PSYCHOMETRIC TOOLS FOR FORMATIVE CLASSROOM ASSESSMENT: TEST CONSTRUCTION AND ITEM POOL DESIGN BASED ON COGNITIVE DIAGNOSTIC MODELS

By

Jiahui Zhang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Measurement and Quantitative Methods—Doctor of Philosophy

2019

ABSTRACT

PSYCHOMETRIC TOOLS FOR FORMATIVE CLASSROOM ASSESSMENT: TEST CONSTRUCTION AND ITEM POOL DESIGN BASED ON COGNITIVE DIAGNOSTIC MODELS

By

Jiahui Zhang

This thesis is concerned with the potential applications of cognitive diagnostic models (CDMs) with hierarchical attributes in supporting formative classroom assessments. The conventional CDM approach that requires large sample sizes is impractical in the classroom setting. Three are three CDM-based approaches that do not involve item calibration and thus are practical in the classroom setting: 1) CDM classifications using non-adaptive tests assembled from a calibrated item pool, 2) nonparametric classifications using non-adaptive tests based on CDMs, and 3) computerized adaptive testing (CAT) combined with CDMs (i.e., CD-CAT). Since most CDMs and their applications assume independent attributes, relevant model parameterizations, and the Q-matrix for hierarchical CDMs were discussed. Three studies were conducted to address the test construction and item pool design issues related to the three CDM-based approaches. Specifically, new indices based on the Kullback-Leibler information are proposed for non-adaptive test construction with a calibrated item pool. Different Q-matrix designs were explored for nonparametric classifications, and recommendations regarding the Q-matrix design were provided for teachers. For CD-CAT, an item pool design method based on simulation was proposed and evaluated. The intended contribution of the thesis consists of psychometric tools for the teachers that help them facilitate formative assessments in the classroom and instrumental guidelines for developers of formative assessment systems.

To my grandpa

ACKNOWLEDGEMENTS

I would like to thank my advisor and chair of my dissertation committee, Dr. William Schmidt, for his guidance and support. His great insight into education has illuminated my graduate study and will continue to guide me in my future career.

I would also like to thank Dr. Richard Houang, Dr. Tenko Raykov, and Dr. Amelia Gotwals for serving on my committee and offering constructive feedback for my proposal and dissertation draft.

The idea for this dissertation was born and developed in the many inspiring conversations I had with my mentors, Dr. Richard Houang and Dr. Leland Cogan. I have benefited tremendously from their knowledge and insight.

I am grateful to my family for their unconditional support and trust. Special thanks go to my husband, Qian Xu, who has made great sacrifices to support my pursuit of knowledge.

I would also like to acknowledge my adviser and mentor at Beijing Normal University, Dr. Tao Xin, who is like family to me. He led me into the field of educational measurement and always guided me in the right direction.

I would also like to thank many other friends and colleagues from Michigan State University, Beijing Normal University, NWEA, and ACT, who supported me along this arduous journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

Assessments are ubiquitous in most education systems. Educational assessments have the potential to provide feedback. The positive effect of feedback on learning has long been established in numerous studies in educational psychology, cognitive science, and learning science (e.g., Fyfe & Rittle-Johnson, 2015; Hattie & Timperley, 2007; Moreno, 2004). Therefore, various types of assessments have been widely used in schools to improve learning and teaching, which can be classified into summative assessment (providing a summary evaluation at the end of an educational program) and formative assessment (providing timely diagnostic information for learning and teaching during an educational program).

Despite its potential usefulness in learning, assessment or testing is among the most debated issues in public education. There have been concerns from teachers and parents that tests take up too much time from teaching and learning (Hefling, 2015; Walsh, 2017). A survey by the Council of the Great City Schools (CGCS) on large urban districts revealed that the average amount of testing time spent on required assessments among eighth-grade students in the 2014-15 school year was 4.22 days or 2.34% of school time (Hart et al., 2015). Examples of required assessments in the CGCS report are (i) state summative assessments for accountability (e.g., the Partnership for Assessment of Readiness for College and Careers [PARCC] assessments), (ii) state and local formative assessments, (iii) local end-of-course exams, and (iv) SAT, ACT, and Advanced Placement (AP) tests (optional in some places). Specific categories of students (including students with disabilities and English language learners) take (v) special assessments in addition to the required and optional tests.

Many of the required tests mentioned above are external, high-stakes, and summative measures for accountability purposes, fueled by important educational policy questions (Baker,

Chung, & Cai, 2016). These tests are not designed for assisting daily classroom learning and teaching. Even if diagnostic information can be extracted, it would be too late to be useful in the classroom (Hart et al., 2015). Too many of such tests would inevitably disrupt the learning process and may lead to problems such as teaching to the tests (e.g., Copp, 2018) and test anxiety (e.g., Schutz & Pekrun, 2007, p.3), both of which result from the misuse and abuse of educational assessments.

To address this issue, the U.S. Department of Education called on states to make assessments fewer and smarter in the Testing Action Plan (U.S. Department of Education, 2015). It calls for more classroom, low-stakes, and formative tests that are "smart" to provide timely feedback to learning and teaching and fewer external, high-stakes, and summative tests. We are entering a new era of K-12 assessments, where both accountability and instructional improvement are emphasized (Chang, 2012), and, correspondingly, both summative and formative educational assessments are required.

Research topics in the psychometric society echo the change in educational policies: the concepts of "assessment for learning" and "assessment as learning" have become popular as researchers emphasize on making assessment truly useful for learning (e.g., Bennett, 2011; Wilson, 2018). If tests are designed for producing feedback for learning and teaching and eventually integrate with the learning process, some problems of educational tests, including disrupting the learning process and teaching to the tests may be solved.

Renewed attention has been brought to the old concepts of classroom assessment and formative assessment (e.g., Bennett, 2015; Black & William, 2008; Gotwals, 2018; Shepard, 2018). Classroom assessment refers to the assessment taking place in the classroom and initiated by the teacher (Shepard, 2006; Wilson, 2018). Formative assessment is designed for providing timely and

constructive feedback that is closely connected to a curriculum and are based on students' learning history. It should be a thoughtful integration of the process to provide feedback and the appropriate measurement instrument or methodology (Bennet, 2011). This thesis concerns formative assessment in the classroom henceforth referred to as formative classroom assessment.

A huge responsibility for implementing formative classroom assessments lies on the shoulders of the teachers. Specifically, teachers need to take two iterated actions that are at the core of formative assessment: one is the identification of the gap between the desired goal and the learner's present state, and the other is the action taken to close the gap (Black & William, 1998). Identifying the gap is a measurement issue per se because the gap is the difference between a student's current state and the goal. However, many teachers do not feel adequately prepared for this assessment task (Mertler, 2003). Despite the increasing emphasis on educational measurement in policies and research, in some states, preservice teachers are not required to take specific coursework in classroom assessment or educational assessment in general (Campbell, 2013). As a result, teachers' formative assessment practices are not without struggles (Black & Wiliam, 1998; Gotwals, 2018). There is a gap between policy and research on one side and teachers' practice on the other side.

Although formative assessment is an attractive concept, the effectiveness of formative assessment hinges on its quality, not on its existence in the classroom (Black & Wiliam, 1998). As it takes time and resources to improve teacher preparation and professional development in assessment, there is an urgent need now to provide teachers with psychometric tools to facilitate formative assessment in the classroom. Teachers especially need assistance in constructing and delivering formative assessments as well as interpreting the results (Bennett, 2015; Campbell, 2013; Gotwals, 2018). Psychometric tools, which has guided and supported most standardized

testing programs, if used appropriately, can also help with constructing, delivering, and interpreting formative assessments (Bennett, 2011; Bennett, 2015).

Note that the use of psychometric tools, especially item response models, inevitably introduces some degree of standardization. Ideally, the teacher would develop his or her own formative assessment because it is the teacher who knows best the learning history of each student and the learning goals. Teachers' self-developed assessment is the exact opposite of standardization. With limited educational resources, therefore, we need to strike a balance between individualization and standardization when thinking of psychometric tools for formative classroom assessment.

In choosing appropriate psychometric tools (e.g., item response models) for formative classroom assessment, the best place to start is the validity, which is mainly decided by the usefulness of the feedback for formative purposes. Therefore, the first question we should ask is: What kind of feedback do teachers need? The needs of teachers were reflected in a survey conducted on a nationally representative sample of 400 elementary and secondary mathematics and English language arts teachers in the U.S. about a decade ago (Goodman & Huff, 2006; Huff & Goodman, 2007). The survey shows that norm-referenced information, standards-based information, and performance information at the item level from large-scale standardized assessments are of comparatively little interest to teachers because the information cannot be used directly in the instruction; what teachers need is detailed information about the strengths and weaknesses of individual students regarding specific knowledge, skill, and competencies.

Various methods have been proposed for providing diagnostic feedback. Some approaches involve extracting information from summative tests based on and calibrated with unidimensional item response theory (IRT) models (e.g., subscores; see Haberman, 2008). However, some

researchers caution that each purpose can be compromised if a single assessment is expected to serve multiple purposes (Pellegrino, Chudowsky, & Glaser, 2001, p2; Reckase, 2017). Although unidimensional IRT models have been successfully applied in summative tests aiming at selecting and differentiating, they might not be the most appropriate ones for formative purposes because the diagnostic nature of formative assessment usually suggests multidimensionality.

1.1 Psychometric solutions for formative classroom assessment

A family of measurement models—cognitive diagnostic models (CDMs; e.g., Rupp, Templin & Henson, 2010), which were developed for modeling diagnostic assessment data, are chosen for formative classroom assessment in this thesis. These models target multiple fine-grained latent constructs (referred to as attributes) that are typical in interim or formative assessments. With categorical latent variables, they are less affected by the high dimensionality as multidimensional IRT (MIRT) models and are more appropriate for finer-grained constructs than MIRT models (Templin & Bradshaw, 2013). The identification of these finer-grained constructs as well as their relationship is often based on cognitive or learning theories, and require collaborations between psychometricians and content experts. This construct space is similar to the concept of a domain in domain-referenced testing (Hively, 1974; Houang, 1980). The assessment developed based on CDMs can be integrated with the learning process through these constructs. Therefore, CDMs have the potential to be an essential part of the solution to formative classroom assessment.

Specifically, this thesis concerns formative classroom assessment that (i) can be linked to an instructional program lasting for several weeks, and (ii) can provide formative information for learning and instruction. The underlying measurement models are CDMs. Note that the assessment of interest does not intend to measure relatively stable traits such as ability or aptitude. Instead, the

targeted construct is the internalized knowledge or skills that the student acquires after particular several days' or weeks' instruction.

Although current CDM methods (i.e., calibration and classification) work well in large-scale assessments with hundreds or thousands of examinees and long tests, the application of CDMs in small-scale test settings in the classroom would be problematic due to limited testing time and the lack of response data required for reliable estimation (Chiu, Sun, & Bian, 2018). There are three alternatives to conventional CDM analysis, which do not require item calibrations and therefore, are practical in the classroom setting:

1) parametric classifications using non-adaptive tests assembled from a calibrated item pool (e.g., Henson & Douglas, 2005),

2) nonparametric classifications using non-adaptive tests based on CDMs (e.g., Chiu, Sun, & Bian, 2018), and

3) cognitive diagnostic computerized adaptive testing (CD-CAT; e.g., Chen, 2009).

The first two approaches use non-adaptive tests, which means the same test is given to all students in a classroom, so test construction is a critical question. The CD-CAT approach uses adaptive tests that are tailored to the state of individual students, the success of which depends on a well-designed item pool. How to design the appropriate item pool for a CD-CAT program remains a research question. Responding to practical needs and gaps in the literature, this thesis addresses the test construction and item pool design issues for these three approaches.

These CDM-based approaches are intended for facilitating formative classroom assessment, which is related to domain-referenced testing and curriculum-based assessment. Therefore, the rest of Chapter 1 reviews these related concepts as well as the broader concept of educational assessment and the so-called next-generation assessment.

The next chapter reviews the fundamentals and previous studies of the three CDM-based approaches with a focus on CDMs with hierarchical attributes. Chapter 3 deals with parameterizations and Q-matrices of CDMs with hierarchical attributes, followed by three chapters addressing three research questions related to the test construction or item pool design issues.

1.2 Related concepts

Formative classroom assessments belong to the broader concept of educational assessment or achievement assessment. The terms educational assessment and achievement assessment have been used interchangeably in the literature. More specifically, Mislevy, Steinberg, and Almond (2003) in their seminal work on assessment design defined an educational assessment to be "*a machine for reasoning about what students know, can do, or have accomplished, based on a handful of things they say, do, or make in particular settings.*" Baker, Chung, and Cai (2016) offered a broader construction: "*A test or an assessment consists of a systematic method of gaining a sample of information about people or programs so as to draw inferences about examinees' knowledge, characteristics, or propensities.*" The definition of Mislevy et al. (2003) focuses on the types of inferences made from the assessment, and the definition of Baker et al. (2016) also highlights the process of making inferences (i.e., via sampling) in educational assessment.

The history of educational assessment has been intertwined with that of psychological assessment. Their connection can be seen from the title of the *Standards for Educational and Psychological Testing* (AERA, American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1985, 1999, 2014) as well as journals and books (e.g., *Educational and Psychological Measurement*). The first generation of standardized achievement tests was developed in the same period and by the same researchers as IQ tests were (Sheperd, 2006). As a result, educational assessments and psychological assessments tend to have the same

item formats and often utilize the same statistical models (e.g., item response theory models), with both having roots in individual differences psychology. In this section, the discussion is limited to IRT-based assessment because most large-scale or commercial achievement tests (e.g., PARCC, NAEP, PISA, SAT, ACT) use IRT models.

More and more researchers in the educational assessment field, however, have realized the critical differences between educational and psychological assessments despite their entwined histories. Among the most discussed issues is the definition of the measured domain, the stability of the unobserved constructs, the dimensionality of the construct space, the normality assumption, and the purpose of assessment.

The unobserved constructs measured in psychological assessments are usually not well-defined. As noted by Brody (2000, p.39), researchers know how to measure the construct called intelligence, but they still do not know what has been measured; what the IQ test does, as a result, is merely trying to differentiate people along a hypothetical scale. In some sense, the test that is supposed to measure intelligence defines what intelligence is. This is not true in education where domains could be well defined according to the instructional goals of a specific instructional program. However, the measured domains are not well delineated for some educational tests (Baker, 2009). In such cases, it can be said that we know how to measure achievement, but we do not know what has been measured, particularly, if and when educational assessments follow the tradition of psychological measurement.

The unobserved constructs in psychological assessments are usually stable traits, such as intelligence, self-efficacy, or personality. These traits are assumed, or believed, to remain stable for an extended period. The purpose of psychological assessments is to reflect the relative location of a person regarding this latent trait, and improvement or change within a short period is not

expected (Baird, Andrich, Hopfenbeck, & Stobart, 2017). However, examinees in educational assessments are expected to show changes in their educational attributes and accomplishments within a short period, which is the primary purpose of any educational program.

The existence of content blueprints complicates the definition of the unobserved constructs in educational assessment. Unlike a psychological test, an educational test is usually developed based on a content blueprint (Luecht, 2013; Reckase, 2017). A content blueprint is usually constructed as a set of test specifications that is independent of the psychometric modeling of test responses (Luecht, 2013). However, a test blueprint with multiple content domains may suggest, and be consistent with, a multidimensional space (Reckase, 2017). Besides content dimensions, cognitive dimensions have also been considered for educational assessments, which further complicates the dimensionality issue (George & Robitzsch, 2018; Harks, Klieme, Hartig, & Leiss, 2014). In an analysis of TIMSS data, content dimensions are number, geometry, and data, and cognitive dimension are knowing, reasoning, and applying (George & Robitzsch, 2018).

For most of the commercial achievement tests, the interpretation of a test score is directly based on the assumed normal distribution of underlying stable psychological characteristics (Baker, 2009). This normality assumption is another inheritance educational measurement inherited from the psychological measurement under the general framework of latent variable modeling (Baker & Kim, 2004). Consistent with the interpretation of scores, a normal distribution is usually assumed in IRT modeling for the unobserved construct. Specifically, the normal distribution is used (i) in the integration step in item calibration and (ii) as a prior distribution in Bayesian IRT-based scoring (Baker & Kim, 2004). While the normality assumption may work well for a variety of stable psychological traits (e.g., intelligence, self-efficacy), whether it is suitable for the

measurement of learning or mastery of educational attributes is questionable (Bloom, 1968; Baker, 2009).

Educational assessment designers, following the guidelines developed for psychological assessments, tend to optimize the test for detecting differences among examinees. It would work well if the goal is selection. However, the test development guidelines may need some adaptations when we consider the purpose of improving student learning because the differences between different test scores could be trivial regarding the subject matter (Bloom, 1968).

One characteristic of educational assessments that is different from psychological assessments, however, is the existence of many dichotomies, such as classroom assessment versus external tests, formative versus summative assessment, domain-referenced (or criterion-referenced) versus norm-referenced testing (assessment).

1.2.1 External and classroom assessment

External assessments are constructed outside of the classroom by measurement and subject experts and are often fueled by educational policies (Baker, Chung & Cai, 2016), also referred to as the large-scale standardized assessments. There is a rich literature on the theories and practices of external assessments. They have served well the purpose of selection and accountability over the past decades. However, the effects of external assessments on learning are difficult to establish (Wilson, 2018).

Educational assessments can be divided into classroom assessments and external assessments, depending on the administration of the assessments. Teachers usually create and grade classroom assessments based on particular instructional goals, and they make short-term decisions based on assessment results (Hanna & Dettmer, 2004, p. 8). Classroom assessments may also be developed out of the classroom but initiated by teachers or students in the classroom.

Classroom assessments, when used in a constructive way by teachers, can send the message to students telling them what is important (Nitko, 2001), and have been shown to have a substantial impact on student success (Shepard, 2006; Wilson, 2018). Some researchers believe that we can make measurement truly important for education through classroom assessments (Wilson, 2018).

1.2.2 Summative and formative assessment

The dichotomy of formative assessment versus summative assessment has been proposed for decades. While great improvement has been seen in the practices and research of summative assessment over the past few decades, formative assessment mostly appears as the subject of theoretical discussion (Scriven, 1967; Bloom, 1968; Bloom, Hastings, & Madaus, 1971). Scriven (1967) and Bloom (1968) were among the first to use the terms "formative evaluation" and "summative evaluation." A summative evaluation judges what students have mastered at the end of an educational program (Bloom, 1968). Defining formative assessments, however, can be much more complicated: There has been debate over conceptualization of formative assessment as a test or a process (Bennet, 2011). For Bennet (2011), neither side of the argument can provide a full picture of formative assessments: He defined formative assessment to be a thoughtful integration of process, on the one hand, and methodology or instrumentation, on the other hand. Other researchers put more emphasis on the process part (e.g., Furtak, Circi, & Heredia, 2018; Gotwals, 2018).

Recently, formative assessment is receiving renewed attention (Bennet, 2011, p. 5). Since formative assessments generally take place in the classroom as a type of classroom assessments, teachers need to take many responsibilities. However, it remains a challenging task for teachers to learn how to do formative assessments (Bennet, 2011; Furtaka, Circib &, Heredia, 2018; Gotwals, 2018; Shavelson, 2008). Teachers need guidance and assistance in various aspects of assessments,

including goal setting, extracting information, providing feedback, and using feedback to modify instructions (Gotwals, 2018, p.157). Bennett (2011, p. 18) argued that teachers need "deep cognitive-domain understanding" and "knowledge of measurement fundamentals" in addition to "pedagogical knowledge", in order to be able to realize effective formative assessments. However, even if teachers can acquire all the knowledge, understanding, and skills needed for formative assessment, they still need a substantial amount of time to put them into practice (Bennet, 2011).

1.2.3 Domain-referenced and norm-referenced testing/interpretations

Another well-known contrast in educational measurement is between domain-referenced (or criterion-referenced) testing and norm-referenced testing (Hively, 1974). Norm-referenced testing (NRT) has its roots in the psychological measurement of individual differences. NRT goes hand in hand with latent trait modeling (Hively, 1974; Houang, 1980). The test construction for NRT based on latent trait modeling places great emphasis on correlation or the so-called internal consistency among a set of items, which plays a significant role in the decisions of including or excluding certain items (Hively, 1974; Houang, 1980). However, this test construction procedure may pose a danger to the validity of measurement because 1) variables that are conceptually disconnected can be correlated (Baird et al., 2017) and 2) the obtained set of items may not be a representative sample from the targeted domain (Houang, 1980).

Domain-referenced testing (DRT), in contrast, bears more educational considerations. More emphasis is placed on validity instead of reliability. Much research is devoted to the discussion of the domain and item sampling within the domain (Baker, 1974; Hively, 1974; Millman, 1974). A domain can be defined by an explicitly specified set of items (Hively, 1974) or a set of rules according to which a large number of test items could be generated (Baker, 1974). A complex domain can be divided into sub-domains. The examinee's measurement of principal

interest in NRT is the examinee's score over all items in domain or sub-domain (Brennan, 1981; Hively, 1974). This score, referred to as the domain score (or the sub-domain score), cannot be directly obtained because it is impossible to administer all the items in the domain (or sub-domain). It can be estimated by the examinee's observed percent of correct responses on a set of items if the set is a representative sample (Brennan, 1981). Estimates for large domains may be obtained by stratified sampling over their constituent sub-domain, and diagnostic profiles may be gathered by sampling within sub-domains (Hively, 1974). IRT-based estimators are available for domain or sub-domain scores, given a large set of calibrated items (Bock, Thissen, & Zimowski, 1997). For a complicated domain, the set of sub-domain scores serves a diagnostic profile (Hively, 1974); alternatively, one can assign sub-domain scores weights to calculate a single domain score (Millman, 1974). The estimated domain or sub-domain scores are then compared to some criterion to decide whether mastery is achieved. In contrast to the two-stage methods, Houang (1980) took a latent class approach to estimate the mastery of a simple domain.

The concept of DRT as an assessment type lost its popularity after the 1970s. Since the 1974 *Standards for Educational and Psychological Tests*, the distinction between two types of test score interpretations—criterion-referenced and norm-(or criterion-)referenced interpretations—have received more attention. Instead of differentiating two different types of assessments (i.e., NRT and DRT), test developers draw from both test development perspectives to ensure the reliability and validity of measurement (Brennan, 2006). Although most standardized testing programs are designed to primarily provide norm-referenced interpretations, there has been an increasing need for domain-referenced or criterion-referenced interpretations.

1.2.4 Curriculum-based assessment

Educational assessments are based on a specific curriculum or not. To be useful for learning, however, assessment needs to be integrated into a coherent process of assessment, instruction, and curriculum based on learning theories (Black, Wilson, & Yao, 2011; Shepard, Penuel, & Pellegrino, 2018). This is especially true for formative classroom assessment. If the assessment is not aligned to the curriculum that students are learning, the validity of the formative feedback will be in doubt.

A link between curriculum and achievement assessment has been well established in the international assessments led by the International Association for the Evaluation of Educational Achievement (IEA). The curriculum-achievement alignment constitutes a vital part of the validity evidence for the subject achievement tests. The validity check (by comparing assessment items with the curriculum students have experienced) has been carried out in some form in all IEA studies (Cogan & Schmidt, 2019). For example, teachers provided validity check on the test items in the pilot study and the First International Mathematics Study (FIMS) and in the second studies, SIMS and SISS (Husén, 1967a; Keeves, 1974; Travers & Westbury, 1989). The 1995 Third International Mathematics and Science Study (TIMSS-95) conducted a more extensive curriculum analysis, and provided evidence for the relationship between assessment, instruction, and curriculum (Schmidt & McKnight, 1995; Schmidt, Jorde, et al., 1996; Schmidt, McKnight, Valverde, Houang, & Wiley, 1996).

A curriculum is structured around subject content. Taking the subject of mathematics as an example, as Schmidt and his colleagues put it, “mathematics, even circumscribed by what is taught in school, encompasses a very large content domain.” The question is then how to model curriculum-sensitive content in the psychometric model for curriculum-based assessment. Under the typical unidimensional IRT modeling framework, content exists in the form of content

constraints, independent of the measured construct (e.g., Kingsbury & Zara, 1991; van der Linden, 2005a). The separation of the measured construct and the curriculum-sensitive contents makes it difficult, if not impossible, to extract formative feedback from the test data regarding the contents.

1.2.5 Next-generation assessment

Since we entered the new millennium, there have been increasing discussion over the so-called next-generation assessment. Questions like "Are we entering a new era for the educational assessment?" are being asked. In the discussion of the next-generation assessment, researchers and measurement practitioners attempt to respond to the critiques on educational measurement mentioned earlier and the needs from learners, parents, and teachers (e.g., Bennett, 2011; Conley, 2018; Embretson, 2003; Heritage, 2010).

A lengthy, but not exhaustive list of next-generation assessment topics includes formative assessment (e.g., Gorin, & Mislevy, 2013; Heritage, 2010), assessment of new constructs such as critical thinking (e.g., Liu, Frankel, & Roohr, 2014), technology-based assessment (e.g., Beatty & Gerace, 2009; Bennett, 2015; Mislevy, 2016), classroom assessment (e.g., Shepard et al., 2018), personalized testing and learning (e.g., Chen, Li, Liu, & Ying, 2018; Clark, 2016), integration of learning and assessment (e.g., Baird et al., 2017), and automatic item generation and scoring (e.g., Bennett, 2015; Gierl & Lai, 2012).

Chapter 2 Literature review of CDM-based approaches

This chapter provides brief literature reviews for the basics of CDM, nonparametric classifications based on CDM, and CD-CAT, which form the foundations of the three CDM-based approaches for formative classroom assessment proposed in Chapter 1.

The CDM-based test construction begins with the identifications of the attribute profile space and the Q-matrix characterizing the relationship between items and attributes (described in detail in Chapter 2). The attribute profile space defines the domain in the language of domain-referenced testing. Test construction based on CDMs has many similarities with domain-referenced testing (Hively, 1974; Houang, 1980). The identifications of the relationships between attributes and items usually depend on cognitive theories and learning theories. In this way, the assessment can be integrated with the learning process.

2.1 CDM

CDMs (cognitive diagnostic models), also known as diagnostic classification models, belong to the confirmatory or constrained latent class modeling framework in which individuals are classified into groups defined by combinations of categorical (usually binary) latent variables (Rupp, Templin & Henson, 2010). The categorical unobserved variables that define the measurement constructs underlying a CDM are often referred to as attributes (Tatsuoka, 1983, 1990), elsewhere called finer-grained proficiencies (de la Torre, & Karelitz, 2009) or facets (Henson, DiBello, & Stout, 2018).

Macready and Dayton (1977) and Houang (1980) were among the first to apply latent class models using only one dichotomous trait to measure mastery of a simple domain. Later, the works of Tatsuoka (1983) and Leighton, Gierl, and Hunka (2004) involve more complex domains with multiple attributes, and they introduced the concepts of Q-matrix and attribute hierarchy. In the

past three decades, a large number of CDMs that employ item response functions (IRFs) and explicit Q-matrices have been proposed and studied intensely (Rupp, Templin, & Henson, 2010; Templin & Bradshaw, 2014) in response to the pressing demand for individualized diagnostic information in education (Center for K-12 Assessment and Performance Management at ETS, 2014; U.S. Department of Education 2014).

2.1.1 Attributes

Since the introduction of attributes to diagnostic assessments by Tatsuoka (1983, 1990), the terminology of attributes has been used in the CDM literature to refer to the unobserved variables that the test aims to measure. Long before the time of diagnostic assessment, Guttman (1944) used "attribute" interchangeably with "qualitative variable" (i.e., categorical variable). Tatsuoka (1990) provided a broad definition of attributes as "production rules, procedural operations, item types, or, more generall, any cognitive tasks" (p. 465). Embretson (1995) viewed attributes as "sources of cognitive complexity" in test performance, which may consist of both cognitive and content components. Leighton, Gierl, and Hunka (1999) defined attributes as the procedural or declarative knowledge needed to perform a task in a specific domain. Most of the above definitions include both cognitive and content components.

