strategy. Second, new item exposure control algorithms for variable-length CD-CAT are desirable. Both the current information-based methods and the stratification method are only feasible for the
fixed-length CD-CAT. Further adaptation is required to extend these methods to variable-length CD-CAT applications, which is the topic in Chapter 5.
Chapter 5 Item Exposure Control in Variable-length CD-CAT

Research has been conducted to investigate how to build CAT upon CDMs (Cheng, 2009; McGlohen & Chang, 2008; Tatsuoka & Ferguson, 2003; Xu, Chang, & Douglas, 2003; Tatsuoka, 2002). However, all of these studies focused on developing item selection algorithms and used a simple rule, namely the fixed-length rule, to terminate the CAT. This fixed-length termination rule may administer an unnecessarily long test to some examinees and an undesirably short test to others. As a consequence, it often yields different degrees of measurement precision for different examinees. In practice, it is desirable that all examinees have the same degree of measurement precision, which is a major advantage of CAT over non-adaptive testing (Weiss & Kingsbury, 1984). Hsu et al. (2013) recently advocated the development of a variable-length CD-CAT, and discussed the termination rule and item exposure control issues for this more flexible version of CD-CAT. Unfortunately, the item exposure control method proposed by them is extremely complicated, as discussed in Chapter 2. In this study, a simple extension of the stochastic methods for variable-length CD-CAT will be described and binary pattern stratification algorithm based on the searching algorithm from Chapter 4 will be proposed in order to deal with the issue of item exposure in variable-length CD-CAT.

5.1 Item selection algorithms for variable-length CD-CAT

Restrictive stochastic item selection methods for variable-length CD-CAT

The original RP and RT methods are only applicable to a fixed-length CD-CAT since the test length must be prescribed in advance, but some simple modifications can be made to the original restrictive stochastic item selection methods by defining a new progressive factor in order to
address item exposure issue in variable-length CD-CAT. The progressive factor in the original methods is \( x/L \) where \( x \) is the number of items currently administered to an examinee and \( L \) is the test length. Since a uniform test length for every examinee is not available in variable-length CD-CAT, this progressive factor cannot be defined in variable-length CD-CAT. One possible alternative which can be derived from the stopping rule proposed by Tatsuoka (2002) states that a CAT can stop if the posterior probability of the cognitive pattern with the greatest probability \( P_{current} \) reaches the prescribed value \( P_{1st} \) which usually is 0.8 or 0.9. A similar progressive factor can be defined as \( P_{current} / P_{1st} \), the ratio of the current posterior probability of the cognitive pattern with the largest probability \( P_{current} \) to the prescribed stopping rule \( P_{1st} \). Thus the modified RP (MRP) is formulated as follows:

\[
MRP_{-PWKL_j} = \left( 1 - \frac{\exp_j}{r} \right) \left[ (1 - P_{current} / P_{1st}) R_j + PWKL_j \times \beta P_{current} / P_{1st} \right].
\]

Similarly, the modified RT (MRT) can be denoted as

\[
\left[ \max(PWKL_j) - \delta, \max(PWKL_j) \right]
\]

with \( f(x) = (1 - P_{current} / P_{1st})^\theta \).

**Binary pattern stratification algorithm for variable-length CD-CAT**

The analogy between b-matching and linear searching is identified in Chapter 4. One major difference between them is that the b-matching in IRT is usually deterministic while linear/binary searching in CDMs is stochastic in general. In IRT, the difficulty parameter is on the same continuous scale as the ability parameter and the b-matching index is usually distinct for each item, so only one item is chosen. The linear/binary searching in CD-CAT, however,
involves selecting an item by determining the optimal Q matrix, which typically ends up identifying a group of items with one particular Q matrix. Therefore binary searching in CD-CAT can be regarded as a pattern stratification method. It involves partitioning the item bank according to distinct patterns in the Q matrix. Some randomness is naturally embedded in the process even without the item bank partitioning via the item discrimination index. This favorable property of the binary searching provides the possibility that the binary pattern stratification (BPS) method is a much simpler alternative for addressing the item exposure issue in the variable-length CD-CAT than is the case for the stochastic methods derived above and the SHTVOR.