In an educational setting, possessing an attribute is often referred to as mastery of an attribute, and lacking an attribute is referred to as non-mastery (Templin & Bradshaw, 2014). Like most CDM research, we restrict the scope of this thesis to attributes with two levels, so that $\alpha_k = 1$ indicates mastery of attribute k and $\alpha_k = 0$ indicates non-mastery of this attribute.

An attribute profile (Templin & Bradshaw, 2014), which is also referred to as an attribute pattern (Ma, Iaconangelo, & de la Torre, 2015) or attribute mastery pattern (Henson & Douglas, 2005), is a specific combination of attribute mastery and non-mastery, with each combination

representing a unique latent class of examinees. Attribute profiles are denoted by column vectors $\boldsymbol{\alpha} = (\alpha_1, \ldots \alpha_k, \ldots, \alpha_K)^T$, where $\alpha_k \in \{0, 1\}$ indicates the absence or presence, respectively, of the kth attribute (mastery vs. non-mastery), and the superscript T denotes transpose.

2.1.1.1 Interaction among attributes in an item

CDMs can be categorized as noncompensatory or compensatory models based on the assumptions about how attributes interact with each other to affect the probability of an item response. According to DiBello, Roussos, and Stout (2006), a noncompensatory (or conjunctive) model assumes that lacking competency on any required attribute poses a severe obstacle to successful performance on the task. In other words, successful performance on a task requires mastery of all the required attributes; mastery of some of the required attributes does not compensate for the non-mastery of other required attributes. The terms of conjunctive models and noncompensatory models are often used interchangeably. Opposite to the noncompensatory nature, compensatory interaction of attributes means that mastering one required attribute can compensate for nonmastery of other required attributes. An extreme case of compensatory models is a disjunctive model in which mastering each subset of the required attributes would lead to the equally high probability of a correct response (DiBello, Roussos, & Stout, 2006).

2.1.1.2 Interdependencies among attributes

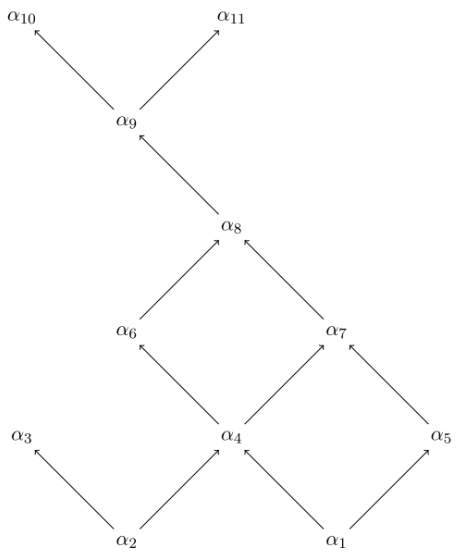
Most CDMs assume independent attributes (Rupp et al., 2010). Nevertheless, there are cases in which data analysis suggested the presence of interdependencies among attributes (Templin & Bradshaw, 2014). To account for the relationships between attributes, de la Torre and Douglas (2004) proposed a higher-order model linking the categorical attributes to an underlying multivariate normal distribution. The interdependencies among attributes are reflected in the correlated dimensions of the multivariate normal distribution. Another approach to modeling the

attribute relationships is to impose a hierarchical structure, in which mastering an attribute could be prerequisite to mastering another attribute (Leighton et al., 2004; Tatsuoka, 2009; Templin & Bradshaw, 2014). This thesis adopts the hierarchical approach, which is reviewed in more details below.

A hierarchy of attributes specifies the relationship between each pair of attributes. For attribute i and attribute j, if $P(\alpha_j = 1|\alpha_i = 0) = 0$, attribute i is called a prerequisite of attribute j. Suppose there are three attributes in a linear relationship. We have $\mathrm{P}(\alpha_2 = 1|\alpha_1 = 0) = 0$, $\mathrm{P}(\alpha_3 = 1|\alpha_1 = 0) = 0$, and $\mathrm{P}(\alpha_3 = 1|\alpha_2 = 0) = 0$.

Attribute hierarchies are often visualized by a tree graph with a set of attributes connected with arrows. An arrow that points from attribute i to attribute j means that mastering attribute i is a prerequisite to mastering attribute j (Gierl, Leighton, & Hunka, 2000; Köhn & Chiu, 2018; Leighton et al., 2004). Attribute i is a lower-level attribute, and attribute j is a higher-level attribute in this case.

These pair-wise prerequisite relationships can be formally defined by a K-by-K binary matrix called the adjacency matrix (A-matrix), in which K is the number of attributes (Tatsuoka, 1983, 2009; Gierl et al., 2000). The A-matrix represents the direct relationships among attributes usually illustrated by one-way arrows. The (i, j)th element of the A matrix indicates whether attribute i is directly connected in the form of a prerequisite to attribute j. The diagonal elements of the A-matrix are zeros. The following is an example of a complex hierarchy in Köhn and Chiu (2018) with its 11-by-11 A-matrix.



A=

α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}
0	0	0	1	1	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0

Figure 1: A complex example of attribute hierarchy in Köhn and Chiu (2018)

In an attribute hierarchy, there are direct and indirect relationships. A direct relationship is characterized by a one-way arrow. In the example below, α_1 and α_4 has a direct relationship because there is an arrow pointing from α_1 to α_4. An indirect relationship can be found between α_1 and α_6, which are connected through two arrows and α_4 in between.

If compared to a road map, an attribute hierarchy consists of at least one path of attributes. A path is defined to be a subset of attributes connected by one-way arrows. The complex attribute hierarchy below has more than one paths, for example, the path $\alpha_1 \rightarrow \alpha_4 \rightarrow \alpha_7 \rightarrow \alpha_8 \rightarrow \alpha_9 \rightarrow \alpha_{10}$. For any hierarchy of K attributes, the longest path involves at most K attributes and has at most $K - 1$ arrows. The maximum is reached when the K attributes form a linear hierarchy.

Note that some attributes appear in the same path while others do not share a common path. For example, α_1 and α_2 in the following hierarchy do not share a common path. Another example is the pair of α_{10} and α_{11}.

The prerequisite relationships between attributes are quite common in content standards for mathematics. As shown in the map of College- and Career-Ready Standards (CCRS - formerly called the Common Core State Standards), content standards do not stand alone but form a complicated network (Zimba, 2011, 2015). Some standards form a linear structure with one standard being the prerequisite of another one (Figure 2a). Some standards serve as prerequisites for several other standards (Figure 2b). There are also standards that are based on several other standards (Figure 2c).

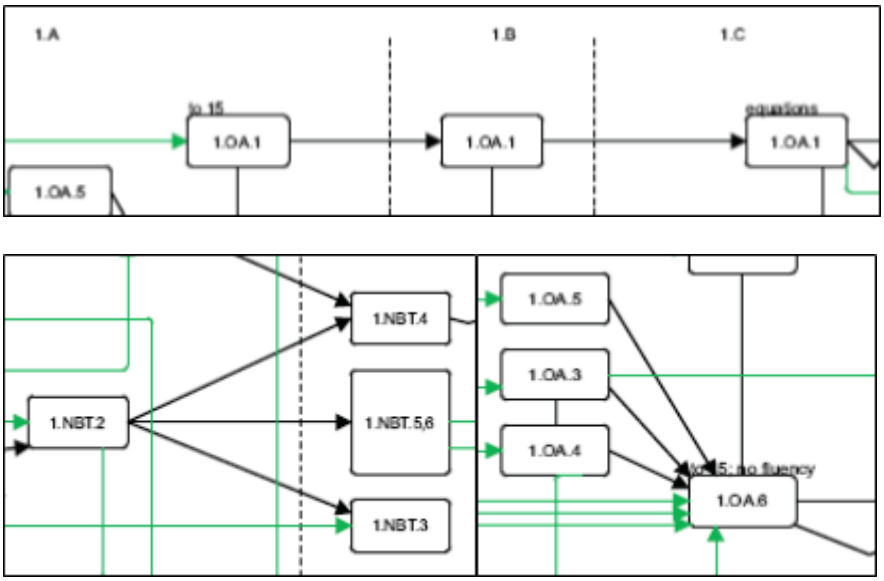


Figure 2: Three types of standard relationships in the Common Core Graph (a: the upper panel, b: left bottom panel, c: right bottom panel)

However, attribute hierarchies have long been poorly represented in the current CD literature, and related studies have begun only recently (e.g., Templin & Bradshaw, 2014). Research on hierarchical attributes has focused on hypothesis testing of the assumed attribute hierarchy (Templin & Bradshaw, 2014) and model estimation (Tu et al., 2018). When attribute hierarchies are proved to be present, it is recommended to incorporate this information in the

modeling process by reparameterizing the original model and excluding certain attribute profiles (Templin & Bradshaw, 2014; Tu et al., 2018).

Hierarchies that have been used in simulation studies are summarized below. Leighton et al. (2004) proposed four types of attribute hierarchies, which have been adopted in many studies—linear, divergent, convergent, and unstructured hierarchies—as illustrated in Figure 3. Liu and Huggins-Manley (2016) renamed the unstructured hierarchy and the convergent hierarchy in Leighton et al. (2004) as the "invert pyramid" and the "diamond hierarchy," respectively. They replaced the divergent hierarchy with the pyramid hierarchy (Figure 4). Tu et al. (2018) added a mixed type to the list, which is a combination of two hierarchies (Figure 5).

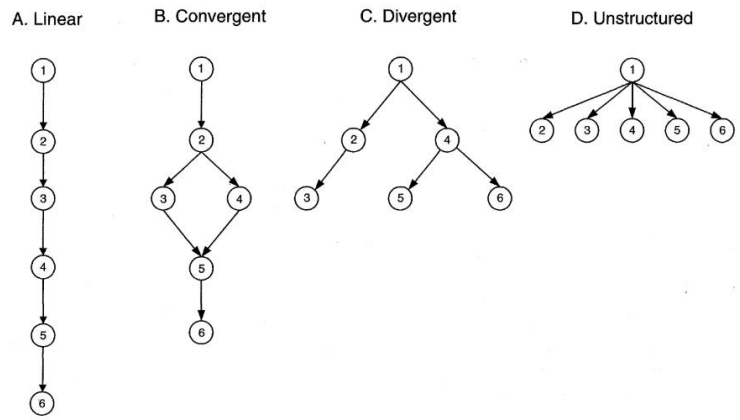


Figure 3: Four hierarchical structures using six attributes (Leighton, Gierl, & Hunka, 2004)

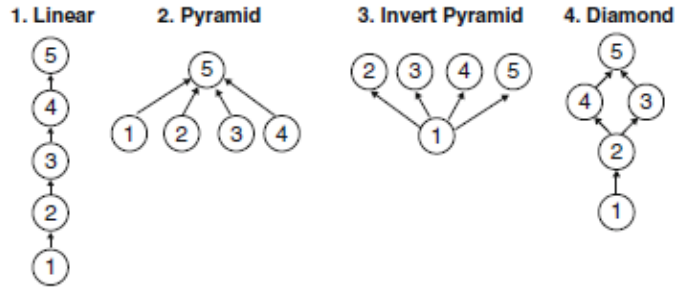


Figure 4: Linear, pyramid, inverted pyramid and diamond structures using five attributes (Liu & Huggins-Manley, 2016)

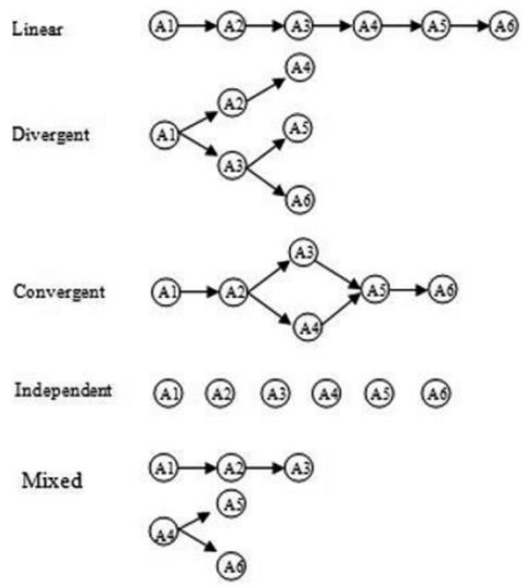


Figure 5: Four types of attribute hierarchies and an independent structure (Tu, Wang, Cai, Douglas, & Chang, 2018)

Note that a pyramid (e.g., Liu & Huggins-Manley, 2016) or a convergent (e.g., Tu, et al., 2018) hierarchy comes with an implicit assumption that all prerequisite attributes must be mastered so that the mastery of the higher-level attribute can be possible.

In application studies of CDMs with hierarchical attributes, the most commonly seen hierarchy is the linear hierarchy (Gierl, Wang, & Zhou, 2008; Gierl, Alves & Majeau, 2010). To get an idea of the hierarchical relationships in real classroom instruction, two CCRS-aligned textbooks for Grade 4 math, Eureka Math (2015) and Engaged NY (2014), were analyzed. The content structures of the textbooks may shed some light on classroom instruction because textbooks provide an essential source of information and guidance for teachers, especially when new standards are introduced. The content analysis results can be found in Appendix A. Generally, three to five attributes (standards) are involved in a period of one to four weeks. Pyramid and invert pyramid structures following the definitions of Liu and Huggins-Manley (2016) are observed besides the linear structure.

2.1.2 Attribute profile space of hierarchical attributes

For a test involving *K* attributes, the set of all possible attribute profiles, subject to the relationship between attributes, is called the attribute profile space (also called latent attribute space or latent space; e.g., Köhn & Chiu, 2018; Tatsuoka, 2009). The attribute profile space, denoted by $\mathcal{L}$, is defined by a matrix with *K* columns representing *K* attributes and each row vector representing an attribute profile.

Identifying the attribute profile space for *K* independent attributes is straightforward. Assuming *K* independent attributes, the attribute profile space $\mathcal{L}$ is a 2^{K}-by-K matrix, representing 2^{K} different classes into which the examinees would be classified.

The hierarchical relationships between attributes constrain the latent attribute space because some attribute profiles become impossible. Specifically, it is not allowed to master an attribute without mastering its prerequisite. Researchers have reached a consensus on restricting the attribute profile space at the presence of hierarchical attributes (e.g., Templin & Bradshaw, 2014; Tu et al., 2018). However, the identification of the attribute profile space is not straightforward, especially when the number of attributes is large (Köhn & Chiu, 2018).

Köhn and Chiu (2018) proposed the lattice-theoretical approach to obtain the latent space. The first step is to derive the *K* basic proficiency classes "by inspection" from the tree graph of the attribute hierarchy. Each basic proficiency class is a K-element vector characterizing a possible path from the lowest-level attribute to a higher-level attribute. The next step is to reconstruct the attribute space as a set of linear combinations of the basic proficiency classes. However, the inspection becomes more difficult as the number of attributes increases and the process is prone to mistakes.

An alternative way to derive the attribute profile space begins with the A-matrix. The first step is to derive the basic proficiency classes as defined in Köhn and Chiu (2018) in the form of column vectors of a matrix, called the reachability matrix (R-matrix; Tatsuoka, 1983, 2009; Gierl et al., 2000). This approach is, therefore, referred to as the R-matrix approach.

2.1.2.1 R-matrix approach

We define some Boolean operations before elaborating the R-matrix approach. A Boolean vector or matrix is one for which all entries are either 0 or 1. The Boolean addition of two Boolean vectors of K elements is defined as

$$\boldsymbol{x}_1 + \boldsymbol{y}_2 = (r_{11} \vee r_{12}, \ldots, r_{K1} \vee r_{K2}), \tag{1}$$

where $\vee$ is the Boolean "or" operator.

The product of the I-by-K Boolean matrix A and the K-by-J Boolean matrix B is defined by a matrix C, the $[i, j]$th element of which is

$$C[i, j] = \bigvee_k A[i, k] \wedge B[k, j], \tag{2}$$

where $\vee$ is the Boolean "or" operator and $\wedge$ is the Boolean "and" operator.

For a square Boolean matrix B, and any $n \geq 0$, the nth Boolean power of B is the Boolean product of n copies of B.

$$B^n \equiv \underbrace{B \odot B \odot \cdots \odot B}_{n \text{ times}} \tag{3}$$

The derivation of the R-matrix from the A-matrix and the derivation of the attribute profile space from the R-matrix are elaborated below.

The R matrix can be calculated as the nth Boolean power of the matrix $A + I$ (Leighton et al., 2004):

$$\mathrm{R} = (\mathrm{A} + \mathrm{I})^{\mathrm{n}}, \quad (4)$$

where n is the integer required for R to reach invariance and can represent the numbers 1 through $\mathrm{K} - 1$. The number n is decided by the number of arrows in the longest path of the hierarchy.

The next step of the R-matrix approach derives the attribute profile space from the R-matrix. Note that the A-matrix and R-matrix are of order (k, k). The attribute profile space $\mathcal{L}$, with k columns indicating different attributes, however, may have more than k rows. The following algorithm produces the transpose of the attribute profile space (Ding, Luo, Cai, Lin, & Wang, 2008).

1) For the ith column of the R-matrix, we take the Boolean addition of the ith column and each column on its right side.
2) When a new column vector is obtained, it is added to the right of the R-matrix.
3) The first two steps are repeated for each column of the original R-matrix, including the last one. Note that the column vectors in the Boolean addition include the new columns.

The obtained matrix is called the expanded R-matrix, denoted as R^*, because it expands the K-by-K R-matrix by adding columns. This algorithm is referred to as the expanding algorithm. The attribute profile space $\mathcal{L}$ is the transpose of the expanded R-matrix (R^{*T}) with an additional row of 0s. The space contains at most 2^{K} rows, representing 2^{K} attribute profiles, denoted as $\boldsymbol{\alpha}$s. The maximum is reached when the attributes are independent. The number of attribute profiles ($\boldsymbol{\alpha}$s) in the space decreases with hierarchical attributes.

The R-matrix approach is equivalent to the lattice-theoretical approach (Köhn & Chiu, 2018), but is easier to apply in practice. Appendix B provides R code for the expanding algorithm.

2.1.2.2 Interpretations of the Boolean operations

The interpretations of the Boolean operations involved in the R-matrix approach are provided below.

Note that the A-matrix only captures the direct relationship between two attributes. Each 1-entry in the A-matrix stands for a one-way arrow that connects two attributes. The R-matrix should also capture indirect relationships. Therefore, the first step is to add the identity matrix to the A-matrix to account for the relationship with an attribute itself. The next step multiplies $A + I$ to itself until invariance is achieved. The $[i, j]$th element of $(A + I)^2$ is

$$(A + I)^2[i, j] = \bigvee_k (A + I)[i, k] \wedge (A + I)[k, j], \tag{5}$$

in which $(A + I)[i, k] \wedge (A + I)[k, j] = 1$ if $i \to k$ and $k \to j$, which means attribute i and attribute j has an indirect relationship through attribute k; else, $(A + I)[i, k] \wedge (A + I)[k, j] = 0$. The disjunction among k attribute $\bigvee_k A^{(1)}[i, k] \wedge A^{(1)}[k, j]$ takes the value of 1 if attribute i and attribute j has an indirect relationship through any attribute.

Consequently, the elements in $(A + I)^2$ capture all indirect relationships between attribute i and attribute j in the form of $i \to k \to j$. Similarly, it can be shown that the $[i, j]$th element of the matrix $(A + I)^3$ takes the value of 1 if attribute i and attribute j has an indirect relationship through two attributes in the form of $i \to m \to n \to j$. Since the longest possible path in an attribute hierarchy has $K - 1$ arrows, the largest number n would take in equation (4) is $K - 1$.

Take the jth column of the R-matrix. The ith element of the jth column takes the value of 1 if there a path from attribute i to attribute j. If the jth attribute is at the lowest level in any path, then the jth column has only one non-zero entry; otherwise, the jth column describes a path which

ends at attribute j. As a result, the columns in the R-matrix correspond to different paths as shown in the tree graph, equivalent to the basic proficiency classes defined in Köhn and Chiu (2018).

We use a linear hierarchy with four attributes to demonstrate the derivation of the R-matrix.

$$\mathrm{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{6}$$

$$\mathrm{A} + \mathrm{I} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{7}$$

$$(\mathrm{A} + \mathrm{I})^2 = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{8}$$

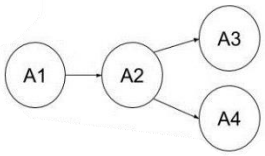
$$R = (\mathrm{A} + \mathrm{I})^3 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{9}$$

The four columns in the R-matrix in equation (9) describe four paths that start from attribute 1 (i.e., the lowest-level attribute) and end with each attribute, respectively. Invariance is achieved at $n = 3$ because the longest path (i.e., $\alpha_1 \rightarrow \alpha_2 \rightarrow \alpha_3 \rightarrow \alpha_4$) has three arrows.

The columns of the R-matrix can be seen as attribute mastery profiles. If the K attributes form a single linear hierarchy, then the R-matrix contains all the possible attribute mastery profiles. However, if there exist two attributes that do not appear in the same path, the R-matrix fails to account for all the possible combinations of states of two such attributes.

Consider the following attribute hierarchy. The first path (column) is nested within the other three paths (columns). The second path is nested within the two paths on the right. However, the last two paths are not nested within each other because A3 and A4 are not connected directly or indirectly in any path. The four columns in the R-matrix also correspond to four profiles.

Another possible profile $[1\ 1\ 1\ 1]^T$, which is not included in the R-matrix, can be obtained by adding the last two columns of the R-matrix.



$$R = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{10}$$

The expanding algorithm involves the Boolean addition of two columns $\boldsymbol{r}_{.i}, \boldsymbol{r}_{.j}$in the R-matrix shown in equation (11) and (12).

$$R = \begin{bmatrix} 1 & r_{1,2} & \cdots & r_{1,K} \\ r_{2,1} & 1 & \cdots & r_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ r_{K,1} & r_{K,2} & \cdots & 1 \end{bmatrix} \tag{11}$$

$$\boldsymbol{r}_{.i} + \boldsymbol{r}_{.j} = (r_{1i} \vee r_{1j}, \ldots, r_{Ki} \vee r_{Kj}) \tag{12}$$

Addition of two nested paths as defined in equation (12) does not produce a new column. Addition of two independent paths, however, produces a new column, which expands the original R-matrix.

Continuing with the complex hierarchy example in Köhn and Chiu (2018), the attribute profile space $\mathcal{L}$ derived from the expanding algorithm contains 31 attribute profiles.

2.1.3 Q-matrix

The relationship between the items and the attributes is described in an indicator matrix, called the Q matrix, which has rows corresponding to items, columns corresponding to attributes, and binary elements indicating whether an attribute is measured by an item (that is, whether mastery of an attribute is required to succeed on an item). The Q-matrix was initially proposed by Tatsuoka (1983) and has been employed in most of the commonly used CDMs.

The Q-matrix reflects the test blueprint (Leighton, Gierl, & Hunka, 2004). Specifically, the Q-matrix operationalizes the substantive and cognitive theories based on which the test has been

developed and provides evidence for the construct and content aspects of validity (Rupp, Templin, & Henson, 2010). It is often considered an analog to the specified factor structure in a confirmatory factor analysis (Henson, DiBello, & Stout, 2018). The row vectors of the Q-matrix are also referred to as q-vectors. Items with a q-vector with only one non-zero entry are called single-attribute items. Others are multiple-attribute items.

An example of Q-matrix is

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \tag{13}$$

which shows that the test measures three attributes with three items, the first item probes the second attribute, the second item targets the first and the third attributes, and the last item requires all three attributes. In other words, an examinee needs to master the second attribute to succeed on item 1 without guessing or slipping.

The specification of the Q-matrix precedes any model fitting and classifying. The Q-matrix is part of the model assumption that can be falsified (e.g., Wang et al., 2018). While most theoretical and empirical studies assume that the Q-matrix is correctly specified (e.g., Henson et al., 2018), recent efforts on Q-matrix construction and validation have pointed out the negative effects of incorrectly identified Q-matrices and proposed solutions (e.g., de la Torre, 2008; Liu, Xu, & Ying, 2012).

2.1.3.1 Reduced versus full Q-matrix

With hierarchical attributes, researchers have reached a consensus on restricting the attribute profile space (e.g., Templin & Bradshaw, 2014; Tu et al., 2018). However, there has not been a consensus on the Q-matrix. Two types of Q-matrices are being used: the *full (or unrestricted) Q-matrices* (Liu et al., 2016; Templin & Bradshaw, 2014) and *the reduced (or restricted) Q-matrices* (Köhn & Chiu, 2018; Leighton et al., 2004; Tu et al., 2018), which are defined below.

Consider a test with three independent attributes. The expanded R-matrix R^* below has seven columns and each column represents an item type:

$$R^* = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}. \quad (14)$$

If we randomly sample from the columns of R^* in equation (14) as the q-vectors, regardless of the attribute hierarchy, the Q-matrix is called a *full Q-matrix*. With any attribute hierarchy, a full Q-matrix could have all seven types of q-vectors or a random subset of them. In a test of three linear attributes, for instance, although the attribute profile $\boldsymbol{\alpha} = (1\ 0\ 1)$ is not allowed, the q-vector $\boldsymbol{q} = (1\ 0\ 1)$ is possible in the full-Q-matrix approach.

Considering that some attributes profiles become illegitimate under a certain hierarchy; particularly, it is impossible to master an attribute without mastering all prerequisite attributes. Therefore, in another line of research, it is assumed that an item probing a higher-level attribute also requires its prerequisite. This assumption would lead to the removal of some q-vectors. For example, $q = (0\ 1\ 0)$ under a linear hierarchy ($\alpha 1 \rightarrow \alpha 2 \rightarrow \alpha 3$) would be unreasonable because the item requires the mastery of the second attribute without requiring its prerequisite. A reduced Q-matrix can only have columns of R^* as q-vectors. A special reduced Q-matrix is the transpose of R^*, denoted as Q_r. For three linear attributes, for example, R^* and Q_r are defined in equation (15) and (16).

$$R^* = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad (15)$$

$$Q_r = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}^T \quad (16)$$

The only difference between $\mathrm{Q_r}$ and the attribute profile space $\mathcal{L}$ is the exclusion or inclusion of the vector of all 0s. Therefore, Q_r can also be derived using the R-matrix approach.

While studies using full Q-matrices tend not to discuss the necessity to make any change in the Q-matrix, researchers using reduced Q-matrices believe that the items should reflect the attribute hierarchy (Köhn & Chiu, 2018; Tu et al., 2018). The choice between the full Q-matrix and the restricted one has not been formally addressed in the literature.

2.1.3.2 Complete Q-matrix

A complete Q-matrix is needed to identify all possible attribute profiles (Chiu, Douglas, & Li, 2009; Chiu & Köhn, 2015). With a complete Q-matrix, we have $S(\boldsymbol{\alpha}) = S(\boldsymbol{\alpha}') \Longrightarrow \boldsymbol{\alpha} = \boldsymbol{\alpha}'$, where $S(\boldsymbol{\alpha})$ denotes the expected response vector $\left(E[Y_1|\boldsymbol{\alpha}], E[Y_2|\boldsymbol{\alpha}] \ldots, E[Y_J|\boldsymbol{\alpha}]\right)$. Completeness of the Q-matrix is evaluated by checking the definition $S(\boldsymbol{\alpha}) = S(\boldsymbol{\alpha}') \Longrightarrow \boldsymbol{\alpha} = \boldsymbol{\alpha}'$ for each pair, $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$, in the attribute profile space. It was proved in Chiu et al. (2009) that a Q-matrix containing the identity matrix (i.e., K single-attribute items) is complete for the DINA model with independent attributes. Köhn and Chiu (2018) later showed that any Q-matrix that contains the transpose of the R-matrix is complete for the DINA model, given any attribute hierarchy. This rule, however, does not apply to more complicated CDMs such as ACDM and GDINA (Köhn & Chiu, 2018).