5.2 Simulation studies

**Item bank and examinees generation.** The item bank consisting of 480 items for the 6-attribute DINA model was generated in the same manner as Cheng (2010). The Q-matrix used in this study was generated item by item and attribute by attribute. Each item has a 20% chance of measuring each attribute. This mechanism was employed to ensure that every attribute is adequately and equally represented in the item pool. The item parameters $s_j$ and $g_j$ were both generated from $U(0.05, 0.25)$. The cognitive patterns for 2000 examinees were generated in the same manner in which 16 distinct cognitive patterns were assumed to be equally likely in the population.

**Item selection algorithms.** Five item selection methods were used in this study including the random selection, PWKL (the baseline condition), SHTVOR, MRP, MRT and binary stratification algorithm. For the MRP and MRT, $\beta = 2$ and the maximum item exposure rate $r = 0.2$; for SHTVOR, $r=0.2$, average test overlap rate was set to be 0.01.
Termination rule. Tatsuoka (2002) recommended that variable-length CD-CAT can stop if the posterior probability value associated with any one cognitive pattern exceeds 0.8. A similar rule was adopted in this study, but the stopping criterion was set on three different levels: 0.7, 0.8 and 0.9.

Evaluation criteria. These item selection algorithms were evaluated in terms of two aspects: estimation accuracy and item bank usage. The evaluation criteria for estimation accuracy include recovery rates of attributes and cognitive patterns and those for item bank usage are the number of items with less than 2% exposure rate (underused items) and the number of items with more than 20% exposure rate (overused items).

Results. The estimation accuracy and test length statistics for all of the algorithms under different stopping criteria are presented in Table 5.1 and item bank use in Table 5.2. The PCCRs for all of the algorithms under each stopping criterion are close to each other, so the results for item bank use are comparable.

As regards the item exposure control, it is expected that the PWKL without the item exposure control mechanism generates the largest test overlap. When item exposure control is adopted, the test overlap rate for the four algorithms is reduced substantially. It is worth noting that all of the three new methods produce a similar test overlap rate to SHTVOR, even though there is no explicit mechanism for controlling it as does SHTVOR.

The number of overused and underused items can provide more information about item bank usage. PWKL has the greatest number of overused and underused items. In particular, underused items constitute as many as 412, 395 and 390 out of 480 items for the stopping criterion of 0.7, 0.8 and 0.9 respectively. There are also about 10 overused items for PWKL. Item
exposure control mechanism can improve these indices significantly. There are no overused items for any of the algorithms with item exposure control methods. In terms of underused items, they exhibit slightly different performances. The numbers of underused items for SHTVOR are 73, 54 and 25, respectively, for three stopping criteria. This is a result of the inability of the SH method to improve the utilization of underused items, even though there is an explicit control over the test overlap rate. By contrast, there are no underused items in MRP, MRT or the binary pattern stratification.

In summary, the three new methods can strike a nice balance between measurement accuracy and item bank use. All three of the new methods, particularly the binary pattern stratification method, are simpler to implement, in comparison with SHTVOR.
Table 5.1 The Measurement Accuracy and Test Length Under Different Stopping Criteria

<table>
<thead>
<tr>
<th>Item Selection</th>
<th>Attribute</th>
<th>Pattern</th>
<th>Test Length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>PWKL</td>
<td>0.942</td>
<td>0.941</td>
<td>0.961</td>
</tr>
<tr>
<td>SHTVOR</td>
<td>0.951</td>
<td>0.951</td>
<td>0.948</td>
</tr>
<tr>
<td>MRT</td>
<td>0.960</td>
<td>0.956</td>
<td>0.957</td>
</tr>
<tr>
<td>MRP</td>
<td>0.950</td>
<td>0.962</td>
<td>0.960</td>
</tr>
<tr>
<td>BPS</td>
<td>0.957</td>
<td>0.961</td>
<td>0.951</td>
</tr>
<tr>
<td>PWKL</td>
<td>0.963</td>
<td>0.971</td>
<td>0.973</td>
</tr>
<tr>
<td>SHTVOR</td>
<td>0.971</td>
<td>0.975</td>
<td>0.961</td>
</tr>
<tr>
<td>MRT</td>
<td>0.974</td>
<td>0.970</td>
<td>0.972</td>
</tr>
<tr>
<td>MRP</td>
<td>0.970</td>
<td>0.975</td>
<td>0.976</td>
</tr>
<tr>
<td>BPS</td>
<td>0.976</td>
<td>0.973</td>
<td>0.968</td>
</tr>
<tr>
<td>PWKL</td>
<td>0.984</td>
<td>0.988</td>
<td>0.993</td>
</tr>
<tr>
<td>SHTVOR</td>
<td>0.990</td>
<td>0.986</td>
<td>0.989</td>
</tr>
<tr>
<td>MRT</td>
<td>0.986</td>
<td>0.981</td>
<td>0.987</td>
</tr>
<tr>
<td>MRP</td>
<td>0.989</td>
<td>0.989</td>
<td>0.984</td>
</tr>
<tr>
<td>BPS</td>
<td>0.987</td>
<td>0.986</td>
<td>0.990</td>
</tr>
</tbody>
</table>
### Table 5.2 Item Exposure and Item Bank Use for Different Stopping Criteria

<table>
<thead>
<tr>
<th>Item selection</th>
<th>test overlap</th>
<th>Overused (&gt;0.2)</th>
<th>Underused (&lt;0.02)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PWKL</td>
<td>0.671</td>
<td>10</td>
<td>412</td>
</tr>
<tr>
<td>SHTVOR</td>
<td>0.033</td>
<td>0</td>
<td>73</td>
</tr>
<tr>
<td>MRT</td>
<td>0.061</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MRP</td>
<td>0.034</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BPS</td>
<td>0.029</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PWKL</td>
<td>0.622</td>
<td>13</td>
<td>395</td>
</tr>
<tr>
<td>SHTVOR</td>
<td>0.037</td>
<td>0</td>
<td>54</td>
</tr>
<tr>
<td>MRT</td>
<td>0.070</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MRP</td>
<td>0.038</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BPS</td>
<td>0.035</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PWKL</td>
<td>0.593</td>
<td>15</td>
<td>390</td>
</tr>
<tr>
<td>SHTVOR</td>
<td>0.045</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>MRT</td>
<td>0.081</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MRP</td>
<td>0.044</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BPS</td>
<td>0.042</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Chapter 6 Discussions and Future Directions

The current studies represent an effort to advance the feasibility of CD-CAT, an intelligent educational measurement tool that was envisioned as enhancing individualized learning over twenty years ago. On one hand, recent developments in cognitive diagnostic modeling and CAT have equipped psychometricians with tools they can use to embark on the development of CD-CAT. On the other hand, the CD assessment component in the PARCC and Smarter Balanced and the pedagogy issue in Moocs present great opportunities for CD-CAT.

The current studies have focused on the crucial element of CD-CAT: item selection algorithms. A comprehensive review of item selection algorithms in CD-CAT was conducted. Several new selection algorithms were proposed to address two important issues in CD-CAT: measurement efficiency and item exposure control. The PWCAI and PWACDI are computationally affordable and highly efficient alternatives to other information index-based algorithms. They can be used as a building block for the development of algorithms to deal with issues such as item exposure control, content balancing and duel-purpose CD-CAT in CD-CAT. All of these can develop into interesting future studies.

Although the binary stratification algorithm is a simpler alternative than the information index-based methods, current research has demonstrated its edge in balancing the item exposure rates in both fixed-length and variable-length CD-CAT. The stratification method has been well studied in traditional CAT. It offers an elegant solution to the item exposure control. It also has the potential to solve item selection problems when multiple constraints must be taken into account.
It appears that the two new proposed approaches in the current studies are competitors, but this is not necessarily the case, because each of them may be a better fit in different scenarios. In general, PWCDI and PWACDI are preferred when measurement efficiency is the top priority, while binary stratification is more advantageous in highly constrained CD-CAT. In some applications that have multiple constraints, there exists the possibility of using a hybrid version of the two proposed approaches to obtain a complementary effect.
REFERENCES


history and predictions for the future (pp. 195-226). Charlotte, NC: Information Age Publishing.


US Department of Education. (2015a). Elementary and secondary education act. from