2.1.4 Item response models and calibration methods

The relationship between each attribute profile and the probability of a correct response is expressed in terms of IRF (de la Torre, 2011; Rupp, Templin, & Henson, 2010). A variety of models with different IRFs for multiple-attribute items have been proposed; most of them are equivalent to each other in the parameterization for a single-attribute item.

Some CDMs are more general models that subsume most other specific models. The general frameworks include the general diagnostic model (GDM; von Davier 2005), the log-linear

cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009), and the generalized DINA (Deterministic Input, Noisy "and" Gate) model (GDINA; de la Torre, 2011).

The rest of the section introduces the GDINA framework and two reduced models from GDINA. The following notations are used:

- K_j^* is the number of required attributes for item j, as in $K_j^* = \sum_{k=1}^{K} q_{ik}$.
- $\boldsymbol{\alpha}_{lj}^*$ is the reduced attribute vector consisting of the columns of the required attributes, where $l = 1, \ldots, 2^{K_j^*}$.
- The probability of a correct response on item *j* by students with attribute pattern $\boldsymbol{\alpha}_{lj}^*$ will be denoted by $P(X_j = 1 | \boldsymbol{\alpha}_{lj}^*) = P(\boldsymbol{\alpha}_{lj}^*)$.

The IRF of the GDINA model (de la Torre, 2011) is given by

$$g[P(\boldsymbol{\alpha}_{lj}^*)] = \phi_{j0} + \sum_{k=1}^{K_j^*} \phi_{jk}\alpha_{lk} + \sum_{k=1}^{K_j^*}\sum_{k=1}^{K_j^*-1} \phi_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \phi_{j12\ldots K_j^*}\prod_{k=1}^{K_j^*}\alpha_{lk} \qquad (17)$$

where $g[P(\boldsymbol{\alpha}_{lj}^*)]$ is $P(\boldsymbol{\alpha}_{lj}^*)$, $\log[P(\boldsymbol{\alpha}_{lj}^*)]$, and $\text{logit}(P(\boldsymbol{\alpha}_{lj}^*))$ in the identity, log, and logit links, respectively; ϕ_{j0} is the intercept for item j; ϕ_{jk} is the main effect due to α_{lk}; $\phi_{jkk'}$ is the interaction effect due to α_{lk} and $\alpha_{lk'}$; $\phi_{j12\ldots K_j^*}$ is the interaction effect due to α_{l1}, …, α_{lK^*}.

The G-DINA model is a saturated model and subsumes several widely used reduced CDMs, including the DINA model (Haertel 1989; Junker and Sjitsma 2001; Macready and Dayton 1977) and the A-CDM (de la Torre, 2011).

To obtain the DINA model, all terms in the GDINA model in identity link, except ϕ_{j0} and $\phi_{j12\ldots K_j^*}$, are constrained to zero, that is,

$$P(\boldsymbol{\alpha}_{lj}^{*}) = \phi_{j0} + \phi_{j12\ldots K_j^{*}} \prod_{k=1}^{K_j^{*}} \alpha_{lk} \tag{18}$$

The A-CDM is the constrained identity-link G-DINA model without the interaction terms. It can be formulated as

$$P(\boldsymbol{\alpha}_{lj}^{*}) = \phi_{j0} + \sum_{k=1}^{K_j^{*}} \phi_{jk}\alpha_{lk}. \tag{19}$$

Current methods for fitting CDMs use either marginal maximum likelihood estimation that relies on the Expectation Maximization algorithm (MMLE-EM) or Markov chain Monte Carlo (MCMC) techniques (Rupp et al., 2010).

2.1.5 Classification methods

The prime objective of CDM data analysis is to classify examinees into one of the attribute profiles. The estimated attribute profile denoted as $\widehat{\boldsymbol{\alpha}}$, takes the value of one of the possible skill patterns $\boldsymbol{\alpha}_l$ for $l = 1, \ldots, L$. When K dichotomous attributes are involved and assumed to be independent, the attribute profile space consists of $L = 2^K$ latent classes. If an attribute hierarchy exists, the number of attribute profiles L decreases with some attribute profiles becoming impossible.

Examinees are often classified via maximum likelihood estimation (MLE; de la Torre, 2008), maximum a posteriori (MAP; Rupp et al., 2010), or expected a posteriori (EAP; de la Torre, 2008; Rupp et al., 2010), which are applicable to any CDM that is a special case of a restricted latent class model. Huebner and Wang (2011) conducted a simulation study comparing the accuracy of the three methods under different testing conditions.

The likelihood function of the responses given the attribute profile $\boldsymbol{\alpha}$ is given by

$$L(\boldsymbol{X}_i|\boldsymbol{\alpha}) = \prod_{j=1}^{J} \mathrm{P}\left(\mathrm{X}_{ij} = 1\middle|\boldsymbol{\alpha}\right)^{X_{ij}} \left[1 - \mathrm{P}\left(\mathrm{X}_{ij} = 1\middle|\boldsymbol{\alpha}\right)\right]^{1-X_{ij}}. \tag{20}$$

The MLE estimator is the attribute profile $\boldsymbol{\alpha}_l$ for $l = 1, \ldots, L$ that maximizes the likelihood, and is formally denoted as

$$\hat{\boldsymbol{\alpha}}_{MLE} = \arg\max_{l} L(\boldsymbol{X}_i|\boldsymbol{\alpha}). \tag{21}$$

If prior probabilities denoted as $P(\boldsymbol{\alpha}_l)$ for $l = 1, \ldots, L$, are available from previous test administrations, the posterior probability $P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i)$ for each $\boldsymbol{\alpha}_l$ can be calculated:

$$P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i) = \frac{L(\boldsymbol{X}_i|\boldsymbol{\alpha}_l)P(\boldsymbol{\alpha}_l)}{\sum_{m=1}^{L} L(\boldsymbol{X}_i|\boldsymbol{\alpha}_m)P(\boldsymbol{\alpha}_m)}. \tag{22}$$

The MAP estimator is then denoted as

$$\hat{\boldsymbol{\alpha}}_{MAP} = \arg\max_{l} P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i). \tag{23}$$

It is generally true that MLE and MAP estimates are equivalent if flat priors are used in MAP estimation (Huebner & Wang, 2011).

For the EAP approach, the probabilities of mastery for each attribute (the marginal skill probabilities), $\tilde{\alpha}_k$ for $k = 1, \ldots K$, are calculated for an examinee and rounded at .50 to obtain binary mastery classifications. The posterior probabilities $P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i)$ are aggregated to obtain the marginal probabilities $\tilde{\alpha}_k$ for $k = 1, \ldots K$:

$$\tilde{\alpha}_k = \sum_{l=1}^{L} P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i) I(\alpha_{l,k} = 1) \tag{24}$$

where $I\left(\alpha_{l,k} = 1\right) = \begin{cases} 1 & \text{if element k of Attribute Profile l equals 1,} \\ 0 & \text{otherwise.} \end{cases}$

The marginal probability $\tilde{\alpha}_k$ is usually rounded at .50 to obtain a binary classification for attribute k ($k = 1, \ldots K$).

With hierarchical attributes, researchers have reached a consensus on restricting the attribute profile space (e.g., Templin & Bradshaw, 2014; Tu et al., 2018). The MLE estimator maximizes the likelihood function over the set of all possible attribute profiles when the item parameters are assumed to be known, which is referred to as unrestricted MLE (Tu et al., 2018). When hierarchical attributes are involved, a restricted MLE is recommended in which the probability of some attribute profiles are fixed to zero due to the hierarchy (Templin & Bradshaw, 2014; Tu et al., 2018). The only difference between unrestricted and restricted MLE is in the attribute profile space. Similarly, restricted MAP and EAP estimators should be used for hierarchical attributes.

2.1.6 Q-matrix design

The CDMs provide guidance for test construction. Cognitive theories could have a real impact on testing practice through CDM model assumptions about relationships between attribute as well as the relationship between attributes and item responses. Given a set of attributes, instead of relying heavily on post hoc item analysis surrounding internal consistency, test development in the CDM context begins with a set of possible item types that are characterized by their q-vectors. For example, a test with three independent attributes can have at most seven different item types. The Q-matrix for a particular test can be obtained by sampling with replacement from the column vectors of the corresponding R^*. The Q-matrix is a core element of the CDM-based test design.

Madison and Bradshaw (2015) defined the Q-matrix design as "the deliberate arrangement of a set of test items according to the specific subset of attributes measured by each individual item." The Q-matrix plays a significant role in the statistical identification of the model (Köhn & Chiu, 2018; Xu & Zhang, 2016). However, Q-matrices that lead to identification may provide varying classification accuracy rates.

Three studies have been done with the effects of Q-matrix design on classification accuracy with independent attributes. Chiu, Douglas, and Li (2009) showed that each attribute needs to be measured by at least one single-structured item in order to obtain acceptable classification accuracy in both DINA (Haertel, 1989; Junker & Sjitsma, 2001; Macready & Dayton, 1977) and DINO (Templin & Henson, 2006) models. Similarly, DeCarlo (2011), in his investigation of the DINA model, found that if an attribute is always measured through interaction terms and never measured in isolation, the classification obtained only reflects the prior probabilities. The finding of DeCarlo (2011) was echoed in Madison and Bradshaw (2015), in which they concluded that attributes measured in isolation could help increase classification accuracy when holding constant the number of times an attribute is measured on a test, based on the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009).

Recent efforts expanded the research on Q-matrix design to testing situations with hierarchical attributes (Liu & Huggins-Manley, 2016; Liu, Huggins-Manley, & Bradshaw, 2017). In Liu, Huggins-Manley, and Bradshaw (2017), different Q-matrix designs were generated using the so-called independent approach, adjacent approach, or reachable approach when the attribute hierarchy was linear, divergent, convergent, or unstructured. The CDM was the hierarchical diagnostic classification model (HDCM; Templin & Bradshaw, 2014). The independent approach only allows for simple-structured items. Each item measures at most two attributes with direct relationships in the adjacent approach. Each item can measure any combination of attributes that are directly or indirectly connected in the reachable approach. Their simulations found that the adjacent approach leads to higher classification accuracy in a shorter test and they recommended using the adjacent approach to design the Q-matrix when a hierarchy is present (Liu et al., 2017). Using the adjacent approach in Liu et al. (2017), Liu and Huggins-Manley (2016) found that

"higher-level attributes were often associated with higher classification accuracy than lower-level attributes" as a result of more information about higher-level attributes from the hierarchical structure.

2.1.7 Criteria for test construction

A research area closely related to Q-matrix design is the development of item and test indices. When estimated item parameters are available for a pool of items, an item index based on the estimated item parameters can be calculated to identify good items that achieve high classification rates with a minimal number of items (Henson, DiBello, & Stout, 2018). This type of item indices is referred to as item discrimination in Henson et al. (2018). The Fisher information is an example of such item indices in the IRT context. For CDMs, a counterpart of the Fisher information is the Kullback-Leibler information (KLI; also called KL divergence or KL distance). Much of the work on item-level and test-level indices in CDMs have been based on KLI.

2.1.7.1 Kullback-Leibler information

KLI measures how far a distribution q is away from the actual distribution p (Gray, 2011; Chang & Ying, 1996; Xu, Chang, & Douglas, 2003). Given a probability space $(\Omega, \mathrm{B}, \mathrm{P})$, with Ω being a finite space, and another measure M on the same space, the KL information of P with respect to M (Gray, 2011) is defined as

$$D(P, M) = \sum_{\omega \in \Omega} P(\omega) \ln \frac{P(\omega)}{M(\omega)}, \tag{25}$$

which ranges from 0 to ∞.

The Fisher information can be used in the test construction because the test information is the sum of item information, and the variability of the maximum likelihood estimate decreases as the information increases. Test construction criteria for CDMs should have similar properties (Henson & Douglas, 2005).

The KL information for an item j for differentiating $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$ is defined as

$$D_{juv} = \sum_{x=0}^{1} P(x|\boldsymbol{\alpha}_u) \ln \frac{P(x|\boldsymbol{\alpha}_u)}{P(x|\boldsymbol{\alpha}_v)}. \quad (26)$$

Note that $D_{juv} \neq D_{jvu}$; $D_{juv} = 0$ for $u = v$. An item is most useful in determining the difference between two attribute profiles, $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$, if D_{juv} and D_{jvu} are large. All D_{juv}s for item j can be recorded in a matrix D_j of L columns and L rows where L is the size of the attribute profile space.

The KL information for a test is defined as

$$D_{.uv} = \sum_{\boldsymbol{X}} P(\boldsymbol{X}|\boldsymbol{\alpha}_u) \ln \frac{P(\boldsymbol{X}|\boldsymbol{\alpha}_u)}{P(\boldsymbol{X}|\boldsymbol{\alpha}_v)}. \quad (27)$$

where $\boldsymbol{X}$ represents the response pattern for J items. The KL information for a test compares the probability distribution for an item response vector $\mathbf{X}$, given $\boldsymbol{\alpha}_u$ when compared to the probability distribution of $\mathbf{X}$ given an alternative attribute pattern, $\boldsymbol{\alpha}_v$. Because of the assumption of local independence among items conditional on $\boldsymbol{\alpha}$, it can be shown that the test information is the sum of the KL information across all items in the exam. The test KL information $D_{.uv}$ for all pairs of (u, v) in the attribute profile space, $\mathcal{L}$, forms an $L \times L$ matrix D where L is the size of $\mathcal{L}$.

of a $2^K \times 2^K$ matrix containing $2^K(2^K - 1)$ possible comparisons because the KL information is not symmetric. The diagonal elements of the matrix are zero. The KL information provides a general method that will apply to all CDMs (Henson & Douglas, 2005), based on which researchers have proposed attribute, item, or test-level indexes for test construction.

2.1.7.2 Cognitive diagnostic index (Henson & Douglas, 2005)

The cognitive diagnostic index (CDI) for an item j is proposed as a weighted average of the off-diagonal elements of D_j since the matrix expands exponentially with the number of

attributes K and makes it difficult in simultaneously evaluating all the elements (Henson & Douglas, 2005). The CDI_j for item j is defined as

$$CDI_j = \frac{1}{\sum_{u \neq v} h(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)^{-1}} \sum_{u \neq v} h(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)^{-1} D_{juv}. \quad (28)$$

where $h(.,.)$ is the Hamming distance and D_{juv} stands for the element of the matrix D_j at (u, v).

The CDI for a test is defined as

$$CDI = \frac{1}{\sum_{u \neq v} h(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)^{-1}} \sum_{u \neq v} h(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)^{-1} D_{uv}. \quad (29)$$

where D_{uv} stands for the element of the matrix $D(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)$ at (u, v). It can be shown that the CDI for a test is the sum of CDI_j for all the items in the test.

Henson and Douglas (2005) showed that the CDI strongly relates to the average correct classification rates across attributes and examinees for a test and they suggest using the cognitive diagnostic index (CDI) as a measure of an item's or test's discrimination power.

Other indexes based on the KL information include the Attribute Discrimination Index (ADI) that is supposed to be related to the correct classification rate of the masters for the *k*th attribute (Henson, Roussos, Douglas & He, 2008), and the modified CDI and modified ADI (Kuo, Pai, & de la Torre, 2016). Note that all the indexes mentioned above are overall indexes that are not conditional on $\boldsymbol{\alpha}$.

2.1.7.3 A unified item and test discrimination approach (Henson, DiBello, & Stout, 2018)

Henson et al. (2018) proposed a probability-based attribute-specific index for items with multiple options. For dichotomous items, the index is reduced to

$$\mathrm{DI}_{jk} = \max_{\boldsymbol{\alpha}} \left(\left| P(X_j = 1 | \boldsymbol{\alpha}) - P(X_j = 1 | \boldsymbol{\alpha}^{-k}) \right| \right), \quad (30)$$

where $\boldsymbol{\alpha}^{-k}$ denotes an attribute pattern that differs from $\boldsymbol{\alpha}$ only on the kth attribute. The maximization is taken over all $\boldsymbol{\alpha}$s. The index DI_{jk} describes the discrimination power of item j in measuring attribute k and has a value between 0 and 1.

2.2 Nonparametric classification based on CDM conception

An alternative to classification when calibrating a parametric CDM is not practical or even possible is the nonparametric approach. The nonparametric approach shares with the conventional CDM approach the conceptions of a Q-matrix, a set of attributes, and different attribute interaction effects on correct responses. The test is still constructed based on a CDM, but a probabilistic model is not used to characterize the correct response probabilities of different attribute profiles. Instead, the examinees are classified into different attribute profiles using a nonparametric method.

Barnes (2010) developed a nonparametric exploratory approach to build the Q-matrix and classify examinees. Some researchers employ cluster analysis for nonparametric classifications (Ayers, Nugent, & Dean, 2008; Chiu, Douglas, & Li, 2009; Willse, Henson, & Templin, 2007). Another stream of research is based on the idea of minimizing the distance between observed item response patterns and the ideal response patterns according to the Q-matrix (Chiu & Douglas, 2013; Chiu, Sun, & Bian, 2018; Wang & Douglas, 2015). The rest of the section reviews the third type of nonparametric methods that minimize distance measures.

2.2.1 The nonparametric (NPC) method

Chiu and Douglas (2013) proposed a simple method to classify examinees by matching observed item response patterns to the nearest ideal response pattern, henceforth referred to as the nonparametric (NPC) method. The ideal response of examinee i on item j is denoted as η_{ij}, and the vector containing ideal responses of examinee i on all the items in a test is denoted as $\boldsymbol{\eta}_i$.

The ideal response patterns are derived from the Q-matrix and the assumption on attribute interactions. Consider a q-vector $\mathbf{q}_* = (1 \quad 1)$ and four possible attribute profiles $\boldsymbol{\alpha}_1 = (0 \quad 0), \boldsymbol{\alpha}_2 = (1 \quad 0), \boldsymbol{\alpha}_3 = (0 \quad 1),$ and $\boldsymbol{\alpha}_4 = (1 \quad 1)$. If we assume a conjunctive model underlying the responses, the ideal responses for the four attribute profiles would be $\eta_{1*} = 0, \eta_{2*} = 0, \eta_{3*} = 0,$ and $\eta_{4*} = 1,$ respectively. For a test with more than one item, each possible attribute profile is associated with an ideal response pattern. The observed response pattern of an examinee is compared with the ideal response patterns. The attribute profile of the closest ideal response pattern is the estimate for the examinee. Three distance measures were proposed by Chiu and Douglas (2013). Denote the observe response pattern as $\boldsymbol{x}$. The hamming distance between $\boldsymbol{x}$ and $\boldsymbol{\eta}$ is given by

$$d_h(\boldsymbol{x}, \boldsymbol{\eta}) = \sum_{j=1}^{J} |x_j - \eta_j|, \tag{31}$$

where *J* stands for the test length. A weighted Hamming distance is defined as

$$d_{wh}(\boldsymbol{x}, \boldsymbol{\eta}) = \sum_{j=1}^{J} \frac{1}{\bar{p}(1-\bar{p})} |x_j - \eta_j|, \tag{32}$$

where $\bar{p}$ denotes the proportion correct on the j th item. They also proposed the penalized Hamming distance for the special cases where the slipping parameter is much less than the guessing parameter or vice versa (Chiu & Douglas, 2013).

Chiu and Douglas (2013) found that accurate classification can be achieved when the true model is DINA and NIDA with slip and guess parameters considerably higher than 0. The estimator of $\boldsymbol{\alpha}$ would be perfect without any slipping or guessing but still performs with good relative efficiency even when this is not the case (Chiu & Douglas, 2013). A formal justification

for the NPC methods was provided in Wang and Douglas (2015), showing that the nonparametric method yields consistent classifications under a variety of underlying conjunctive models.

2.2.2 The general nonparametric classification (GNPC) method

The general nonparametric classification (GNPC) method (Chiu, Sun, & Bian, 2018) was proposed as an extension of the NPC methods (Chiu & Douglas, 2013). The example in 3.2.1 is revisited to illustrate the need for this extension. The ideal responses for the four attribute profiles are $\eta_{1*} = 0, \eta_{2*} = 0, \eta_{3*} = 0,$ and $\eta_{4*} = 1$, respectively, assuming an underlying conjunctive model. The ideal responses would become $\eta_{1*} = 0, \eta_{2*} = 1, \eta_{3*} = 1,$ and $\eta_{4*} = 1$ if the underlying model is a disjunctive one. In the NPC method, either the conjunctive ideal response patterns (denoted as $\boldsymbol{\eta}^{(c)}$) or the disjunctive ideal response patterns (denoted as $\boldsymbol{\eta}^{(d)}$) are used according to the assumptions about the cognitive process. However, using $\boldsymbol{\eta}^{(c)}$ or $\boldsymbol{\eta}^{(d)}$ may not be adequate if the underlying CDM is a complex one, such as a saturated GDINA model. Consider a set of GDINA parameters for this item $(\beta_0, \beta_1, \beta_2, \beta_3) = (0.1, 0.4, 0.6, -0.2)$. The probabilities for the four possible attribute profiles to answer the item correctly are $(0.1, 0.5, 0.7, 0.9)$. Obviously, neither the ideal responses (0, 0, 0, 1) nor (0, 1, 1, 1) would be appropriate.

Besides, before any analysis of the response data, we cannot decide which of the ideal response patterns is more suitable. Therefore, the GNPC method defines the weighted ideal response on item j for the lth attribute profile in the attribute profile space as

$$\eta_{lj}^{(w)} = w_{lj}\eta_{lj}^{(c)} + (1 - \mathrm{w}_{lj})\eta_{lj}^{(d)}, \tag{33}$$

in which $0 \leq w_{lj} \leq 1$ is a weight calculated from the data in an iterative procedure. Conceptually, the weight is found when the total distance between the observed responses and the weighted ideal responses is minimized. Denote the attribute profiles as C_l for $l = 1, \ldots L$. The total distance can be denoted as

$$\mathrm{d}_{lj} = \sum_{i \in C_l} \left(x_{ij} - \eta_{lj}^{(w)} \right)^2. \tag{34}$$

$\widehat{w}_{lj}$ is obtained by minimizing d_{lj}:

$$\widehat{w}_{lj} = \frac{\sum_{i \in C_l} \left(x_{ij} - \eta_{lj}^{(d)} \right)}{n_l \left(\eta_{lj}^{(c)} - \eta_{lj}^{(d)} \right)}, \tag{35}$$

where n_l is the number of examinees classified to attribute profile C_l. The $\widehat{w}_{lj}$ can be computed via an iterative procedure described in Chiu et al. (2018). The NPC method can be used to provide a set of initial classifications to calculate the initial $\widehat{w}_{lj}$.

The NPC (Chiu & Douglas, 2013; Wang & Douglas, 2015) and the GNPC (Chiu et al., 2018) methods do not have limitations regarding the number of attributes, the sample size or the test length as the conventional CDMs do, which makes them a practical option for small-scaled classroom assessments.

2.3 CD-CAT

2.3.1 From IRT-based CAT to CD-CAT

Computerized adaptive testing (CAT), built on the idea of "individualized testing," can tailor both items in the test form and the test length to an individual examinee. The maximum information criterion is usually adopted in IRT-based CAT's item selection to optimize test efficiency in terms of shorter test length or higher measurement precision or both compared to linear testing. There have been many operational CAT programs since the 1980s and rich literature in the past decades (Reckase, 2010).

CAT algorithms based on CDMs (denoted as CD-CAT) have been developed with the same motivation behind the IRT-based CAT, that is, to increase testing efficiency (Cheng, 2009; McGlohen & Chang, 2008; Xu, Chang, & Douglas, 2003). When the cognitive diagnosis is

combined with CAT, we can proceed from "individualized testing" to a new stage of "individualized learning." As technologies become more available in the classroom, CD-CAT can play a more important role in learning and teaching. Chang (2015) reported an experimental CD-CAT program was implemented in Zhengzhou, China and a survey suggested that "CD-CAT encourages critical thinking, making students more independent in problem solving, and offers easy to follow individualized remedy, making learning more interesting. (p. 15)"

Similar to the CATs based on other measurement models, a CD-CAT algorithm consists of a measurement model (e.g., the DINA model), a method for selecting the first item(s) to administer, a scoring method, a rule to select the next item conditional on examinee responses to the previous item(s), and a termination rule to end the test. An item pool with calibrated items is needed for the implementation of the CAT algorithm.

2.3.2 Item selection methods for CD-CAT

Item selection is a core element of CAT algorithms. Three item selection indices based on the KL information are reviewed in this section because they will be used in the simulation study. There are item selection methods based on other criteria such as the Shannon entropy (Wang, 2013; Xu et al., 2003) and mutual information (Huebner, Finkelman, & Weissman, 2018).

The following notations are used for the CD-CAT context:

$\widehat{\boldsymbol{\alpha}}_i^{(t)}$ denotes the attribute profile estimate for examinee i after t items have been administered;

$\boldsymbol{x}_i^{(t)}$ denotes the observed response pattern for examinee i when t items have been administered;

L denotes the size of the attribute profile space;

$\boldsymbol{\alpha}_l$ ($l = 1, 2, \ldots, L$) denotes the lth attribute profile in the attribute profile space;

$R^{(t)}$ denotes the available items in the item pool when t items have been administered; and X_{ih} denotes the response of examinee i to item h from $R^{(t)}$.

The KL algorithm. Xu, Chang, and Douglas (2003) proposed using the straight sum of the KL distances between $f(X_{ih}|\widehat{\boldsymbol{\alpha}}_i^{(t)})$ and all the $f(X_{ih}|\boldsymbol{\alpha}_l)s$ for $l = 1, 2, \ldots, L$. Note that $L = 2^K$ when there are K independent attributes. The KL index is defined as

$$KL_h\left(\widehat{\boldsymbol{\alpha}}_i^{(t)}\right) = \sum_{l=1}^{L} D_h(\widehat{\boldsymbol{\alpha}}_i^{(t)} \parallel \boldsymbol{\alpha}_l) \tag{36}$$

where

$$D_h(\widehat{\boldsymbol{\alpha}}_i^{(t)} \parallel \boldsymbol{\alpha}_l) = \sum_{q=0}^{1} \log\left(\frac{P(X_{ih} = q|\widehat{\boldsymbol{\alpha}}_i^{(t)})}{P(X_{ih} = q|\boldsymbol{\alpha}_l)}\right) P(X_{ih} = q|\widehat{\boldsymbol{\alpha}}_i^{(t)}). \tag{37}$$

Then the $(t+1)$th item for the ith examinee is the item in $R^{(t)}$ that maximizes $KL\left(\widehat{\boldsymbol{\alpha}}_i^{(t)}\right)$. The KL index $KL\left(\widehat{\boldsymbol{\alpha}}_i^{(t)}\right)$ is referred to as the global discrimination index (GDI) in Xu et al. (2003). This item selection method is referred to as the KL algorithm in Cheng (2009).

The KL algorithm selects items that are the most powerful in distinguishing the current attribute profile estimate from all other possible attribute profiles on average (Cheng, 2009). Cheng (2010) points out that the KL algorithm does not consider attribute coverage. Another drawback is that this algorithm may not be effective at the early stage with inaccurate $\widehat{\boldsymbol{\alpha}}_i^{(t)}$.

The posterior-weighted KL (PWKL) index. The PWKL index weights the KL index by the posterior distribution (Cheng, 2009). If informative priors π_{0l} are available for each attribute profile, posterior distributions can be obtained at each step t:

$$\pi_{i,t}\left(\boldsymbol{\alpha}_c|\boldsymbol{x}_i^{(t)}\right) \propto \pi_{0l} L\left(\boldsymbol{x}_i^{(t)}\middle|\boldsymbol{\alpha}_l\right). \tag{38}$$

Denote $\pi_{i,t}(\boldsymbol{\alpha}_l|\boldsymbol{x}_i^{(t)})$ by $\pi_{i,t}(\boldsymbol{\alpha}_l)$ for simplicity in notation. The PWKL index is defined as

$$PWKL_h\left(\widehat{\boldsymbol{\alpha}}_i^{(t)}\right) = \sum_{l=1}^{L} D_h\left(\widehat{\boldsymbol{\alpha}}_i^{(t)} \parallel \boldsymbol{\alpha}_l\right)\pi_{i,t}(\boldsymbol{\alpha}_l) = \sum_{l=1}^{L} D_h\left(\widehat{\boldsymbol{\alpha}}_i^{(t)} \parallel \boldsymbol{\alpha}_l\right)\pi_{0l}L\left(\boldsymbol{x}_i^{(t)}\middle|\boldsymbol{\alpha}_l\right). \quad (39)$$

Assuming local independence, the likelihood function $L(\boldsymbol{x}_i^{(t)}|\boldsymbol{\alpha}_l)$ can be written as

$$L\left(\boldsymbol{x}_i^{(t)}\middle|\boldsymbol{\alpha}_l\right) = \prod_{j=1}^{t}[P(\boldsymbol{\alpha}_l)]^{x_{ij}}\,[1 - P(\boldsymbol{\alpha}_l)]^{1-x_{ij}} \quad (40)$$

where $P(\boldsymbol{\alpha}_l)$ is the IRF defined by a CDM. Then the $(t+1)$th item for the ith examinee is the item in $R^{(t)}$ that maximizes PW$KL\left(\widehat{\boldsymbol{\alpha}}_i^{(t)}\right)$. If the prior is discrete uniform, the PWKL index is reduced to the likelihood-weighted KL (LWKL) index:

$$LWKL_h\left(\widehat{\boldsymbol{\alpha}}_i^{(t)}\right) = \sum_{l=1}^{L} D_h\left(\widehat{\boldsymbol{\alpha}}_i^{(t)} \parallel \boldsymbol{\alpha}_c\right) L\left(\boldsymbol{x}_i^{(t)}\middle|\boldsymbol{\alpha}_c\right). \quad (41)$$

*The modified posterior-weighted Kullback-Leibler (MPWKL) ind*ex. The KL and PWKL index use the current estimate $\widehat{\boldsymbol{\alpha}}_i^{(t)}$ with an implicit assumption that the point estimate is a good summary of the current information. However, the point estimate $\widehat{\boldsymbol{\alpha}}_i^{(t)}$ may be inaccurate especially at the early stages of a test. To solve this problem, Kaplan, de la Torre, and Barrada (2015) used the entire posterior distribution instead of a point estimate. The MPWKL index is given as

$$MPWKL_h^{(t)} = \sum_{m=1}^{L}\left[\sum_{l=1}^{L}\left[\sum_{q=0}^{1} log\left(\frac{P(X_{ih}=q|\boldsymbol{\alpha}_m)}{P(X_{ih}=q|\boldsymbol{\alpha}_l)}\right)P(X_{ih}=q|\boldsymbol{\alpha}_m)\pi_{i,t}(\boldsymbol{\alpha}_l)\right]\pi_{i,t}(\boldsymbol{\alpha}_m)\right]. \quad (42)$$

2.3.3 Item pool design

The potential benefits of CAT cannot be realized without a well-constructed item pool (Reckase, 2010). There are some studies on item pool design for CAT based on IRT models (e.g., Reckase, 2010; Thissen, Reeve, Bjorner, & Chang, 2007), and more research is needed in this area. Considering the difference between items based on IRT and CDM, the findings from IRT-based

CAT cannot be directly applied to CD-CAT. However, the item pool design for CD-CAT has not been addressed in the literature despite its importance.

Simulation findings on item usage in CD-CAT might inform the item pool design process (Kaplan et al., 2015). For example, a CD-CAT based on the DINA model tends to use items with a q-vector matching the examinee's true attribute profile and items that required single attributes which were not mastered by the examinee, which implies that it is important to include sufficient single-attribute items in the item pool.

Since there is no published research on item pool design for CD-CAT, the studies on the IRT-based CAT are reviewed below. There is a body of literature on selecting operational pools from a larger pool called a "master pool" (Belov & Armstrong, 2009; Swanson and Stocking, 1998; van der Linden, Ariel, & Veldkamp, 2006; Way, Steffen, & Anderson, 1998). The problem they address is related to item pool design but is more appropriately described as item pool assembly (van der Linden et al., 2006).

van der Linden et al. (2006) argues that an item pool design problem occurs before actual items are available and the output is a blueprint for an item pool that defines the distribution of numbers of items over the space of all possible combinations of statistical and nonstatistical item attributes (e.g., item difficulty parameter and word count). The goal of item pool design is to guide the item writing and pool maintenance process (Reckase, 2010; Veldkamp and van der Linden, 2000).

Item pool design studies for IRT-based CAT focuses on different aspects of an item pool. Veldkamp and van der Linden (2000) proposes a method for item pool design that minimizes item-writing costs subject to test constraints. Test constraints are represented in the classification table that contains all possible combinations of item attributes such as word counts, difficulty parameters,

difficulty parameters, and discrimination indices (Veldkamp & van der Linden, 2000). Quantitative attributes are transformed to categorical variables represented by intervals, for example, $(-\infty, -2.5), (-2.5, -2), \ldots, (2, 2.5), (2.5, \infty)$ for the difficulty parameter. The goal of the item pool design process is to find out the number of items needed for each cell of the classification table. The number of items in each cell of a previous item pool, however, is needed to define item writing costs as the inverse of that number, based on the idea that items written more frequently tend to be less costly.

Another stream of research based on the bin-and-union method (Reckase, 2010) explores item pool design without any information of existing item pools as a starting point (He & Reckase, 2014; Mao, 2014). This family of research focuses on the psychometric performances of item pools instead of the item-writing costs. Reckase (2010) thinks an optimal item pool should always provide the desired item for every item selection. An optimal item pool for a CAT procedure based on 1PL model, for example, is "one that has an item in the pool that has a b-parameter exactly equal to the current θ estimate for every item selection." (Reckase, 2010) The size of an optimal item pool is $2^n - 1$ where n is the test length, which is too large to be practical. If the latent scale is divided into bins and the items with b-parameters within a bin are treated equivalent, the item pool size will be greatly decreased to a reasonable level. The definition of "bins" is similar to the categorization of the difficulty parameter in Veldkamp and van der Linden (2000).

The item pool design methods of Veldkamp and van der Linden (2000) and Recakse (2010; also see He & Reckase, 2014; Mao, 2014) are based on different definitions of optimal item pool, but a common feature they share is the use of computer simulation. The simulations in Veldkamp and van der Linden (2000) are carried out using integer programming and the shadow test approach (van der Linden, 2005a, 2005b; van der Linden & Diao, 2014; van der Linden & Reese, 1998) and

sampling examinees from a hypothetical examinee distribution. The goal is to record the counts of the number of times items from each cell in the classification table are used, and the final blueprint is calculated from these counts (Veldkamp & van der Linden, 2000). The bin-and-union method (Reckase, 2010) takes a more direct approach by simulating an operational CAT and sampling from an examinee population.

Chapter 3 CDM parameterization and Q-matrix with hierarchical attributes

3.1 Introduction

The CDMs with a restricted attribute profile space due to the attribute hierarchy is henceforth referred to as hierarchical CDMs. This section addresses parameterizations and the Q-matrix of hierarchical CDMs. Parameterizations for hierarchical CDMs have not been formally discussed except for the HDCM (Liu et al., 2017; Templin & Bradshaw, 2014) and the DINA model. When it comes to the Q-matrix, two types of Q-matrices are being used by two groups of researchers, respectively: the full (or unrestricted) Q-matrices (Liu et al., 2016; Templin & Bradshaw, 2014) and the reduced (or restricted) Q-matrices (Köhn & Chiu, 2018; Leighton et al., 2004; Tu et al., 2018). The choice between the full Q-matrix and the restricted one has not been formally addressed.

Therefore, the first set of research questions is about the parametrization of hierarchical CDMs and the difference between reduced and full Q-matrix. These questions are important because the test constructions and item pool designs all depend on correctly-defined CDMs and Q-matrices.

In this thesis, it is assumed that the hierarchical relationship and the Q-matrix have been established and validated, and we focus on test construction or item pool design for different types of attribute hierarchies.

3.2 Attribute hierarchies

Before discussing parameterizations and Q-matrices, we define the attribute hierarchies studied in this thesis. The formative assessment is designed for a period of two to four weeks. Therefore, we consider situations with three, four, or five attributes in this study. The subsets of attribute hierarchies chosen for 3-attribute, 4-attribute, or 5-attribute conditions, respectively, are

listed in Table 1 and illustrated in Figure 6-Figure 8. Most of the selected attribute hierarchies can be found in the textbook analysis, as well as previous empirical and simulation studies.

Table 1: Subsets of attribute hierarchies for 3-attribute, 4-attribute, or 5-attribute conditions

ID	Number of attributes	Size of attribute profile space	Attribute hierarchy
H3.1	3	8	Independent
H3.2	3	4	Linear
H3.3	3	5	Inverted pyramid
H3.4	3	5	Pyramid
H4.1	4	16	Independent
H4.2	4	5	Linear
H4.3	4	8	Linear + single
H4.4	4	6	Inverted pyramid
H4.5	4	6	Pyramid
H5.1	5	32	Independent
H5.2	5	6	Linear
H5.3	5	10	Inverted pyramid I
H5.4	5	11	Inverted pyramid II
H5.5	5	10	Pyramid I
H5.6	5	11	Pyramid II

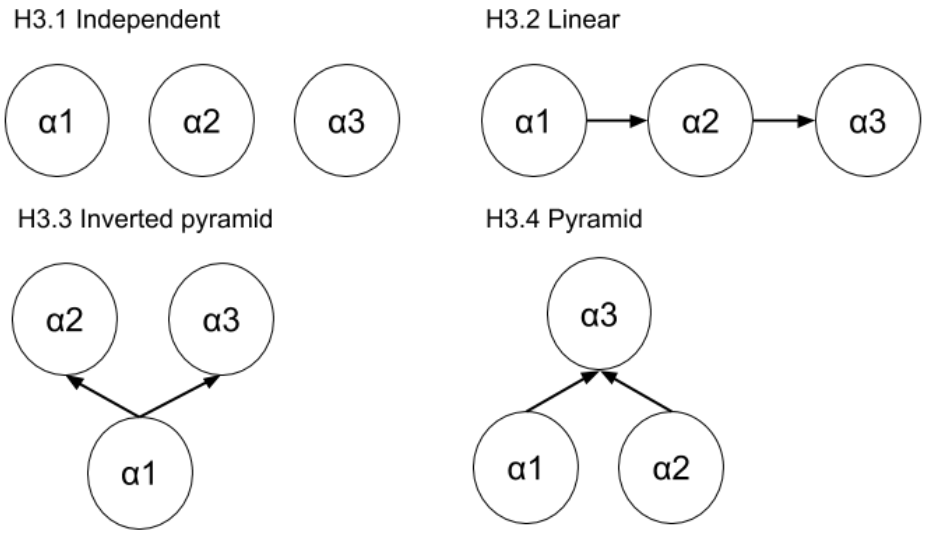


Figure 6: A subset of attribute hierarchies with 3 attributes

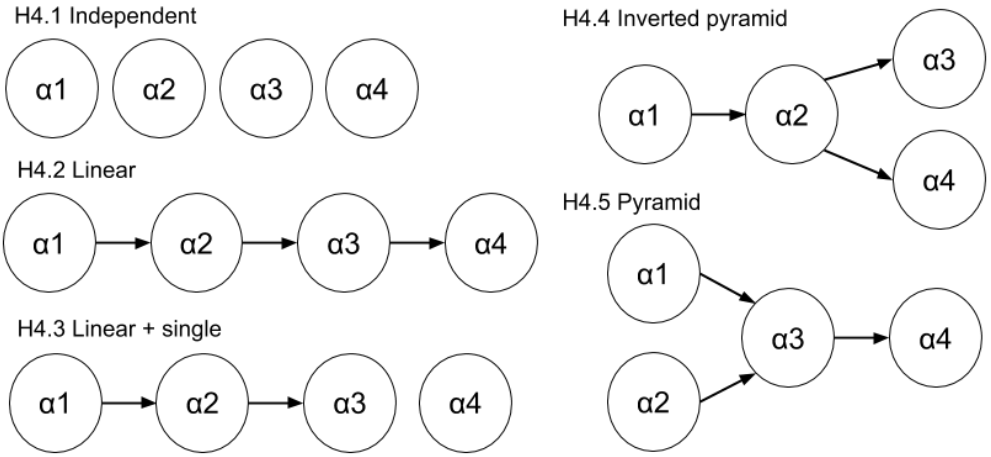


Figure 7: A subset of attribute hierarchies with 4 attributes

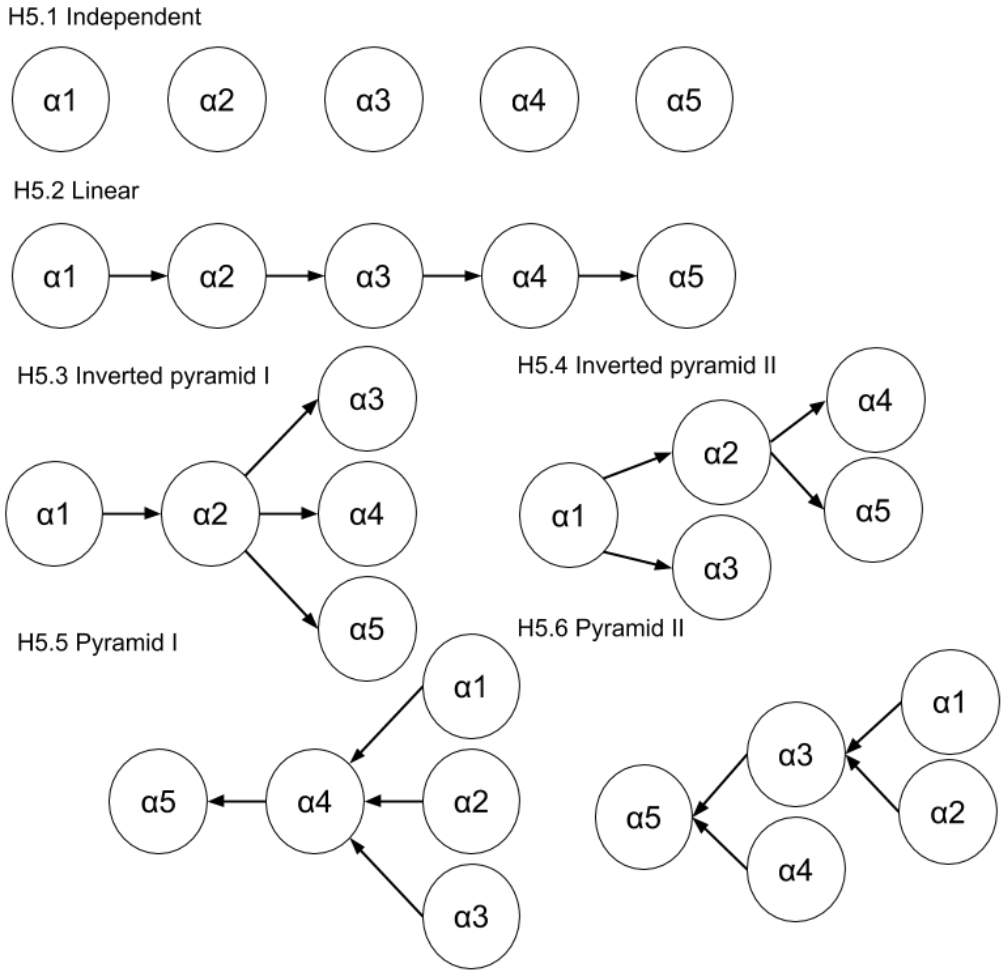


Figure 8: A subset of attribute hierarchies with 5 attributes

3.3 Parameterizations of hierarchical CDMs

We discuss the parameterizations for the DINA (Junker & Sijtsma, 2001), ACDM (de la Torre, 2011), and GDINA model with the identity link function (de la Torre, 2011) when the attributes are hierarchical.

An item requiring K attributes can classify students into at most 2^K classes. A hierarchical relationship among attributes leads to fewer than 2^K classes. A saturated model for an item requiring K independent attributes can have 2^K item parameters including an intercept, K main effect terms, and $2^K - K - 1$ interaction terms. The number of item parameters cannot exceed the number of classes.

The parameterizations for DINA and ACDM do not change with hierarchical attributes. The DINA model has two parameters for each item disregarding the q-vector of the item: an intercept and an interaction term (or a guessing parameter and a slipping parameter in an alternative parameterization). Under the A-CDM, an item requiring K independent attributes has $K + 1$ item parameters (i.e., one intercept and K main effect terms).

For GDINA, some item parameters (i.e., the main effects of nested attributes and some interaction terms) need to be fixed at zero, which is parallel to the parameterizations of the Hierarchical Diagnostic Classification Model (HDCM; Templin & Bradshaw, 2014).

Before demonstrating the parameterizations of hierarchical models, we present the parameterizations of three models—DINA, ACDM, and GDINA—for a single-attribute item and a two-independent-attribute item. The three models are equivalent regarding a single-attribute item but have different parameterizations for an item requiring two independent attributes, which are shown in Table 2 in the form of expected response $E\left[Y_j \middle| \boldsymbol{\alpha}\right]$.

Table 2: Expected responses on two items with two independent attributes

$\boldsymbol{\alpha}$	$\boldsymbol{q}_j = (0\ 1)$	$\boldsymbol{q}_j = (1\ 1)$		
	Any model	DINA	ACDM	GDINA
(00)	ϕ_{j0}	ϕ_{j0}	ϕ_{j0}	ϕ_{j0}
(10)	ϕ_{j0}	ϕ_{j0}	$\phi_{j0} + \phi_{j1}$	$\phi_{j0} + \phi_{j1}$
(01)	$\phi_{j0} + \phi_{j2}$	ϕ_{j0}	$\phi_{j0} + \phi_{j2}$	$\phi_{j0} + \phi_{j2}$
(11)	$\phi_{j0} + \phi_{j2}$	$\phi_{j0} + \phi_{j,12}$	$\phi_{j0} + \phi_{j1} + \phi_{j2}$	$\phi_{j0} + \phi_{j1} + \phi_{j2} + \phi_{j,12}$

Note: Item j involves two independent attributes α_1 and α_2; all models the identity link; DINA = deterministic input noisy "and" gate; ACDM = additive cognitive diagnosis modeling; GDINA = generalized DINA; ϕ_{j0} = intercept; ϕ_{jk} = main effect of the kth attribute ($k = 1, 2$); $\phi_{j,12}$ = two-way interaction.

Suppose α_1 is the prerequisite of α_2 (i.e., $\alpha_1 \rightarrow \alpha_2$). The item $\boldsymbol{q}_j = (0\ 1)$, under each model (DINA, A-CAM, or GDINA), classifies examinees into two groups: those who master both α_2 and its prerequisite α_1 and those who have not mastered α_2. The parameterizations of the three hierarchical models are in Table 3. Under the DINA model, the item $\boldsymbol{q}_j = (1\ 1)$ has the same parameterizations as $\boldsymbol{q}_j = (0\ 1)$. For the parameterizations of the item $\boldsymbol{q}_j = (1\ 1)$ under GDINA, the main effect of the higher-level attribute (i.e., α_2) needs to be fixed at zero. Both ACDM and GDINA have three item parameters. ACDM has an intercept and two main effects. GDINA has an intercept, a main effect, and an interaction effect. Although parameterized differently, the two models become mathematically equivalent for an item measuring two linear attributes.

Table 3: Expected responses on two items with two linear attributes ($\alpha_1 \rightarrow \alpha_2$)

$\boldsymbol{\alpha}$	$\boldsymbol{q}_j = (0\ 1)$	$\boldsymbol{q}_j = (1\ 1)$		
	Any model	DINA	ACDM	GDINA
(00)	ϕ_{j0}	ϕ_{j0}	ϕ_{j0}	ϕ_{j0}
(10)	ϕ_{j0}	ϕ_{j0}	$\phi_{j0} + \phi_{j1}$	$\phi_{j0} + \phi_{j1}$
(11)	$\phi_{j0} + \phi_{j2}$	$\phi_{j0} + \phi_{j,12}$	$\phi_{j0} + \phi_{j1} + \phi_{j2}$	$\phi_{j0} + \phi_{j1} + \phi_{j,12}$

Note: Item j involves two attributes α_1 and α_2 under a linear hierarchy; all models the identity link; DINA = deterministic input noisy "and" gate; ACDM = additive cognitive diagnosis modeling; GDINA = generalized DINA; ϕ_{j0} = intercept; ϕ_{jk} = main effect of the kth attribute ($k = 1, 2$); $\phi_{j,12}$ = two-way interaction.

Next, we consider a situation involving three attributes with one attribute being the prerequisite of the other two as in an inverted pyramid hierarchy (H3.3). Table 4 presents the parameterizations of three models for $\boldsymbol{q}_j = (1\ 1\ 1)$. For this item, the three models have different parameterizations. The difference between ACDM and GDINA lies in the interaction effect between α_2 and α_3.

Table 4: Expected responses on $\boldsymbol{q}_j = (1\ 1\ 1)$ under an inverted pyramid hierarchy (H3.3)

$\boldsymbol{\alpha}$	$\boldsymbol{q}_j = (1\ 1\ 1)$		
	DINA	ACDM	GDINA
(000)	ϕ_{j0}	ϕ_{j0}	ϕ_{j0}
(100)	ϕ_{j0}	$\phi_{j0} + \phi_{j1}$	$\phi_{j0} + \phi_{j1}$
(110)	ϕ_{j0}	$\phi_{j0} + \phi_{j1} + \phi_{j2}$	$\phi_{j0} + \phi_{j1} + \phi_{j,12}$
(101)	ϕ_{j0}	$\phi_{j0} + \phi_{j1} + \phi_{j3}$	$\phi_{j0} + \phi_{j1} + \phi_{j,13}$
(111)	$\phi_{j0} + \phi_{j,123}$	$\phi_{j0} + \phi_{j1} + \phi_{j2} + \phi_{j3}$	$\phi_{j0} + \phi_{j1} + \phi_{j,12} + \phi_{j,13} + \phi_{j,123}$

Note: The inverted pyramid hierarchy defines $\alpha_1 \rightarrow \alpha_2$, $\alpha_1 \rightarrow \alpha_3$. α_2 and α_3 do not share a common path.

We then consider a situation involving three attributes with two attributes being the prerequisite of the third one as in a pyramid hierarchy (H3.4). Table 5 presents the parameterizations of three models for $\boldsymbol{q}_j = (1\ 1\ 1)$. For this item, the three models have different parameterizations. The difference between ACDM and GDINA lies in the interaction effect between α_1 and α_2.

Table 5: Expected responses on $\boldsymbol{q}_j = (1\ 1\ 1)$ under a pyramid hierarchy (H3.4)

$\boldsymbol{\alpha}$	$\boldsymbol{q}_j = (1\ 1\ 1)$		
	DINA	ACDM	GDINA
(000)	ϕ_{j0}	ϕ_{j0}	ϕ_{j0}
(100)	ϕ_{j0}	$\phi_{j0} + \phi_{j1}$	$\phi_{j0} + \phi_{j1}$
(010)	ϕ_{j0}	$\phi_{j0} + \phi_{j2}$	$\phi_{j0} + \phi_{j2}$
(110)	ϕ_{j0}	$\phi_{j0} + \phi_{j1} + \phi_{j2}$	$\phi_{j0} + \phi_{j1} + \phi_{j2} + \phi_{j12}$
(111)	$\phi_{j0} + \phi_{j3}$	$\phi_{j0} + \phi_{j1} + \phi_{j2} + \phi_{j3}$	$\phi_{j0} + \phi_{j1} + \phi_{j2} + \phi_{j12} + \phi_{j,123}$

Note: The pyramid hierarchy defines $\alpha_1 \rightarrow \alpha_3$, $\alpha_2 \rightarrow \alpha_3$. α_1 and α_2 do not share a common path.

3.4 Q-matrix of hierarchical CDMs

3.4.1 Reduced or full Q-matrix

In previous studies, either a reduced Q-matrix or a full Q-matrix is used. With hierarchical attributes, the argument is around whether it is possible for an item to measure a higher-level attribute without measuring its prerequisite(s). A full Q-matrix allows all types of q-vectors as in an independent-attribute situation. A reduced Q-matrix requires that items that measure a higher-level attribute also require all its prerequisite(s). In other words, a reduced Q-matrix can only

contain q-vectors in Q_r (the transpose of the expanded R-matrix R^*). We will demonstrate that the reduced Q-matrix approach is equivalent to the full Q-matrix approach under the DINA model.

It can be shown that, under the DINA model, a multiple-attribute item $\mathbf{q_1} = (q_1 \ldots q_{k-1}\ 1_k\ 0_{k+1} \ldots 0_K)$ is equivalent to the single-attribute item $\mathbf{q_2} = (0_1 \ldots 0_{k-1}\ 1_k\ 0_{k+1} \ldots 0_K)$, in which q_i $(i = 1, \ldots, k-1)$ takes the value 1 or 0 if the previous $k-1$ attributes are the direct or indirect prerequisites of the kth attribute, or takes the value 0 if the ith attribute is not connected with the kth attribute in any path. The multiple-attribute item $\mathbf{q_1}$ and the single-attribute item $\boldsymbol{q}_2$ are equivalent because they classify attribute profiles into the same two groups (i.e., $\boldsymbol{\alpha}$s mastering the kth attribute or not), and they have the same expected response for each group as shown in Table 6.

Therefore, under the DINA model with a linear hierarchy, the reduced Q-matrix $\mathrm{Q_r}$ is equivalent to an identity matrix consisting of K single-attribute q-vectors. Table 7 presents the equivalent q-vectors for each row of Q_r in the case of three linear attributes.

Under the DINA model and any attribute hierarchy, each q-vector in $\mathrm{Q_r}$ represents a unique type of items (Table 7-Table 10). Other q-vectors can find their equivalent one in $\mathrm{Q_r}$. Consequently, there would be no difference between the reduced Q-matrix approach and the full Q-matrix approach under the DINA model. However, it is noteworthy that there are less than $2^K - 1$ distinctive q-vectors with hierarchical attributes.

Note that all the single-attribute items are included in Q_r under the DINA model. Under the ACDM or GDINA, however, each q-vector is distinctive, and consequently Q_r does not include all the single-attribute items. We use H3.2 under the ACDM to demonstrate this in Table 11. If the reduced Q-matrix approach is used with ACDM or GDINA, it means that some single-attribute q-vectors will be excluded from the Q-matrix.

Table 6: The expected responses of two groups of attribute profiles on $\mathbf{q_1}$ and $\mathbf{q_2}$ under the DINA model

$\boldsymbol{\alpha}$	$\boldsymbol{q_1} = (q_1 \ldots q_{k-1}\ 1_k\ 0_{k+1} \ldots 0_K)$	$\boldsymbol{q_2} = (0_1 \ldots 0_{k-1}\ 1_k\ 0_{k+1} \ldots 0_K)$
$\boldsymbol{\alpha}s\ with\ \alpha_k = 0$	ϕ_{j0}	ϕ_{j0}
$\boldsymbol{\alpha}s\ with\ \alpha_k = 1$	$\phi_{j0} + \phi_{j,all}$	$\phi_{j0} + \phi_{j,all}$

Note: α_k stands for the kth attribute; q_i $(i = 1, \ldots, k-1)$ takes the value 1 or 0 if the previous $k-1$ attributes are the direct or indirect prerequisites of the kth attribute, or takes the value 0 if the i th attribute is not connected with the k th attribute in any path; ϕ_{j0} = intercept; $\phi_{j,all}$ = interaction.

Table 7: The q-vectors in Q_r and their equivalent q-vectors under the DINA model with three linear attributes (H3.2)

Q_r	Equivalent $\boldsymbol{q}$s	Attribute Profiles $\boldsymbol{\alpha}$	
		$E[Y_j \mid \boldsymbol{\alpha}] = \phi_{j0}$	$\phi_{j0} + \phi_{j,all}$
(1 0 0)		(0 0 0)	(1 0 0) (1 1 0) (1 1 1)
(1 1 0)	**(0 1 0)**	(0 0 0) (1 0 0)	(1 1 0) (1 1 1)
(1 1 1)	**(0 0 1)** (1 0 1) (0 1 1)	(0 0 0) (1 0 0) (1 1 0)	(1 1 1)

Note: Single-attribute items are bolded; ϕ_{j0} = intercept; $\phi_{j,all}$ = interaction.

Table 8: The q-vectors in Q_r and their equivalent q-vectors under the DINA model with three inverted pyramid attributes (H3.3)

Q_r	Equivalent $\boldsymbol{q}$s	Attribute Profiles $\boldsymbol{\alpha}$	
		$E[Y_j \mid \boldsymbol{\alpha}] = \phi_{j0}$	$\phi_{j0} + \phi_{j,all}$
(1 0 0)		(0 0 0)	(1 0 0) (1 1 0) (1 0 1) (1 1 1)
(1 1 0)	**(0 1 0)**	(0 0 0) (1 0 0)	(1 1 0) (1 0 1) (1 1 1)
(1 0 1)	**(0 0 1)**	(0 0 0) (1 0 0) (1 1 0)	(1 0 1) (1 1 1)
(1 1 1)*	(0 1 1)	(0 0 0) (1 0 0) (1 1 0) (1 0 1)	(1 1 1)

Note: Single-attribute items are bolded; ϕ_{j0} = intercept; $\phi_{j,all}$ = interaction; * = q-vector that is not in the R-matrix.

Table 9: The q-vectors in Q_r and their equivalent q-vectors under the DINA model with three pyramid attributes (H3.4)

Q_r	Equivalent $\boldsymbol{q}$s	Attribute Profiles $\boldsymbol{\alpha}$	
		$E[Y_j \mid \boldsymbol{\alpha}] = \phi_{j0}$	$\phi_{j0} + \phi_{j,all}$
(1 0 0)		(0 0 0) (0 1 0)	(1 0 0) (1 1 0) (1 1 1)
(0 1 0)		(0 0 0) (1 0 0)	(0 1 0) (1 1 0) (1 1 1)
(1 1 0)*		(0 0 0) (1 0 0) (0 1 0)	(1 1 0) (1 1 1)
(1 1 1)	**(0 0 1)** (1 0 1) (0 1 1)	(0 0 0) (1 0 0) (0 1 0) (1 1 0)	(1 1 1)

Note: Single-attribute items are bolded; ϕ_{j0} = intercept; $\phi_{j,all}$ = interaction; * = q-vector that is not in the R-matrix.

Table 10: The q-vectors in Q_r and their equivalent q-vectors under the DINA model with four or five attributes

Hierarchy	Q_r	Equivalent $\boldsymbol{q}$s	Hierarchy	Q_r	Equivalent $\boldsymbol{q}$s
H4.2	**(1000)**		H5.4	**(10000)**	
	(1100)	**(0100)**		(11000)	**(01000)**
	(1110)	(qq10), e.g.,**(0010)**		(10100)	**(00100)**
	(1111)	(qqq1), e.g.,**(0001)**		(11100)	(01100)
H4.3	**(1000)**			(11010)	(qq010), e.g., **(00010)**
	(0001)			(11001)	(qq001), e.g., **(00001)**
	(1100)	**(0100)**		(11110)	(qq110)
	(1001)			(11101)	(qq101)
	(1110)	(qq10), e.g., **(0010)**		(11011)	(qq011)
	(1101)	(0101)		(11111)	(qq111)
	(1111)	(0111) (1011) (0011)	H5.5	**(10000)**	
H4.4	**(1000)**			**(01000)**	
	(1100)	**(0100)**		**(00100)**	
	(1110)	(qq10), e.g.,**(0010)**		(11000)	
	(1101)	(qq01), e.g., **(0001)**		(10100)	
	(1111)	(qq11)		(01100)	
H4.5	**(1000)**			(11100)	
	(0100)			(11110)	(qqq10), e.g., **(00010)**
	(1100)			(11111)	(qqqq1), e.g., **(00001)**
	(1110)	(qq10), e.g., **(0010)**	H5.6	**(10000)**	
	(1111)	(qqq1), e.g., **(0001)**		**(01000)**	
H5.2	**(10000)**			**(00010)**	
	(11000)	**(01000)**		(11000)	
	(11100)	(qq100), e.g., **(00100)**		(10010)	
	(11110)	(qqq10), e.g., **(00010)**		(01010)	
	(11111)	(qqqq1) e.g., **(00001)**		(11100)	(qq100), e.g., **(00100)**
H5.3	**(10000)**			(11010)	
	(11000)	**(01000)**		(11110)	(qq110)
	(11100)	(qq100), e.g., **(00100)**		(11111)	(qqqq1), e.g., **(00001)**
	(11010)	(qq010), e.g., **(00010)**			
	(11001)	(qq001), e.g., **(00001)**			
	(11110)	(qq110)			
	(11101)	(qq101)			
	(11011)	(qq011)			
	(11111)	(qq111)			

Note: q takes the value of 0 or 1. Single-attribute items are bolded.

Table 11: The q-vectors in Q_r and their equivalent q-vectors under the ACDM with three linear attributes (H3.2)

Q_r	Other $\boldsymbol{q}$	Attribute Profiles $\boldsymbol{\alpha}$			
		$E[Y_j \mid \boldsymbol{\alpha}] = \phi_{j0}$	$\phi_{j0} + \phi_{jk}$	$\phi_{j0} + \phi_{jk_1} + \phi_{jk_2}$	$\phi_{j0} + \Sigma\phi_{jk}$
(100)		(000)	(100) (110) (111)		
(110)		(000)	(100)	(110) (111)	
	(010)	(000) (100)	(110) (111)		
	(001)	(000) (100) (110)	(111)		
	(101)	(000)	(100) (110)	(111)	
	(011)	(000) (100)	(110)	(111)	
(111)		(000)	(100)	(110)	(111)

Note: Single-attribute items are bolded; ϕ_{j0} = intercept; ϕ_{jk} = main effect of attribute k.

If the reduced Q-matrix approach is taken, there are only three q-vectors under ACDM. However, if the model selection is made at the item level and an item pool of mixed models can be constructed (Ma et al., 2015), items calibrated with the DINA model can be included in this item pool. For the linear hierarchy H3.2, for example, the mixed item pool has five distinct item types in Table 12. If the full Q-matrix approach is taken instead, the mixed item pool can have two more item types: $\boldsymbol{q} = (1\ 0\ 1)$ and $\boldsymbol{q} = (0\ 1\ 1)$ calibrated by the ACDM.

Table 12: Distinct q-vectors in a mixed item pool under DINA and ACDM for H3.2 using the reduced Q-matrix approach

q	Model	Attribute Profiles $\boldsymbol{\alpha}$			
		$E[Y_j \mid \boldsymbol{\alpha}] = \phi_{j0}$	$\phi_{j0} + \phi_{jk}$	$\phi_{j0} + \phi_{jk_1} + \phi_{jk_2}$	$\phi_{j0} + \Sigma\phi_{jk}$
(100)	-	(000)	(100) (110) (111)		
(110)	ACDM	(000)	(100)	(110) (111)	
(110)	DINA	(000) (100)	(110) (111)		
(111)	DINA	(000) (100) (110)	(111)		
(111)	ACDM	(000)	(100)	(110)	(111)

Note: Single-attribute items are bolded; ϕ_{j0} = intercept; ϕ_{jk} = main effect of attribute k.

3.4.2 Complete Q-matrix for hierarchical attributes

A Q-matrix containing the identity matrix is complete for the DINA model with independent attributes, according to Chiu et al. (2009). Since the completeness of a Q-matrix is evaluated by checking whether it holds that $S(\boldsymbol{\alpha}) = S(\boldsymbol{\alpha}') \Longrightarrow \boldsymbol{\alpha} = \boldsymbol{\alpha}'$ for each pair of $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ in the attribute profile space, the completeness will not change if some $\boldsymbol{\alpha}$s are excluded from the attribute profile space. Since there is only one way to define single-attribute items under different models, it is safe to conclude that the identity matrix is complete for any attribute hierarchy under any model. Under the DINA model, Q_r is complete since Q_r equals to or contains the identity matrix; another type of complete matrix is the transpose of the R-matrix that equals to the identity matrix, consistent with the conclusion of Köhn and Chiu (2018). The expected response vectors given $\boldsymbol{\alpha}$ are presented in Table 13.

Table 13: Expected response vectors given $\boldsymbol{\alpha}$ of two Q-matrices (Q_r and I) for the inverted pyramid (H3.3) under the DINA model

$\boldsymbol{\alpha}$	Q_r				I		
	$\boldsymbol{q}_1 = (100)$	$\boldsymbol{q}_2 = (110)$	$\boldsymbol{q}_3 = (101)$	$\boldsymbol{q}_4 = (111)$	$\boldsymbol{q}_5 = (100)$	$\boldsymbol{q}_6 = (010)$	$\boldsymbol{q}_7 = (001)$
	$S(\boldsymbol{\alpha})$				$S(\boldsymbol{\alpha})$		
(000)	ϕ_{10}	ϕ_{20}	ϕ_{30}	ϕ_{40}	ϕ_{50}	ϕ_{60}	ϕ_{70}
(100)	$\phi_{10} + \phi_{1,\text{all}}$	ϕ_{20}	ϕ_{30}	ϕ_{40}	$\phi_{50} + \phi_{5,\text{all}}$	ϕ_{60}	ϕ_{70}
(110)	$\phi_{10} + \phi_{1,\text{all}}$	$\phi_{20} + \phi_{2,\text{all}}$	ϕ_{30}	ϕ_{40}	$\phi_{50} + \phi_{5,\text{all}}$	$\phi_{60} + \phi_{6,\text{all}}$	ϕ_{70}
(101)	$\phi_{10} + \phi_{1,\text{all}}$	ϕ_{20}	$\phi_{30} + \phi_{3,\text{all}}$	ϕ_{40}	$\phi_{50} + \phi_{5,\text{all}}$	ϕ_{60}	$\phi_{70} + \phi_{7,\text{all}}$
(111)	$\phi_{10} + \phi_{1,\text{all}}$	$\phi_{20} + \phi_{2,\text{all}}$	$\phi_{30} + \phi_{3,\text{all}}$	$\phi_{40} + \phi_{4,\text{all}}$	$\phi_{50} + \phi_{5,\text{all}}$	$\phi_{60} + \phi_{6,\text{all}}$	$\phi_{70} + \phi_{7,\text{all}}$

Note: Single-attribute items are bolded; ϕ_{j0} = intercept; ϕ_{jk} = main effect of attribute k.

Under ACDM, one type of items— $\boldsymbol{q}_4 = (1\ 1\ 1)$ —alone would be sufficient for completeness by definition as long as the three main effects (ϕ_{41}, ϕ_{42}, and ϕ_{44}) are different from each other (Table 14). Without assuming the differences between ϕ_{41}, ϕ_{42}, and ϕ_{44}, an inspection of Table 14 shows that Q_r of each attribute hierarchy is a complete Q-matrix disregarding the attribute hierarchy.

Table 14: Expected response vectors given $\boldsymbol{\alpha}$ of five q-vectors for independent attributes under ACDM

$\boldsymbol{\alpha}$	$\boldsymbol{q}_1 = (100)$	$\boldsymbol{q}_2 = (110)$	$\boldsymbol{q}_3 = (101)$	$\boldsymbol{q}_4 = (111)$	$\boldsymbol{q}_5 = (010)$
(000)	ϕ_{10}	ϕ_{20}	ϕ_{30}	ϕ_{40}	ϕ_{50}
(100)	$\phi_{10} + \phi_{11}$	$\phi_{20} + \phi_{21}$	$\phi_{30} + \phi_{31}$	$\phi_{40} + \phi_{41}$	ϕ_{50}
(010)	ϕ_{10}	$\phi_{20} + \phi_{22}$	ϕ_{30}	$\phi_{40} + \phi_{42}$	$\phi_{50} + \phi_{52}$
(001)	ϕ_{10}	ϕ_{20}	$\phi_{30} + \phi_{33}$	$\phi_{40} + \phi_{43}$	ϕ_{50}
(110)	$\phi_{10} + \phi_{11}$	$\phi_{20} + \phi_{21} + \phi_{22}$	$\phi_{30} + \phi_{31}$	$\phi_{40} + \phi_{41} + \phi_{42}$	$\phi_{50} + \phi_{52}$
(101)	$\phi_{10} + \phi_{11}$	$\phi_{20} + \phi_{21}$	$\phi_{30} + \phi_{31} + \phi_{33}$	$\phi_{40} + \phi_{41} + \phi_{43}$	ϕ_{50}
(011)	ϕ_{10}	$\phi_{20} + \phi_{22}$	$\phi_{30} + \phi_{33}$	$\phi_{40} + \phi_{41} + \phi_{43}$	$\phi_{50} + \phi_{52}$
(111)	$\phi_{10} + \phi_{11}$	$\phi_{20} + \phi_{21} + \phi_{22}$	$\phi_{30} + \phi_{31} + \phi_{33}$	$\phi_{40} + \phi_{41} + \phi_{42} + \phi_{43}$	$\phi_{50} + \phi_{52}$

Note: q_1, q_2, and q_4 form the Q_r for the linear hierarchy (H3.2); q_1, q_2, q_3, and q_4 form the Q_r for the inverted pyramid hierarchy (H3.3); q_1, q_2, q_4, and q_5 form the Q_r for the pyramid hierarchy (H3.4).

3.5 Summary

In discussing the parameterizations of hierarchical CDMs, we identified equivalent models when an attribute hierarchy is present. The three models in the GDINA family parameterize single-attribute items in the same way regardless of the attribute hierarchy. The hierarchical ACDM and hierarchical GDINA model are equivalent to each other but different from the hierarchical DINA model when two linear attributes are involved in an item. The hierarchical ACDM and GDINA

model have different parameterizations when two "independent" attributes are involved. Independence refers to the fact that the two attributes are not on the same path in the tree graph.

Under the hierarchical DINA model, the q-vectors in Q_r represent distinct item types. Since the number of q-vectors in Q_r is smaller than $2^K - 1$, a full Q-matrix may have two seemingly different q-vectors that are are equivalent. By equivalence, we mean that the items have the same parameterizations and would thus lead to the same classifications of examinees given the same item parameters. For example, $\boldsymbol{q} = (1\ 0\ 1)$ and $\boldsymbol{q} = (1\ 1\ 1)$ are equivalent to $\boldsymbol{q} = (0\ 0\ 1)$ in the hierarchical DINA model if attribute α_1 and attribute α_2 are prerequisite to attribute α_3. As a result, the choice between the reduced and the full Q-matrix approaches does not make a difference under the hierarchical DINA model.

Under the ACDM or GDINA model, any combination of attributes is a distinct q-vector so there are in theory $2^K - 1$ different item types. A reduced Q-matrix under the hierarchical ACDM or GDINA model inevitably excludes the single-attribute items for the higher-level attributes. For example, a reduced Q-matrix Q_r in H3.4 (pyramid hierarchy) only includes two single-attribute q-vectors corresponding to the two lower-level attributes. The single-attribute q-vector for the other attribute is excluded from a reduced Q-matrix. The absence of single-attribute q-vectors in the recuded Q-matrices may have serious impact on the classifications, which is discussed in the next chapter.

Chapter 4 Conditional KLI-based indexes for hierarchical CDMs

4.1 Introduction

In the previous chapter, we discuss two approaches to constructing Q-matrices with hierarchical attributes. We mainly talk about equivalent q-vectors and complete Q-matrices. There are, however, numerous ways to construct the Q-matrix for a test from all the available q-vectors. Previous studies in Q-matrix design simulate tests with different Q-matrices to compare the classification results (Chiu et al., 2009; Liu & Huggins-Manley, 2016; Liu et al., 2017; Madison & Bradshaw, 2015). We address the issue of Q-matrix design from the perspective of item-level and test-level indices. The indices can be used to automate test assembly with a calibrated item pool. The indices also provide a basis for comparing different Q-matrix designs.

The existing item-level and test-level indexes based on the KL information are overall indexes for a population of examinees, and they are found to be positively correlated with the overall classification rates (Henson & Douglas, 2005; Kuo et al., 2016). However, the correct classification rates (CCRs) could vary substantially across different attribute profiles within the same test regardless of independent or hierarchical attributes. The CCRs conditional on the attribute profile are usually not reported as most studies only calculate an overall CCR for the population of examinees.

With independent attributes, the conditional CCRs are different for different attribute profiles when each attribute is measured in different numbers of items. In this situation, attribute-level indexes could compensate for overall indices for items or tests (Henson et al., 2008; Kuo et al., 2016). However, the attribute-level index ADI fails to consider the dependency between attributes as a result of attribute hierarchies. To address this problem, the modified ADI proposed

by Kuo and colleagues (2016) add weights on the original ADI but remains to be an overall index for a population of examinees.

The following examples show the necessity for conditional indices instead of an overall index. Suppose items are calibrated with the DINA model and the intercept $\phi_0 = 0.1$ and the interaction effect $\phi_{all} = 0.8$ for all items. The Q-matrix Q_3 contains a multiple-attribute item in addition to three single-attribute items. When Q_3 is used for three independent attributes, different attribute profiles have substantially different conditional CCRs. Another example Q_4 is the identity matrix but is used for measuring three linear attributes (i.e., $\alpha_1 \rightarrow \alpha_2 \rightarrow \alpha_3$). The CCRs for complete mastery and complete non-mastery are higher than other profiles.

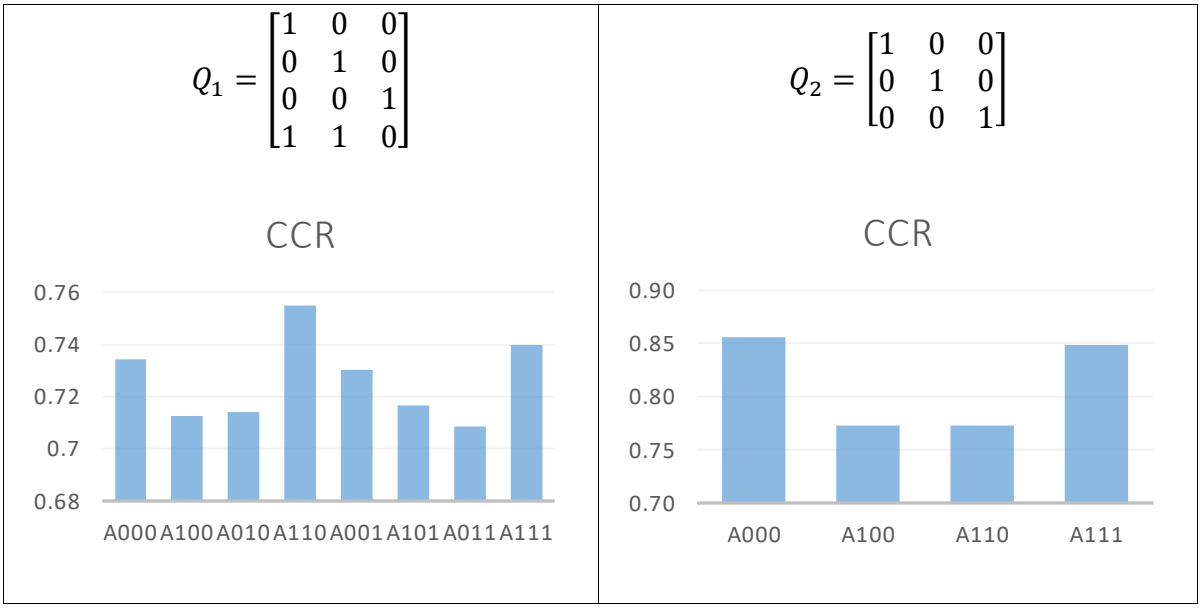


Figure 9: Correct classification rates under two conditions

Since the goal is to estimate the attribute profile for every examinee accurately, it is necessary to develop an index conditional on the attribute profile, especially when hierarchical attributes are present. This thesis proposes two conditional indices based on the KL information that can be used for non-adaptive test construction and Q-matrix design.

In this chapter, it is assumed that a large number of items have been developed for a well-defined domain and that the Q-matrix, as well as the relationship between attributes, are correctly specified. We take the full Q-matrix approach and allow all types of q-vectors. It is also assumed that item parameter estimates have been obtained from previous calibrations.

4.2 Conditional KL indices for test construction

A set of two indices is proposed, conditional on the attribute profile. The two conditional indices summarize the information from the L-by-L test KLI matrix D as defined in equation (27) in 2.1.7, where L is the size of the attribute profile space. The first index is the average of the elements in the uth column and the uth row of the test KLI matrix. The second index is the

$$mean\ KLI(\boldsymbol{\alpha}_u) = \ln\left(\frac{1}{2(L-1)}\left(\sum_{l=1}^{L} D_{.ul} + \sum_{l=1}^{L} D_{.lu}\right)\right), \tag{43}$$

$$range\ KLI(\boldsymbol{\alpha}_u) = \ln\left(\max_{l}(D_{.ul}, D_{.lu})\right) - \ln\left(\min_{l}(D_{.ul}, D_{.lu})\right), \tag{44}$$

where $D_{.uv}$ is the (u, v)th element of the test KLI matrix D and L represents the size of the attribute profile space. The two KLI-based indices were log-transformed to get a linear relationship with CCR (Henson et al., 2008; Henson et al., 2018).

The index $mean\ KLI(\boldsymbol{\alpha}_u)$ describes the average discrimination power of a test to differentiate $\boldsymbol{\alpha}_u$ from other attribute profiles. It is supposed to be positively correlated with the conditional CCR for $\boldsymbol{\alpha}_u$. However, this index alone is not sufficient for predicting CCR due to the multidimensional nature of the CDMs. When the $mean\ KLI_j(\boldsymbol{\alpha}_u)$ is fixed, if the test does not differentiate well between two particular attribute profiles $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$, the CCR for $\boldsymbol{\alpha}_u$ or $\boldsymbol{\alpha}_v$ suffers (Cheng, 2010). This phenomenon was mentioned in Cheng (2010)'s CD-CAT study and compared to Liebig's "law of the minimum," or Liebig's barrel. Therefore, a second index $range\ KLI(\boldsymbol{\alpha}_u)$ was defined in (44) to characterize the weakest point of a test. One particularly

low KLI between two $\boldsymbol{\alpha}$s leads to a relatively large range given the same $mean\ KLI_j(\boldsymbol{\alpha}_u)$. A range measure was used instead of the minimum measure to control the effect of $mean\ KLI(\boldsymbol{\alpha}_u)$. The index $range\ KLI(\boldsymbol{\alpha}_u)$ is negatively correlated with the conditional CCR for $\boldsymbol{\alpha}_u$ but has a low or insignificant correlation with $mean\ KLI(\boldsymbol{\alpha}_u)$.

The need for the second index is best illustrated by comparing the following two Q-matrices under the DINA model. Three independent attributes are measured with nine items.

$$Q_1 = \begin{bmatrix} I \\ I \\ I \end{bmatrix}, \qquad Q_2 = \begin{bmatrix} & I & \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$

where I is the identity matrix.

Assuming the intercept $\phi_0 = 0.1$ and the interaction effect $\phi_{all} = 0.8$ for all items, the two indices were calculated for the two tests. The CCRs were also obtained from the simulation. In the Table 15, the two tests have the same $mean\ KLI$ for each attribute profile but the second test has substantially lower CCRs.

Table 15: KLI indices and the CCRs for two Q-matrices

	Q_1			Q_2		
$\boldsymbol{\alpha}$	$mean\ KLI$	$range\ KLI$	CCR	$mean\ KLI$	$range\ KLI$	CCR
000	2.20	1.10	0.92	2.20	2.20	0.81
100	2.20	1.10	0.92	2.20	2.20	0.81
010	2.20	1.10	0.92	2.20	2.20	0.81
001	2.20	1.10	0.92	2.20	2.20	0.81
110	2.20	1.10	0.92	2.20	2.20	0.80
101	2.20	1.10	0.91	2.20	2.20	0.80
011	2.20	1.10	0.91	2.20	2.20	0.81
111	2.20	1.10	0.91	2.20	2.20	0.81

The difference between the two Q-matrices in Table 15 was referred to as an issue of content balancing in Cheng (2010) since the number of items for each attribute is not balanced in Q_2. Given the same $mean\ KLI$, the second index $range\ KLI$ is needed in this case to account for the different CCRs. A larger range index corresponds to a lower CCR.

The two conditional KL indices would be good predictors of the conditional CCR of $\boldsymbol{\alpha}$ with a fixed test length. To make them useful for between various test lengths, the following two conditions need to be satisfied:

1. For each $\boldsymbol{\alpha}$, there are no zero off-diagonal entries in the test KLI matrix D because $\ln(0)$ is not defined;
2. There is an odd number of items in each item type (i.e., a distinct q-vector).

The first condition is satisfied when the Q-matrix is complete. The second condition is necessary for the indices to be useful because when the examinee correctly respond to half of the items, the examinee is likely to be misclassified. For example, if the test has two items with $\boldsymbol{q} = (1\ 0\ 0)$, for examinees who master attribute α_1, the likelihood function is $L_1 = \prod_{j=1}^{2}(\phi_{j0} + \phi_{j,all})^{x_j}(1 - \phi_{j0} - \phi_{j,all})^{1-x_j}$; for examinees who do not possess attribute α_1, the likelihood function is $L_0 = \prod_{j=1}^{2}(\phi_{j0})^{x_j}(1 - \phi_{j0})^{1-x_j}$. It is possible that an examinee correctly answers item 1 but fails at item 2. Then $L_1 = (\phi_{10} + \phi_{1.all})(1 - \phi_{20} - \phi_{2,all})$, $L_0 = \phi_{20}(1 - \phi_{20})$. When the items are homogenous in quality, the difference between L_1 and L_0 would be very small. In an extreme case when $\phi_{j0} = 0.1$and $\phi_{j,all} = 0.8$ for all items, $L_1 = L_2 = 0.1 \times 0.9$.

KLI-based item selection in CD-CAT uses indices similar to $mean\ KLI(\boldsymbol{\alpha}_u)$ and ignores the minimum effect. As a result, researchers found it necessary to add extra constraints to the item selection algorithm in order to improve the CCR (Cheng, 2010). Such constraints intend to balance attribute coverage, and this process is also referred to as content balancing (Cheng, 2010). The

result of content balancing is a smaller KLI range given the same $mean\ KLI(\boldsymbol{\alpha}_u)$. When attribute hierarchies are present, content balancing becomes tricky. Using the two indices together in test construction becomes more practical with hierarchical attributes than content balancing.

4.3 Simulation design

A simulation study was conducted to assess the performance of the two indices. Random tests were generated as described below with items calibrated using DINA or A-CDM. The hierarchical GDINA model is equivalent to A-CDM in most cases, so the GDINA model is excluded from the simulations. The attribute hierarchies shown in 3.2 were used to simulate the examinee responses. The assessment tasks may be embedded in the classroom instruction and scattered in multiple class sessions. As a result, the assessment is not necessarily a concise one. We consider test lengths of three, five, and seven times the number of attributes, respectively.

For each combination of test length (e.g., nine items) and attribute hierarchy (e.g., H3.2), three sets of tests were simulated. The first set of 25 tests consists of single-attribute items, the second set of 25 tests consists of q-vectors from the full Q-matrix calibrated by the DINA model, and the third set of 50 tests consists of q-vectors from the full Q-matrix calibrated by the ACDM. The actual Q-matrix for each random test was constructed by randomly sampling from all the possible q-vectors with replacement if the full Q-matrix approach is used or from the identity matrix if only single-attribute items are wanted. Each Q-matrix contained the identity matrix to ensure completeness. There was an odd number of items in each item type (i.e., a distinct q-vector).

For all items, the intercept parameter ($\phi_0 = P(X = 1|\boldsymbol{\alpha} = \mathbf{0})$) was generated from the uniform distribution $U(0.1, 0.4)$ and $P(X = 1|\boldsymbol{\alpha} = \mathbf{1})$ was generated from $U(0.6, 0.9)$.

A total of 5,000 examinees are simulated for each true attribute profile for each random test. Given each examinee's attribute profile, item scores are generated based on the chosen model.

A random $U(0,1)$ variable u is generated. The correct response probability $P(X_{ij}=1|\boldsymbol{\alpha})$ is compared with u to decide the response of examinee i to item j:

$$Y_{ij}=\begin{cases}1 \ \ if \ u \leq P(Y_{ij}=1|\boldsymbol{\alpha}) \\ 0 \ \ \text{otherwise}\end{cases}. \tag{45}$$

The two conditional indices were calculated for each attribute profile for each random test. Classifications were accomplished via MLE for independent attributes or restricted MLE for hierarchical attributes because the item parameters are known. Conditional profile-wise CCR were recorded for each $\boldsymbol{\alpha}$.

The index of means is supposed to be positively correlated with the CCR, and the index of range is supposed to be negatively correlated with the CCR. For each attribute hierarchy, a linear regression model with normal errors was fit using the two indices to predict the CCR:

$$\text{CCR}=\beta_1 range\ KLI(\boldsymbol{\alpha}_u)+\beta_2 mean\ KLI(\boldsymbol{\alpha}_u)+\epsilon \tag{46}$$

The regression estimates were used to produce a linear combination of the two indices as a combined index, cKLI:

$$cKLI=\widehat{\beta_1} range\ KLI(\boldsymbol{\alpha}_u)+\widehat{\beta_2} mean\ KLI(\boldsymbol{\alpha}_u) \tag{47}$$

The combined index cKLI is expected to be highly correlated with the CCR.

4.4 Simulation results

The regression estimates and the R^2 for each attribute hierarchy were summarized in Table 16. A combined index was calculated as a linear combination of the two indices using the regression estimates as weights. This combined index (cKLI) was plotted against the CCR conditional on $\boldsymbol{\alpha}$ in the following scatter plots to visualize the relationships (see Figure 10-Figure 24). For brevity, we only present the scatter plots for a subset of $\boldsymbol{\alpha}$s when there are more than $K+1$ attribute profiles in the space.

Table 16: Regression estimates and R^2 for each attribute hierarchy

Attribute hierarchy		$range\ KLI(\boldsymbol{\alpha}_u)$	$mean\ KLI(\boldsymbol{\alpha}_u)$	R^2
H3.1	Independent	-0.07	0.24	0.76
H3.2	Linear	-0.07	0.19	0.78
H3.3	Inverted pyramid	-0.07	0.20	0.74
H3.4	Pyramid	-0.06	0.21	0.79
H4.1	Independent	-0.08	0.27	0.82
H4.2	Linear	-0.08	0.21	0.80
H4.3	Linear+single	-0.08	0.24	0.80
H4.4	Inverted pyramid	-0.08	0.22	0.81
H4.5	Pyramid	-0.07	0.22	0.81
H5.1	Independent	-0.07	0.28	0.81
H5.2	Linear	-0.09	0.21	0.82
H5.3	Inverted pyramid I	-0.08	0.26	0.82
H5.4	Inverted pyramid II	-0.08	0.26	0.81
H5.5	Pyramid I	-0.08	0.25	0.80
H5.6	Pyramid II	-0.08	0.25	0.81

Table 17: The overall correlation and the correlations for different test lengths between cKLI and the CCR

Attribute hierarchy		All	Test length		
			$3 \times K$	$5 \times K$	$7 \times K$
H3.1	Independent	0.87	0.60	0.76	0.85
H3.2	Linear	0.88	0.83	0.87	0.87
H3.3	Inverted pyramid	0.86	0.79	0.82	0.84
H3.4	Pyramid	0.89	0.81	0.88	0.86
H4.1	Independent	0.90	0.77	0.82	0.88
H4.2	Linear	0.89	0.86	0.86	0.86
H4.3	Linear+single	0.90	0.81	0.84	0.85
H4.4	Inverted pyramid	0.90	0.83	0.87	0.89
H4.5	Pyramid	0.90	0.84	0.87	0.87
H5.1	Independent	0.90	0.77	0.87	0.85
H5.2	Linear	0.90	0.85	0.91	0.87
H5.3	Inverted pyramid I	0.91	0.84	0.88	0.91
H5.4	Inverted pyramid II	0.90	0.81	0.88	0.89
H5.5	Pyramid I	0.90	0.80	0.87	0.90
H5.6	Pyramid II	0.90	0.84	0.87	0.88

Note: K is the number of attributes.

The overall correlation between cKLI and the CCR is presented in Table 17. All the overall correlations are around 0.9. The correlations for different test lengths are also calculated (Table

17). The correlation generally increases substantially as the test length goes up from three times of K to five or seven times of K where K is the number of attributes. This trend can also be seen in the scatter plots.

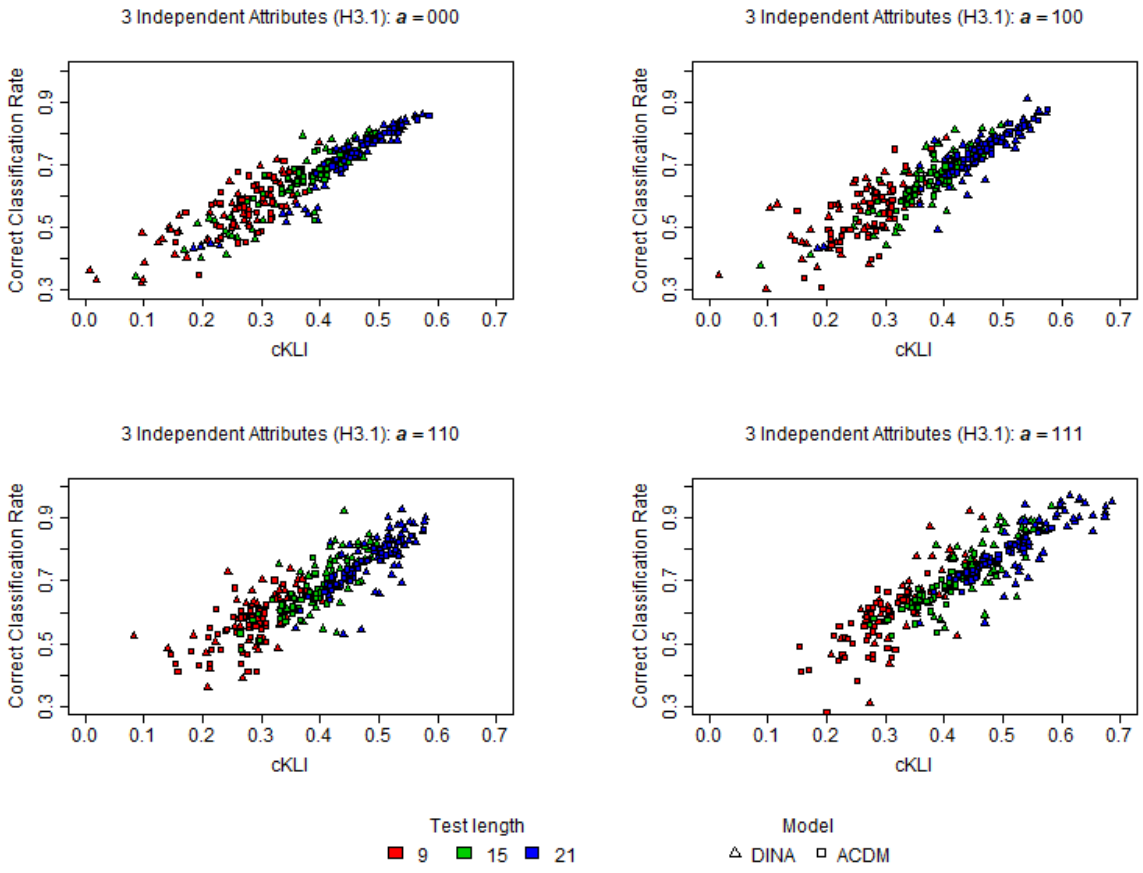


Figure 10: A plot for tests with three independent attributes (H3.1) of the combined index with CCRs

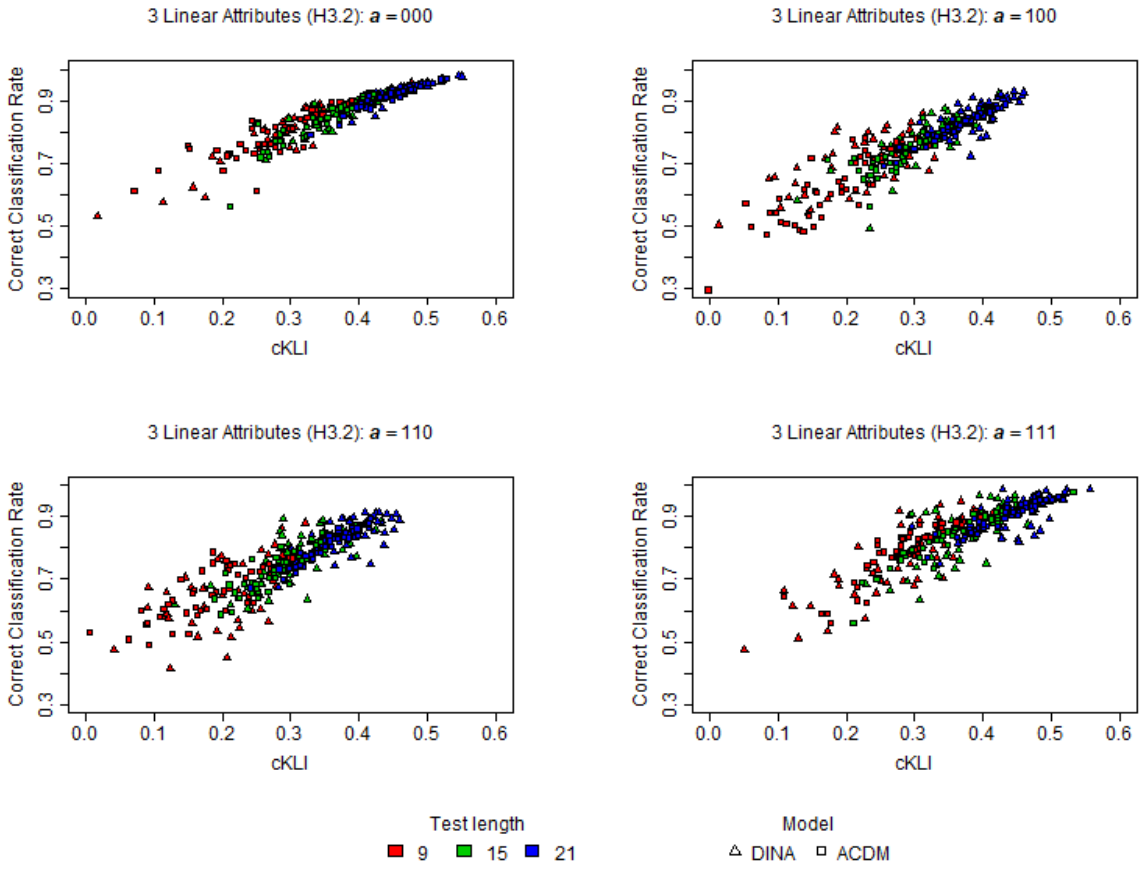


Figure 11: A plot for tests with three linear attributes (H3.2) of the combined index with CCRs

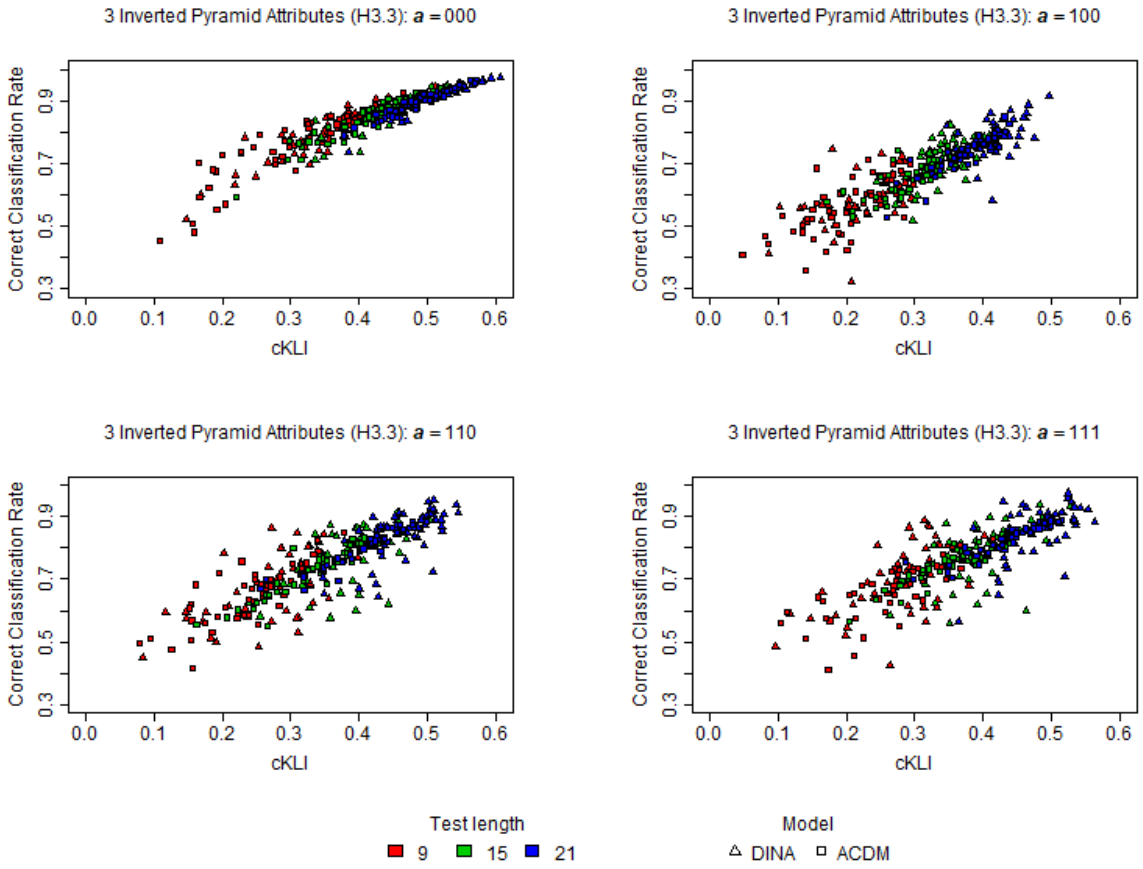


Figure 12: A plot for tests with three inverted pyramid attributes (H3.3) of the combined index with CCRs

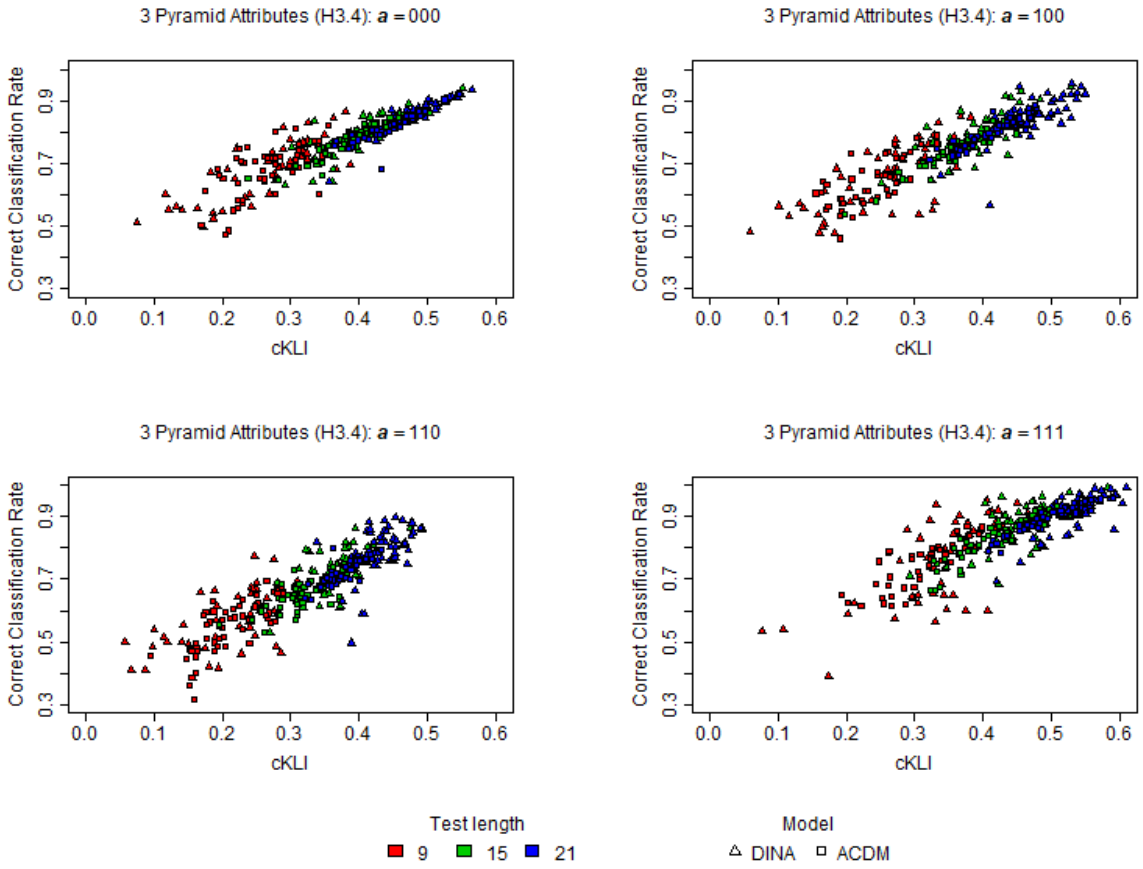


Figure 13: A plot for tests with three pyramid attributes (H3.4) of the combined index with CCRs

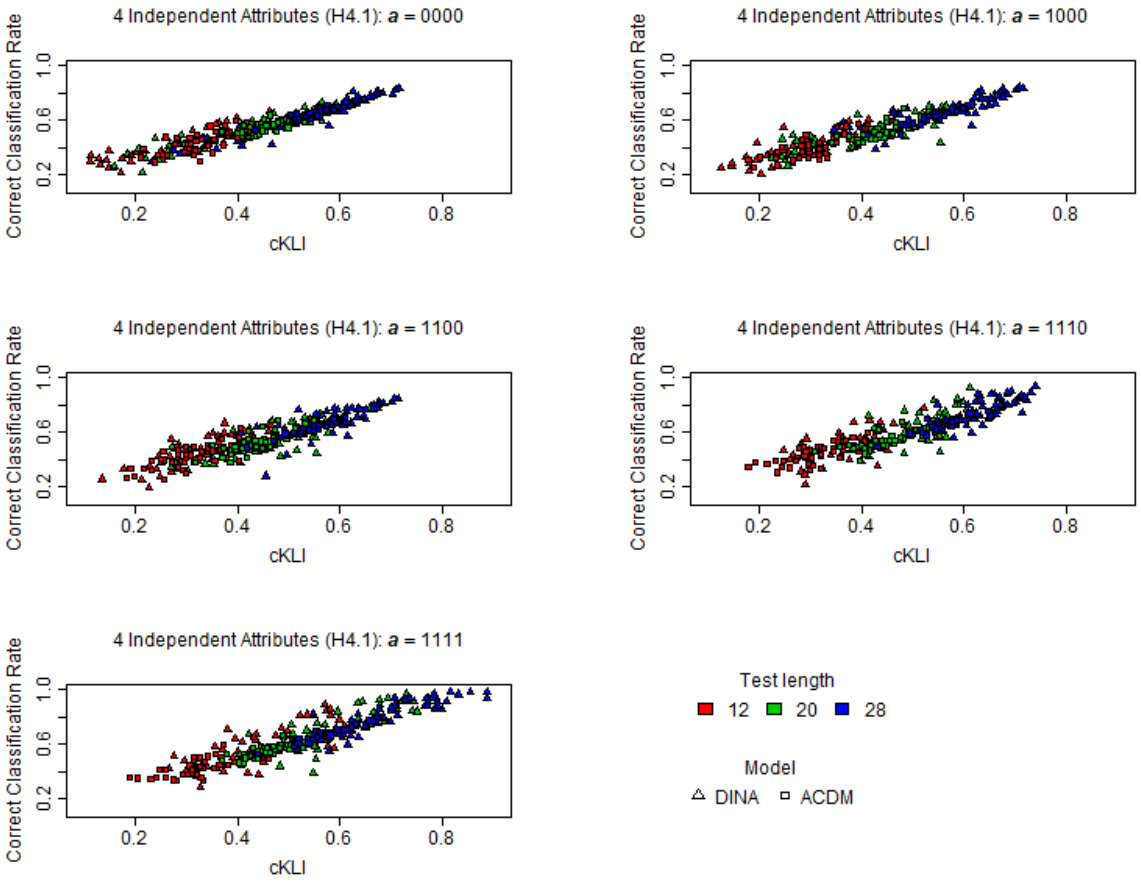


Figure 14: A plot for tests with four independent attributes (H4.1) of the combined index with CCRs

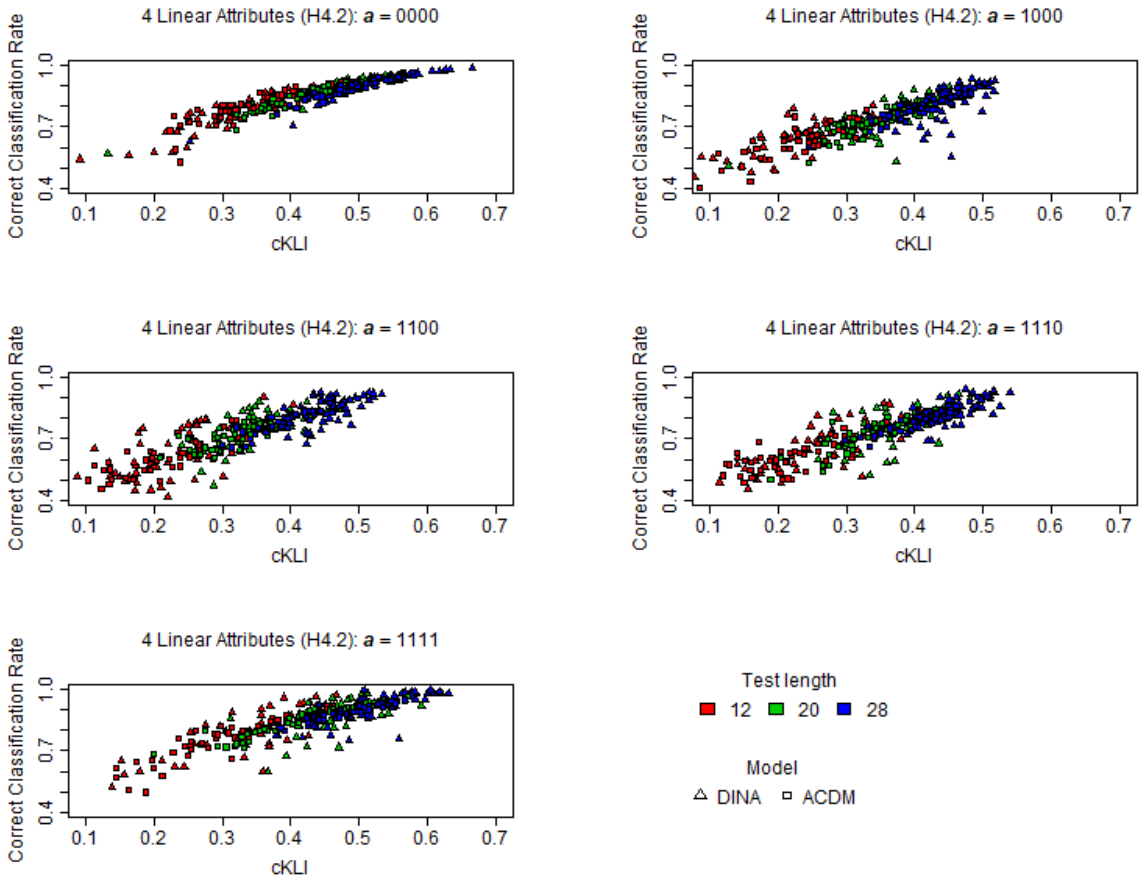


Figure 15: A plot for tests with four linear attributes (H4.2) of the combined index with CCRs

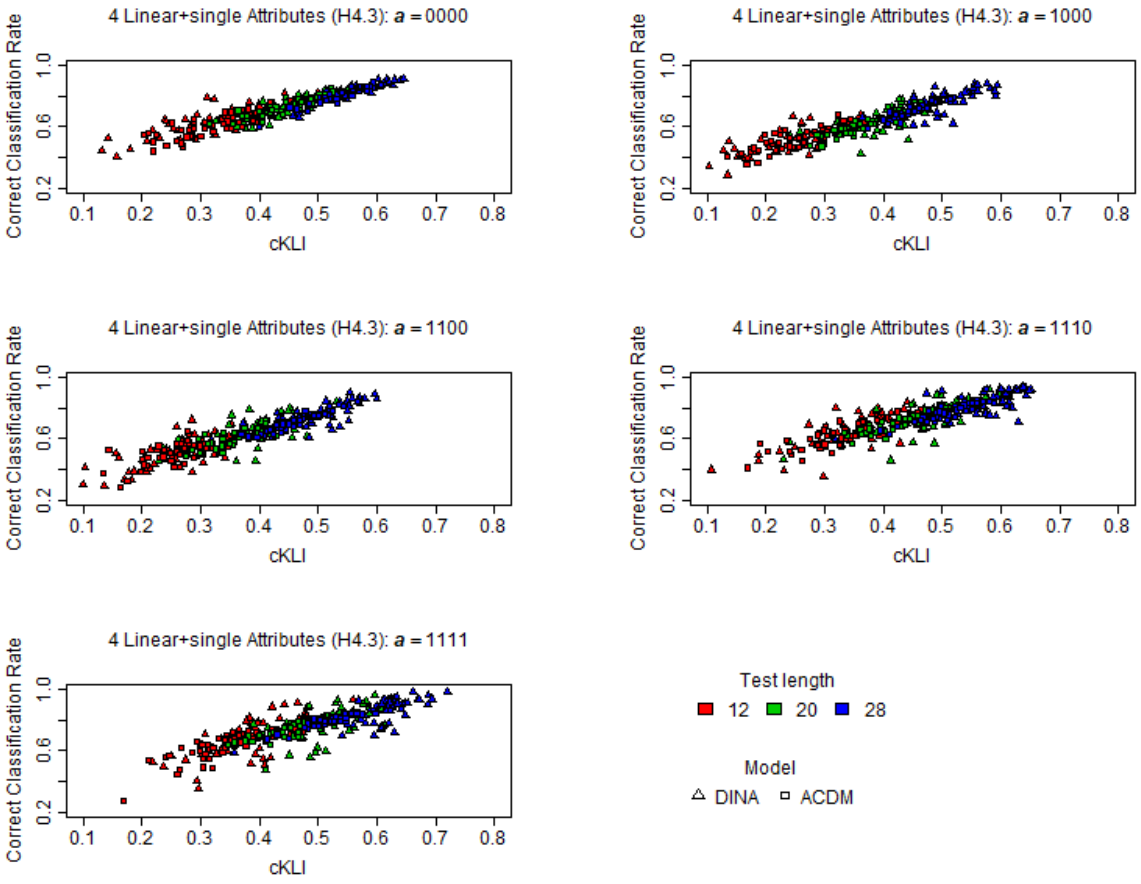


Figure 16: A plot for tests with three linear attributes + one single attribute (H4.3) of the combined index with CCRs

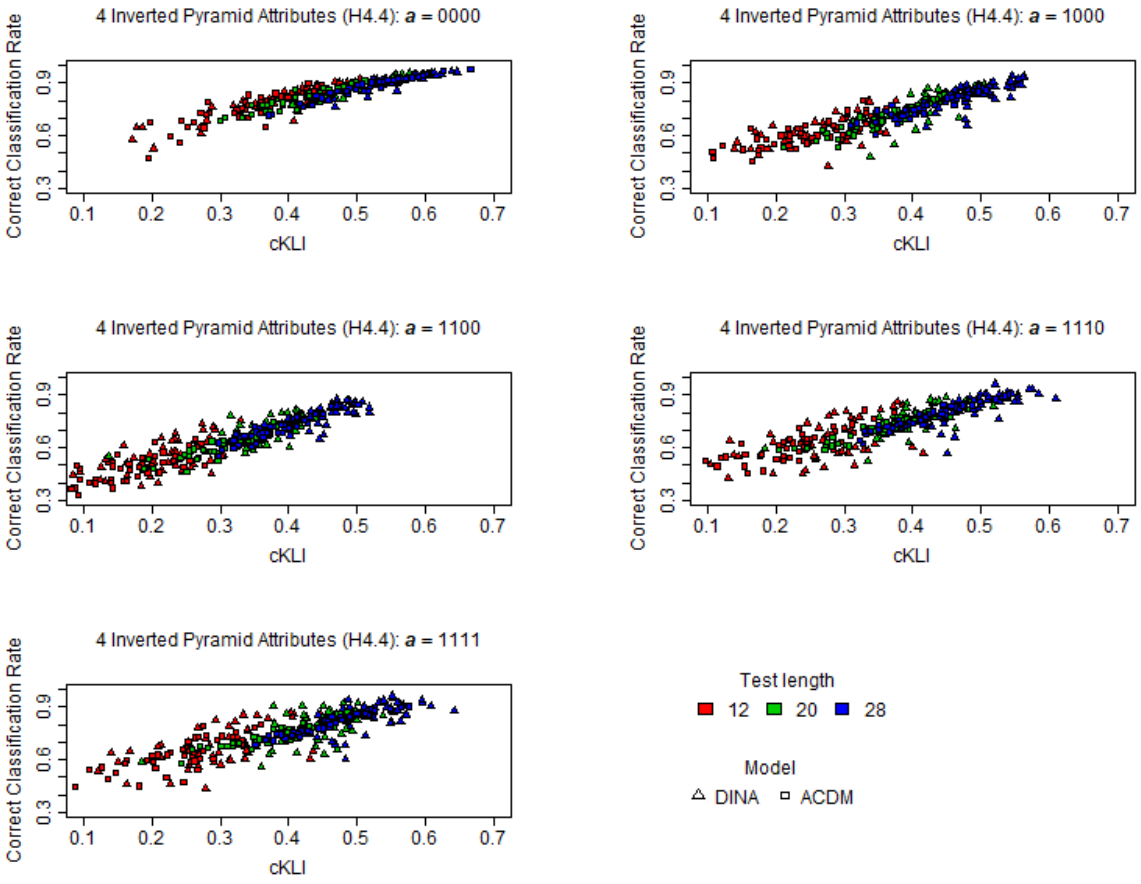


Figure 17: A plot for tests with four inverted pyramid attributes (H4.4) of the combined index with CCRs

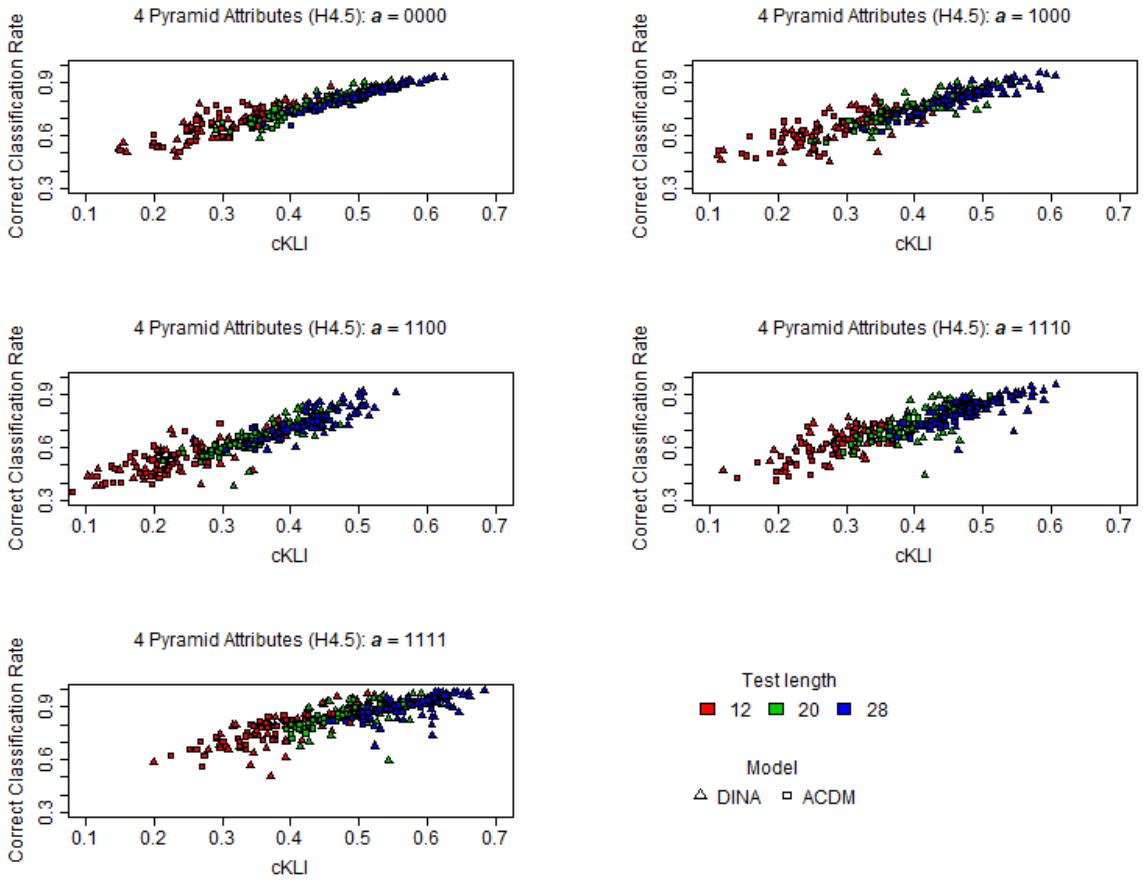


Figure 18: A plot for tests with four pyramid attributes (H4.5) of the combined index with CCRs

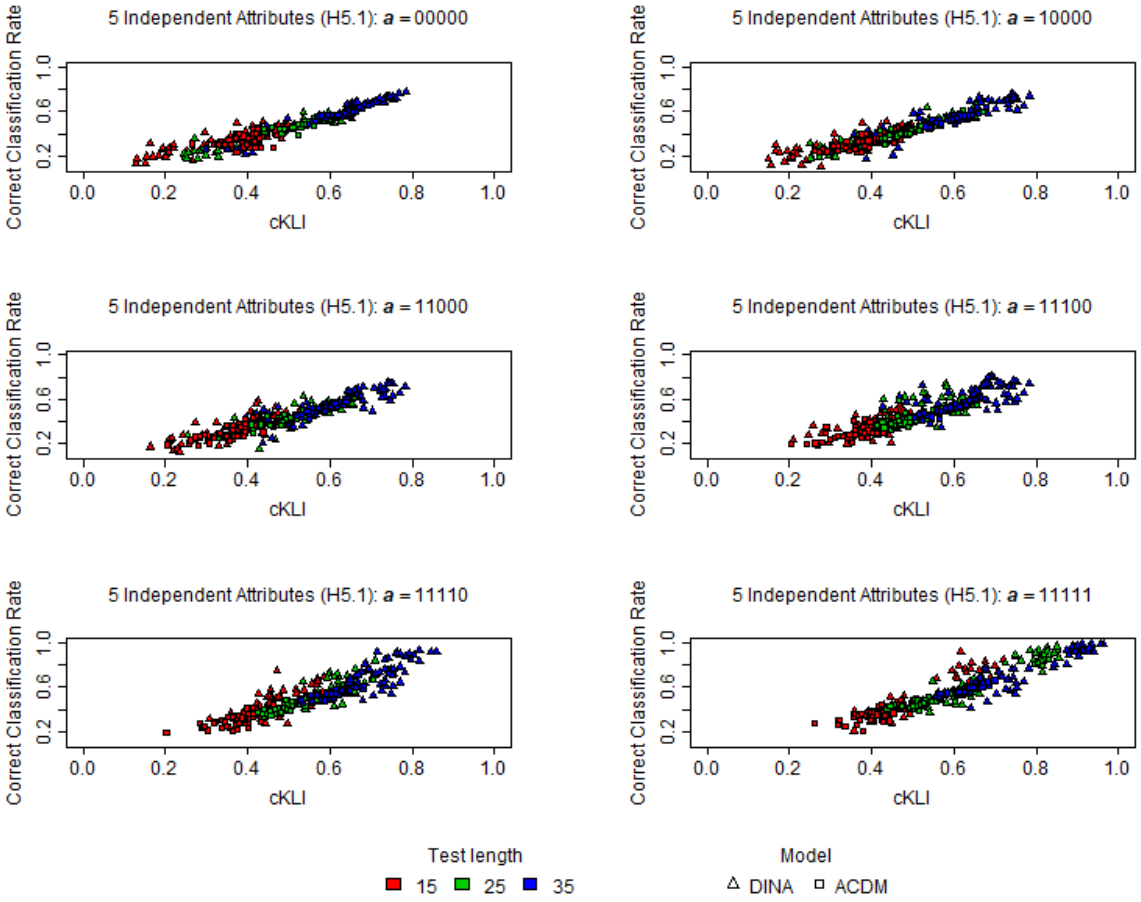


Figure 19: A plot for tests with five independent attributes (H5.1) of the combined index with CCRs

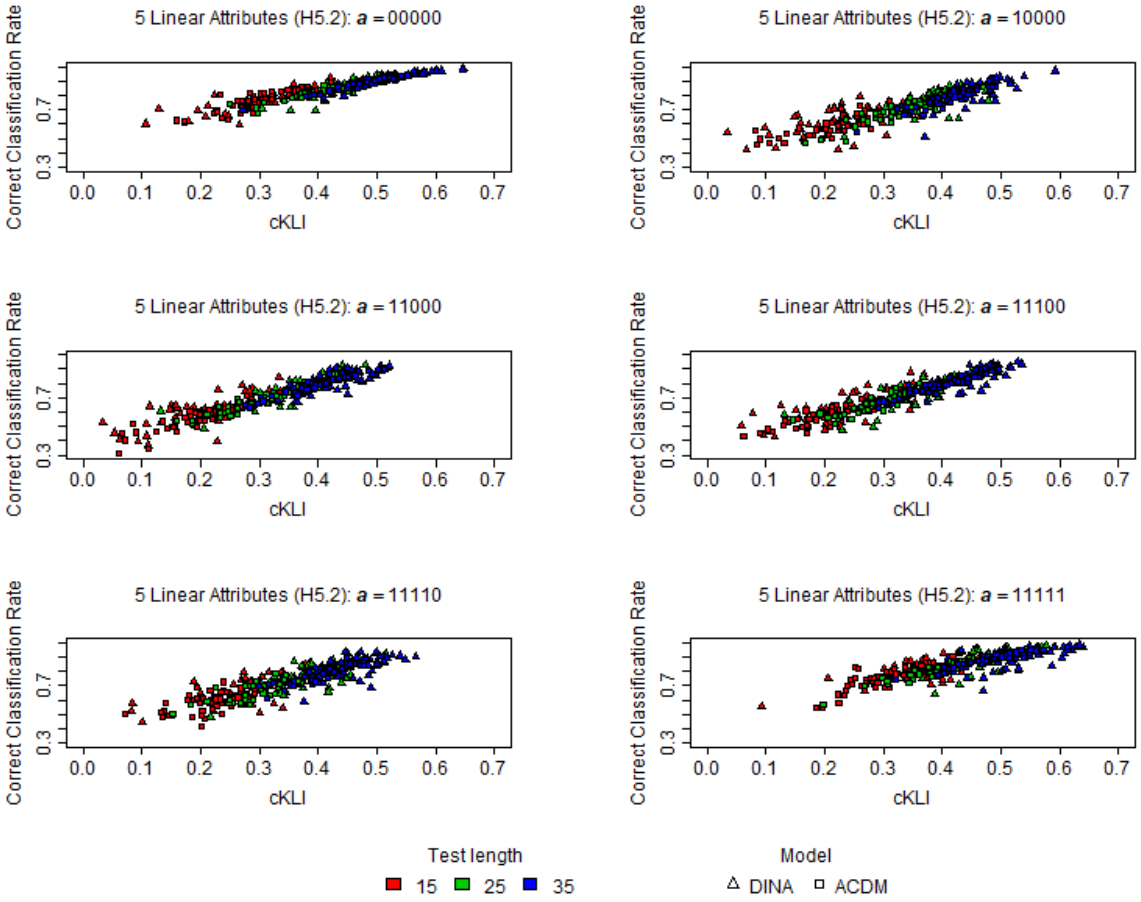


Figure 20: A plot for tests with five linear attributes (H5.2) of the combined index with CCRs

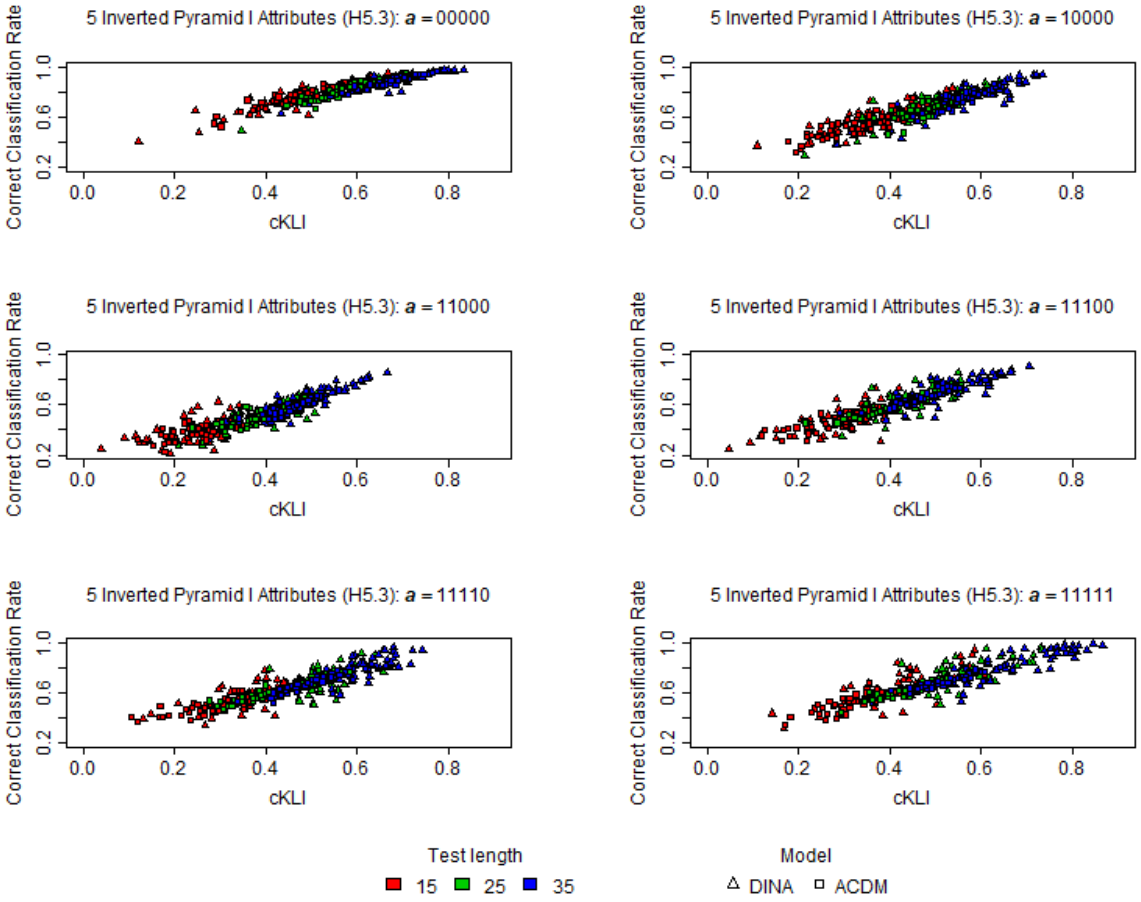


Figure 21: A plot for tests with five inverted pyramid attributes (H5.3) of the combined index with CCRs

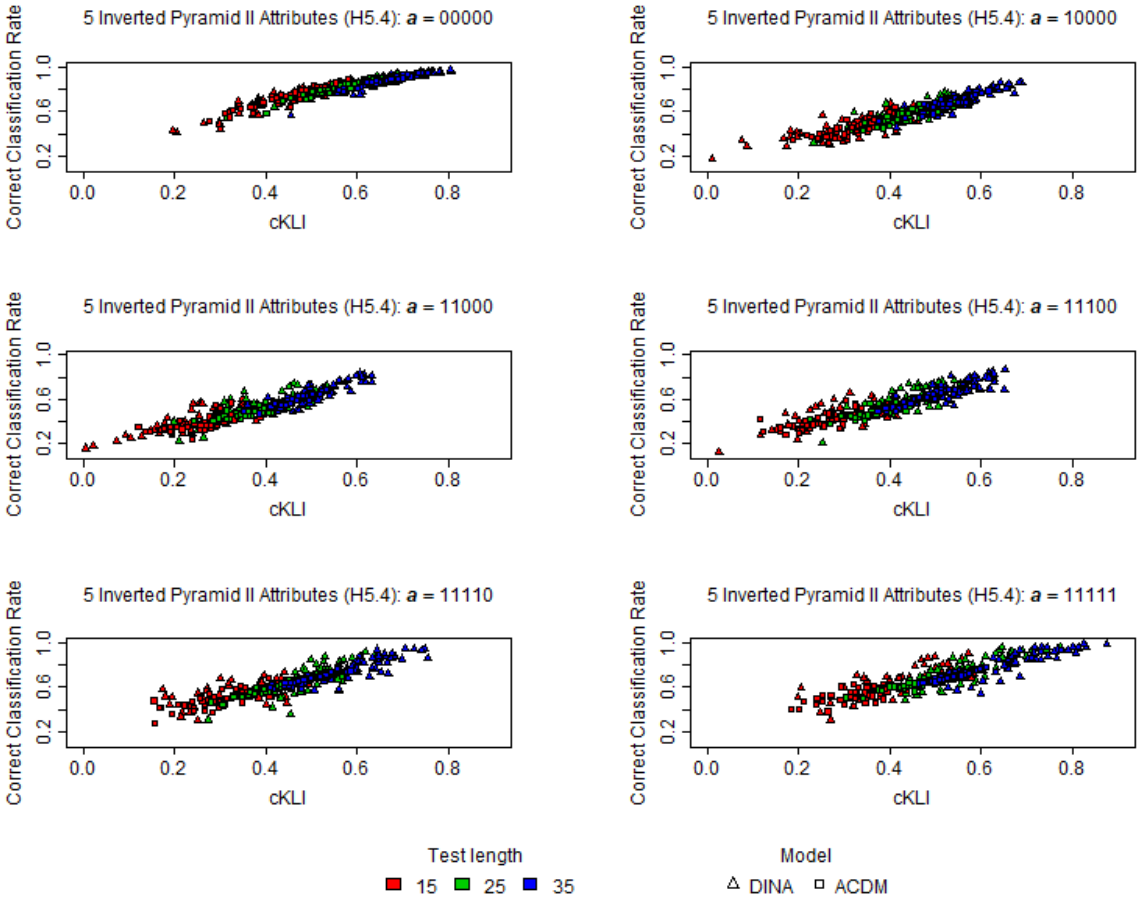


Figure 22: A plot for tests with five inverted pyramid attributes (H5.4) of the combined index with CCRs

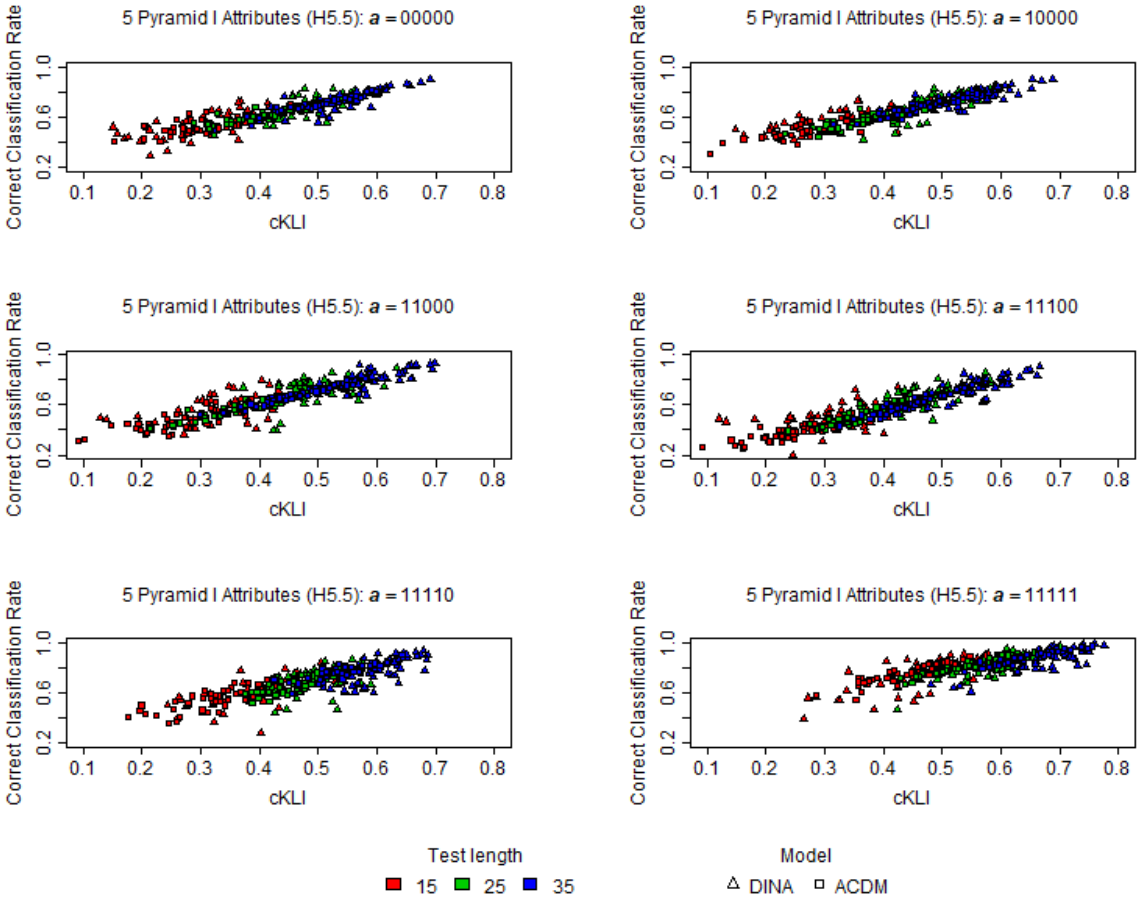


Figure 23: A plot for tests with five pyramid attributes (H5.5) of the combined index with CCRs

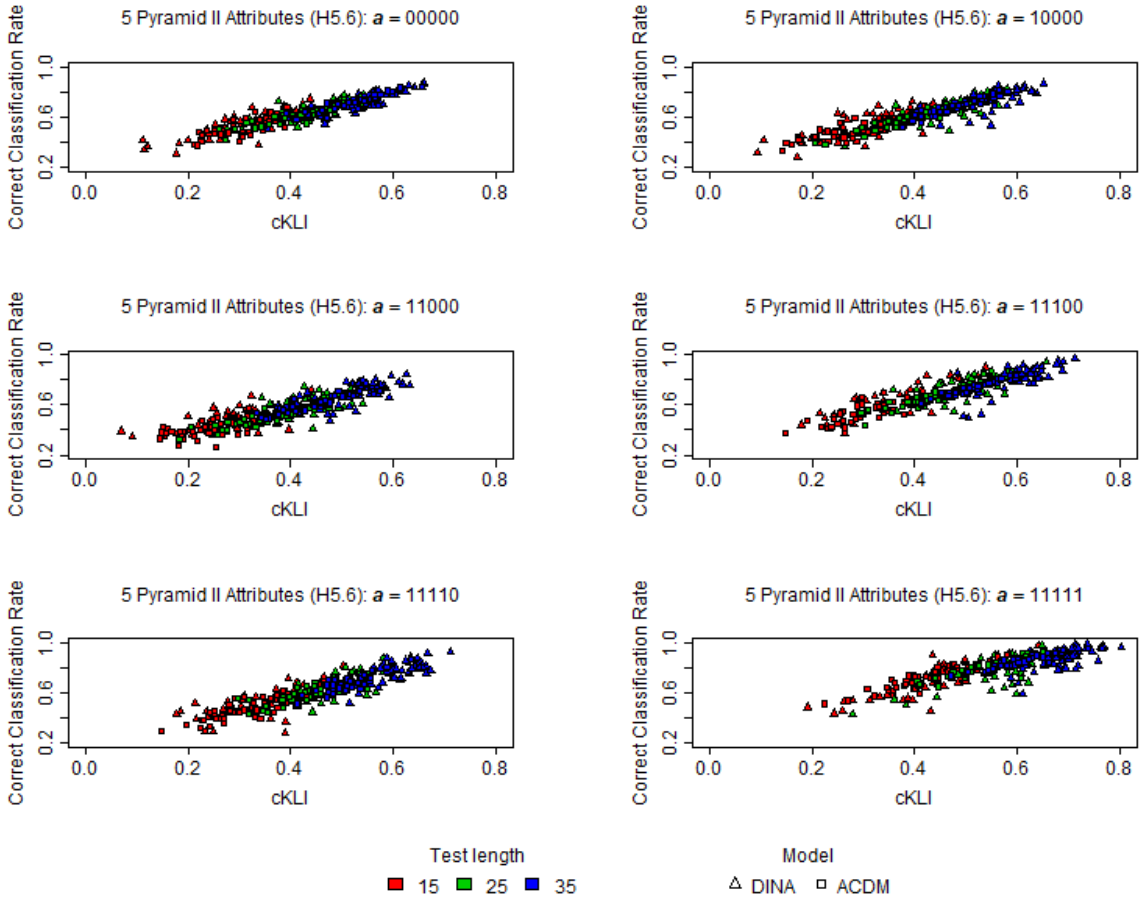


Figure 24: A plot for tests with five pyramid attributes (H5.6) of the combined index with CCRs

4.5 Discussion

The two indices can predict the CCR well according to the linear regression results showing that abut 80% of the total variance was explained. The prediction of two indices was a substantial improve from the prediction of either index alone. This relationship was also reflected by the high correlation between the combined index *cKLI* and the CCR. These results suggest that using an averaged KLI may not be sufficient for predicting CCR. Therefore, any single index based on the maximum or the mean of KLI would have serious limitations as a test construction index. As mentioned earlier, it has been found necessary to add extra constraints to the item selection algorithm based on a single KLI index, in order to improve the CCR in CD-CAT research (Cheng, 2010). Such constraints would lead to a decreased range index, and balanced attribute coverage

would be observed with independent attributes. In other words, content balancing could have the same effect as having a range index when attributes are independent. With hierarchical attributes, however, there is no clear way to define content balancing. Therefore, using the two indices together in test construction would be more appropriate with hierarchical attributes than content balancing. This applies to both non-adaptive and adaptive test construction.

It is important to note that the ($cKLI$, CCR) relationship does not depend on the model selected (DINA or ACDM) or test length. However, the relationship between the two indices and the CCR may depend on the attribute hierarchy, more specifically, the number of attribute profiles as suggested by the different regression estimates in Table 15. Moreover, the indices lead to better predictions of the CCR as the test length increases.

The proposed index can be used to assemble tests from an item pool by setting an information target or a fixed test length. Setting an appropriate information target may not be easy because on the one hand, a target needs to be set for each attribute profile and on the other hand, also noted in Henson et al. (2018), the threshold value that would ensure a certain CCR may depend on the number of attributes and the attribute hierarchy.

If the test length is fixed, the test assembly algorithm could take two steps: a set of tests with largest mean KLI is identified first, and then the one with the largest minimum KLI or smallest range index is chosen. Alternatively, the regression estimates in Table 16 an be used to calculate the combined index. The test assembly can be automated in various ways. With the two information indices, we do not need extra constraints like "each cognitive attribute is measured by an adequate number of items (Cheng, 2010, p. 903)."

The ($cKLI$, CCR) relationship is visualized for each attribute profile in Figure 11 - Figure 24 because the CCR could vary substantially between $\boldsymbol{\alpha}$s. We chose four random tests in the condition H4.2 to demonstrate the variation of CCR in Figure 25.

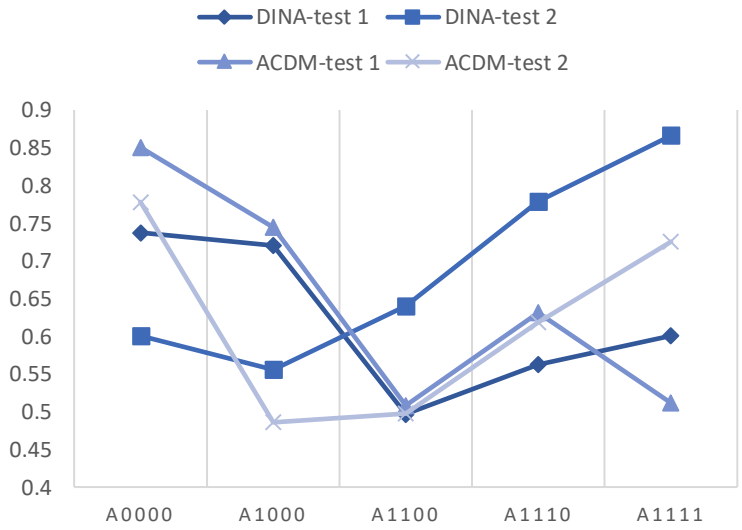


Figure 25: The conditional CCRs from four random tests in H4.2

With a linear hierarchy and an identity matrix as the Q-matrix, the attribute profiles that master some but not all attributes are easier to be misclassified than the two attribute profiles on the two ends (i.e., the one with all 0s and the one with all 1s). This pattern can also be explained in terms of the KL indices (Table 18). Another way to see the various CCRs for a linear hierarchy is the item with $\boldsymbol{q} = (1\ 0\ 0)$ differentiates $\boldsymbol{\alpha} = (0\ 0\ 0)$ with other $\boldsymbol{\alpha}$**s** and the item with $\boldsymbol{q} = (0\ 0\ 1)$ differentiates $\boldsymbol{\alpha} = (1\ 1\ 1)$ with other $\boldsymbol{\alpha}$**s**, and as a result these two $\boldsymbol{\alpha}$**s** have a higher classification rates than the $\boldsymbol{\alpha}$**s** in the middle.

We use the two KLI-based indices to compare the full and reduced Q-matrix approaches under the ACDM. As mentioned earlier, the two approaches are equivalent under the ACDM. The major difference between a full Q-matrix with ACDM and a reduced Q-matrix with ACDM is the exclusion of some single-attribute items from the reduced Q-matrices. Therefore, we compare the

identity matrix with Q_r with ACDM in terms of the two indices for a linear hierarchy of three attributes (H3.2). The item parameters are presented in Table 18. The indices for the two three-item tests are shown in Table 19.

If the reduced Q-matrix approach is adopted and all the items are calibrated with the ACDM, classifications for the attribute profiles $\boldsymbol{\alpha} = (100)$, $\boldsymbol{\alpha} = (110)$, and $\boldsymbol{\alpha} = (111)$ become much more difficult. As suggested by the combined index, much longer tests are required to achieve comparable classification rates for most of the attribute profiles if two types of single-attribute items, $\boldsymbol{q} = (0\ 1\ 0)$ and $\boldsymbol{q} = (0\ 0\ 1)$, are excluded from the candidate pool.

In addition to the consideration of classification efficiency, the choice between a reduced Q-matrix and a full Q-matrix should depend on answers to questions such as whether it is possible to a mixed item pool, whether it is possible to develop a certain item type, and the model-data fit at the item level.

Table 18: Item parameters of five items for H3.2

	$\boldsymbol{q}$	Model	ϕ_{j0}	ϕ_{j1}	ϕ_{j2}	ϕ_{j3}
	(010)	-	0.1		0.8	
	(001)	-	0.1			0.8
	(100)	-	0.1	0.8		
$Q_r + ACDM$	(110)	ACDM	0.1	0.4	0.4	
	(111)	ACDM	0.1	0.27	0.27	0.27

Table 19: Comparison between two three-item tests in terms of the two indices

$\boldsymbol{\alpha}$	I_3			$Q_r + ACDM$		
	mean KLI	*range KLI*	cKLI	*mean KLI*	*range KLI*	cKLI
000	1.26	1.10	0.05	1.38	0.82	0.12
100	0.85	0.69	0.04	0.33	1.35	-0.17
110	0.85	0.69	0.04	0.52	2.84	-0.38
111	1.26	1.10	0.05	0.80	3.34	-0.42

Chapter 5 Q-matrix design for nonparametric classifications with hierarchical attributes

5.1 Introduction

Without a calibrated item pool, the nonparametric classification (NPC) method (Chiu, Sun, & Bian, 2018) provide an alternative approach for classifications. The NPC method allows the teachers to develop their own items based on CDMs if they can identify the attribute hierarchy and the Q-matrix. There is no need for item calibration, and students are classified based on their response data without the need to estimate item parameters.

The Q-matrix design plays an even more important role in nonparametric classifications than in parametric classifications, but it has not been formally addressed in the literature. Related studies explore different Q-matrix designs with hierarchical attributes in the context of parametric classifications (Liu, Huggins-Manley, & Bradshaw 2017; Tu, Wang, Cai, Douglas, & Chang, 2018). There is a consensus on the effect of single-structured items on accurate classifications regardless of the attribute hierarchy (Chiu et al., 2009; DeCarlo, 2011; Madison & Bradshaw, 2015). However, the role of items with multiple attributes is not clear. Other factors in Q-matrix design that receive less attention in existing research include test length and the number of items in each item type.

In this study, the NPC method (Chiu & Douglas, 2013) was used Because it is assumed that the teacher develops a CDM-based test for a particular classroom. Prior data are not expected to be available. Therefore, the general nonparametric classification method (Chiu, Sun, & Bian, 2018) that requires some prior response data is not considered.

5.2 Ties in NPC

There is a tie when the observed response pattern of an examinee is at an equal distance to more than one ideal response pattern. Some Q-matrices lead to more ties than others. With an ideal Q-matrix, the item responses of high probabilities are always closest to the ideal response pattern of the true $\boldsymbol{\alpha}$, and there would be no ties in the hamming distance. In this study, if a tie occurs, the examinee would be randomly classified into one attribute profile with the minimal hamming distance.

We present a comparison between two Q-matrices as an example. The underlying model is the DINA model. The item quality is assumed to be high: $1 - P(X = 1|\boldsymbol{\alpha} = \mathbf{1}) = P(X = 1|\boldsymbol{\alpha} = \mathbf{0}) = 0.1$. Three independent attributes are involved. We focus on an examinee with $\boldsymbol{\alpha} = (1\ 1\ 1)$. The hamming distances between several likely response patterns of $\boldsymbol{\alpha}_{111}$ and each of the ideal response patterns are shown in the cells of Table 20 and Table 21. With an identity matrix as the Q-matrix, there are no ties in the hamming distance and the probability of correctly classifying the examinee with $\boldsymbol{\alpha} = (1\ 1\ 1)$ equals to the probability of observing the response pattern of $(1\ 1\ 1)$, which is 0.9^3.

When the Q-matrix contains the identity matrix I_3 and an item probing all three attributes, ties are observed when the examinee slips on one of the items (Table 21). The probability of a tie is the probability of observing such a response patter, which is $0.9^3 \times 0.1 \times 4 = 0.29$. It is still possible to clarify the examinee with a tie in the hamming distance. The CCR for $\boldsymbol{\alpha}_{111}$ can be calculated as a weighted sum of probabilities: $0.9^4 + 0.9^3 \times 0.1 \times 0.25 + 0.9^3 \times 0.1 \times 0.33 + 0.9^3 \times 0.1 \times 0.5 + 0.9^3 \times 0.1 \times 0.5 = 0.77$. Comparing the two Q-matrices reveals that adding an item with $\boldsymbol{q} = (1\ 1\ 1)$ to the identity matrix leads to a slight increase in the CCR for $\boldsymbol{\alpha}_{111}$ from 0.73 to 0.77. The second Q-matrix leads to a probability of 0.29 to obtain a tie.

Table 20: Hamming distances for $\boldsymbol{\alpha}_{111}$ with $Q = I_3$ (H3.1)

Response pattern $\boldsymbol{X}$	Probability $Pr(\boldsymbol{X})$	$\boldsymbol{\alpha}$: (Ideal response pattern)							
		$\boldsymbol{\alpha}_{000}$: (000)	$\boldsymbol{\alpha}_{100}$: (100)	$\boldsymbol{\alpha}_{010}$: (010)	$\boldsymbol{\alpha}_{001}$: (001)	$\boldsymbol{\alpha}_{110}$: (110)	$\boldsymbol{\alpha}_{101}$: (101)	$\boldsymbol{\alpha}_{011}$: (011)	$\boldsymbol{\alpha}_{111}$: (111)
(111)	0.9^3	3	2	2	2	1	1	1	**0**
(110)	$0.9^2 \times 0.1$	2	1	1	3	**0**	2	2	1
(101)	$0.9^2 \times 0.1$	2	1	3	1	2	**0**	2	1
(011)	$0.9^2 \times 0.1$	2	3	1	1	2	2	**0**	1

Table 21: Hamming distances for $\boldsymbol{\alpha}_{111}$ with $Q = [I_3, q_{111}]^T$ (H3.1)

Response pattern $\boldsymbol{X}$	Probability $Pr(\boldsymbol{X})$	$\boldsymbol{\alpha}$: (Ideal response pattern)							
		$\boldsymbol{\alpha}_{000}$: (000)	$\boldsymbol{\alpha}_{100}$: (100)	$\boldsymbol{\alpha}_{010}$: (010)	$\boldsymbol{\alpha}_{001}$: (001)	$\boldsymbol{\alpha}_{110}$: (110)	$\boldsymbol{\alpha}_{101}$: (101)	$\boldsymbol{\alpha}_{011}$: (011)	$\boldsymbol{\alpha}_{111}$: (111)
(1111)	0.9^4	4	3	3	3	2	2	2	**0**
(1110)	$0.9^3 \times 0.1$	3	2	2	2	**1**	**1**	**1**	**1**
(1101)	$0.9^3 \times 0.1$	3	2	2	4	**1**	**1**	3	**1**
(1011)	$0.9^3 \times 0.1$	3	2	4	2	3	**1**	3	**1**
(0111)	$0.9^3 \times 0.1$	3	4	2	2	3	3	**1**	**1**

5.3 Simulation design

The identity matrix served as the baseline Q-matrix. We considered the following situations: 1) adding one or two simple-attribute items to the baseline, 2) adding one or two multiple-attribute items to be baseline, and 3) adding an identity matrix. A total of 15, 19, and 23 Q-matrices are obtained for $K = 3$, 4, and 5, respectively, presented in Table 22.

The computations of CCRs and the probability of a tie become more complicated with a longer test or more attributes. Therefore, a simulation study was conducted to compare Q-matrices. Item parameters were simulated based on $1 - P(X = 1|\boldsymbol{\alpha} = \mathbf{1}) = P(X = 1|\boldsymbol{\alpha} = \mathbf{0}) = 0.1$. A total of 5,000 examinees are simulated for each true attribute profile for each Q-matrix. Given each examinee's attribute profile, item scores are generated based on the DINA. A random $U(0, 1)$ variable u is generated. The correct response probability $P(X_{ij} = 1|\boldsymbol{\alpha})$ is compared with u to decide the response of examinee i to item j:

$$Y_{ij} = \begin{cases} 1 & if\ u \leq P(Y_{ij} = 1|\boldsymbol{\alpha}) \\ 0 & \text{otherwise} \end{cases}. \tag{48}$$

Examinee responses were classified using the nonparametric classification method (Chiu & Douglas, 2013). Conditional profile-wise CCR were recorded for each $\boldsymbol{\alpha}$. The percent of ties was recorded for each simulation condition as an estimate of the probability of getting a tie.

Table 22: Q-matrix designs for the simulation study of nonparametric classifications

Q-matrix		Q-matrix		Q-matrix	
3-1	I_3	4-1	I_4	5-1	I_5
3-2	$[I_3, \boldsymbol{q}_{\{100\}}]^T$	4-2	$[I_4, \boldsymbol{q}_{\{1000\}}]^T$	5-2	$[I_5, \boldsymbol{q}_{\{10000\}}]^T$
3-3	$[I_3, \boldsymbol{q}_{\{110\}}]^T$	4-3	$[I_4, \boldsymbol{q}_{\{1100\}}]^T$	5-3	$[I_5, \boldsymbol{q}_{\{11000\}}]^T$
3-4	$[I_3, \boldsymbol{q}_{\{111\}}]^T$	4-4	$[I_4, \boldsymbol{q}_{\{1110\}}]^T$	5-4	$[I_5, \boldsymbol{q}_{\{11100\}}]^T$
3-5	$[I_3, \boldsymbol{q}_{\{100\}}, \boldsymbol{q}_{\{100\}}]^T$	4-5	$[I_4, \boldsymbol{q}_{\{1111\}}]^T$	5-5	$[I_5, \boldsymbol{q}_{\{11110\}}]^T$
3-6	$[I_3, \boldsymbol{q}_{\{110\}}, \boldsymbol{q}_{\{110\}}]^T$	4-6	$[I_4, \boldsymbol{q}_{\{1000\}}, \boldsymbol{q}_{\{1000\}}]^T$	5-6	$[I_5, \boldsymbol{q}_{\{11111\}}]^T$
3-7	$[I_3, \boldsymbol{q}_{\{111\}}, \boldsymbol{q}_{\{111\}}]^T$	4-7	$[I_4, \boldsymbol{q}_{\{1100\}}, \boldsymbol{q}_{\{1100\}}]^T$	5-7	$[I_5, \boldsymbol{q}_{\{10000\}}, \boldsymbol{q}_{\{10000\}}]^T$
3-8	$[I_3, I_3]^T$	4-8	$[I_4, \boldsymbol{q}_{\{1110\}}, \boldsymbol{q}_{\{1110\}}]^T$	5-8	$[I_5, \boldsymbol{q}_{\{11000\}}, \boldsymbol{q}_{\{11000\}}]^T$
3-9	$[I_3, I_3, \boldsymbol{q}_{\{100\}}]^T$	4-9	$[I_4, q_{\{1111\}}, q_{\{1111\}}]^T$	5-9	$[I_5, \boldsymbol{q}_{\{11100\}}, \boldsymbol{q}_{\{11100\}}]^T$
3-10	$[I_3, I_3, \boldsymbol{q}_{\{110\}}]^T$	4-10	$[I_4, I_4]^T$	5-10	$[I_5, \boldsymbol{q}_{\{11110\}}, \boldsymbol{q}_{\{11110\}}]^T$
3-11	$[I_3, I_3, \boldsymbol{q}_{\{111\}}]^T$	4-11	$[I_4, I_4, \boldsymbol{q}_{\{1000\}}]^T$	5-11	$[I_5, \boldsymbol{q}_{\{11111\}}, \boldsymbol{q}_{\{11111\}}]^T$
3-12	$[I_3, I_3, \boldsymbol{q}_{\{100\}}, \boldsymbol{q}_{\{100\}}]^T$	4-12	$[I_4, I_4, \boldsymbol{q}_{\{1100\}}]^T$	5-12	$[I_5, I_5]^T$
3-13	$[I_3, I_3, \boldsymbol{q}_{\{110\}}, \boldsymbol{q}_{\{110\}}]^T$	4-13	$[I_4, I_4, \boldsymbol{q}_{\{1110\}}]^T$	5-13	$[I_5, I_5, \boldsymbol{q}_{\{10000\}}]^T$
3-14	$[I_3, I_3, \boldsymbol{q}_{\{111\}}, \boldsymbol{q}_{\{111\}}]^T$	4-14	$[I_4, I_4, \boldsymbol{q}_{\{1111\}}]^T$	5-14	$[I_5, I_5, \boldsymbol{q}_{\{11000\}}]^T$
3-15	$[I_3, I_3, I_3]^T$	4-15	$[I_4, I_4, \boldsymbol{q}_{\{1000\}}, \boldsymbol{q}_{\{1000\}}]^T$	5-15	$[I_5, I_5, \boldsymbol{q}_{\{11100\}}]^T$
		4-16	$[I_4, I_4, \boldsymbol{q}_{\{1100\}}, \boldsymbol{q}_{\{1100\}}]^T$	5-16	$[I_5, I_5, \boldsymbol{q}_{\{11110\}}]^T$
		4-17	$[I_4, I_4, \boldsymbol{q}_{\{1110\}}, \boldsymbol{q}_{\{1110\}}]^T$	5-17	$[I_5, I_5, \boldsymbol{q}_{\{11111\}}]^T$
		4-18	$[I_4, I_4, \boldsymbol{q}_{\{1111\}}, \boldsymbol{q}_{\{1111\}}]^T$	5-18	$[I_5, I_5, \boldsymbol{q}_{\{10000\}}, \boldsymbol{q}_{\{10000\}}]^T$
		4-19	$[I_4, I_4, I_4]^T$	5-19	$[I_5, I_5, \boldsymbol{q}_{\{11000\}}, \boldsymbol{q}_{\{11000\}}]^T$
				5-20	$[I_5, I_5, \boldsymbol{q}_{\{11100\}}, \boldsymbol{q}_{\{11100\}}]^T$
				5-21	$[I_5, I_5, \boldsymbol{q}_{\{11110\}}, \boldsymbol{q}_{\{11110\}}]^T$
				5-22	$[I_5, I_5, \boldsymbol{q}_{\{11111\}}, \boldsymbol{q}_{\{11111\}}]^T$
				5-23	$[I_5, I_5, I_5]^T$

5.4 Simulation results

Simulation results for the conditions with three attributes are summarized in Table 22-Table 25. For brevity, we only present the results for four attribute profiles. Comparing each Q-matrix to the baseline (Q3-1), we found that a very high probability of obtaining a tie usually suggests no increase in the CCR and a lack of ties suggests an increased CCR for some $\boldsymbol{\alpha}$s**.** A longer test does not necessarily lead to higher CCR for each attribute profile.

As shown in Table 22, adding a single-attribute item to the baseline Q-matrix does not lead to an increased CCR with three independent attributes. The lack of change can be explained by the ties in hamming distances that cancel the effect of adding one more item. It is more likely to obtain a tie when there are an even number of $\boldsymbol{q}_{\{100\}}$, $\boldsymbol{q}_{\{010\}}$, or $\boldsymbol{q}_{\{001\}}$in the Q-matrix. Adding $\boldsymbol{q}_{\{110\}}$ slightly increases the CCR of $\boldsymbol{\alpha}_{110}$ and $\boldsymbol{\alpha}_{111}$ and adding $\boldsymbol{q}_{\{111\}}$ slightly increases in the CCR of $\boldsymbol{\alpha}_{111}$. In the above conditions, ties are likely to occur for all or some attribute profiles. However, when two items of each q-vector are added to the baseline, as in Q3-5, Q3-6, and Q3-7, the CCRs of all or some attribute profiles increase substantially, and almost no ties are observed.

With a linear hierarchy, all q-vectors have their equivalent single-attribute q-vectors. Therefore, all the Q-matrices contain single-attribute q-vectors. The comparison between Q3-2 and Q3-5, between Q3-3 and Q3-6, and between Q3-4 and Q3-7 in Table 24 suggests that a large probability for getting ties would hurt the classifications. For example, the CCR for $\boldsymbol{\alpha}_{000}$ increases slightly after adding a $q_{\{100\}}$ (Q3-2) but the classifications for other attribute profiles are not benefited. When two $\boldsymbol{q}_{\{100\}}$s are added (Q3-5), the CCR for $\boldsymbol{\alpha}_{000}$ and $\boldsymbol{\alpha}_{000}$ increase substantially. The probability of ties decreases from 0.23 (Q3-2) to 0.08 (Q3-5) with another $q_{\{100\}}$ added to the Q-matrix. Similar patterns can be found for the inverted pyramid or pyramid hierarchies in Table 25 and Table 26.

The negative effect of having even numbers of items in an item type is highlighted in the comparison between Q3-1, Q3-8, and Q3-15 in Table 23-Table 26. When the Q-matrix consists of two identity matrices, the CCR for each $\boldsymbol{\alpha}$ does not change or increase slightly compared to the baseline. However, when Q-matrix consists of three identity matrices, the CCR for each $\boldsymbol{\alpha}$ increases substantially.

Summarizing simulation results for three attributes, we conclude that tests with even number of items from each q-vector are less efficient than tests with each q-vector in odd times. When a q-vector appears in an even number and the item quality is homogeneous, it is more likely to have ties compared to the baseline situation of each attribute hierarchy, and consequently, the effect of extra test length is partially or completely canceled out. This conclusion also applies to conditions of four or five attributes, shown in Table 27-Table 37.

Table 23: NPC results for H3.1

Q	J	J_q					CCR				Pr(tie)			
		$q_{\{100\}}$	$q_{\{010\}}$	$q_{\{001\}}$	$q_{\{110\}}$	$q_{\{111\}}$	$\boldsymbol{\alpha}_{000}$	$\boldsymbol{\alpha}_{100}$	$\boldsymbol{\alpha}_{110}$	$\boldsymbol{\alpha}_{111}$	$\boldsymbol{\alpha}_{000}$	$\boldsymbol{\alpha}_{100}$	$\boldsymbol{\alpha}_{110}$	$\boldsymbol{\alpha}_{111}$
3-1	3	1	1	1	0	0	0.73	0.74	0.74	0.74	0.00	0.00	0.00	0.00
3-2	4	2	1	1	0	0	0.73	0.74	0.73	0.73	0.18	0.18	0.18	0.18
3-3	4	1	1	1	1	0	0.72	0.71	0.76	0.76	0.03	0.16	0.24	0.25
3-4	4	1	1	1	0	1	0.73	0.72	0.72	0.77	0.00	0.02	0.15	0.28
3-5	5	3	1	1	0	0	0.78	0.79	0.77	0.79	0.00	0.00	0.00	0.00
3-6	5	1	1	1	2	0	0.73	0.75	0.86	0.86	0.01	0.07	0.02	0.02
3-7	5	1	1	1	0	2	0.74	0.72	0.75	0.93	0.00	0.02	0.07	0.03
3-8	6	2	2	2	0	0	0.74	0.73	0.74	0.71	0.46	0.45	0.46	0.45
3-9	7	3	2	2	0	0	0.79	0.79	0.78	0.78	0.32	0.33	0.33	0.33
3-10	7	2	2	2	1	0	0.74	0.78	0.85	0.85	0.44	0.32	0.18	0.18
3-11	7	2	2	2	0	1	0.73	0.73	0.78	0.93	0.46	0.44	0.33	0.02
3-12	8	4	2	2	0	0	0.78	0.79	0.79	0.79	0.36	0.37	0.36	0.36
3-13	8	2	2	2	2	0	0.73	0.79	0.86	0.85	0.45	0.36	0.25	0.25
3-14	8	2	2	2	0	2	0.73	0.73	0.78	0.93	0.44	0.44	0.36	0.11
3-15	9	3	3	3	0	0	0.92	0.92	0.91	0.92	0.00	0.00	0.00	0.00

Note: J = test length; J_q = number of items with a certain q-vector; CCR = correct classification rate.

Table 24: NPC results for H3.2

Q	J	J_q			CCR				Pr(tie)			
		$q_{\{100\}}$	$q_{\{110\}}$	$q_{\{111\}}$	$\boldsymbol{\alpha}_{000}$	$\boldsymbol{\alpha}_{100}$	$\boldsymbol{\alpha}_{110}$	$\boldsymbol{\alpha}_{111}$	$\boldsymbol{\alpha}_{000}$	$\boldsymbol{\alpha}_{100}$	$\boldsymbol{\alpha}_{110}$	$\boldsymbol{\alpha}_{111}$
3-1	3	1	1	1	0.85	0.77	0.77	0.86	0.09	0.10	0.08	0.09
3-2	4	2	1	1	0.90	0.77	0.80	0.85	0.17	0.23	0.03	0.09
3-3	4	1	2	1	0.89	0.81	0.80	0.89	0.04	0.15	0.15	0.03
3-4	4	1	1	2	0.85	0.80	0.77	0.89	0.10	0.03	0.23	0.17
3-5	5	3	1	1	0.95	0.84	0.79	0.86	0.03	0.08	0.03	0.08
3-6	5	1	3	1	0.89	0.87	0.87	0.89	0.02	0.02	0.03	0.03
3-7	5	1	1	3	0.85	0.80	0.83	0.96	0.08	0.03	0.09	0.03
3-8	6	2	2	2	0.89	0.82	0.81	0.89	0.18	0.33	0.33	0.19
3-9	7	3	2	2	0.97	0.86	0.81	0.89	0.01	0.18	0.32	0.18
3-10	7	2	3	2	0.89	0.88	0.88	0.90	0.18	0.18	0.17	0.17
3-11	7	2	2	3	0.89	0.81	0.86	0.97	0.19	0.31	0.19	0.01
3-12	8	4	2	2	0.97	0.86	0.82	0.89	0.05	0.23	0.33	0.19
3-13	8	2	4	2	0.90	0.87	0.88	0.90	0.18	0.22	0.22	0.19
3-14	8	2	2	4	0.89	0.81	0.87	0.97	0.20	0.31	0.22	0.05
3-15	9	3	3	3	0.97	0.94	0.94	0.97	0.01	0.01	0.01	0.01

Note: J = test length; J_q = number of items with a certain q-vector; CCR = correct classification rate.

Table 25: NPC results for H3.3

Q	J	J_q				CCR				Pr(tie)			
		$q_{\{100\}}$	$q_{\{110\}}$	$q_{\{001\}}$	$q_{\{111\}}$	$\boldsymbol{\alpha}_{000}$	$\boldsymbol{\alpha}_{100}$	$\boldsymbol{\alpha}_{110}$	$\boldsymbol{\alpha}_{111}$	$\boldsymbol{\alpha}_{000}$	$\boldsymbol{\alpha}_{100}$	$\boldsymbol{\alpha}_{110}$	$\boldsymbol{\alpha}_{111}$
3-1	3	1	1	1	0	0.81	0.72	0.78	0.81	0.17	0.02	0.08	0.02
3-2	4	2	1	1	0	0.88	0.75	0.80	0.80	0.16	0.14	0.02	0.01
3-3	4	1	2	1	0	0.85	0.72	0.80	0.81	0.11	0.18	0.17	0.18
3-4	4	1	1	1	1	0.81	0.72	0.76	0.83	0.17	0.05	0.24	0.24
3-5	5	3	1	1	0	0.95	0.78	0.80	0.81	0.04	0.00	0.02	0.00
3-6	5	1	3	1	0	0.85	0.79	0.87	0.87	0.11	0.01	0.02	0.00
3-7	5	1	1	1	2	0.81	0.72	0.80	0.95	0.17	0.04	0.15	0.02
3-8	6	2	2	2	0	0.88	0.75	0.80	0.80	0.19	0.44	0.33	0.33
3-9	7	3	2	2	0	0.97	0.78	0.81	0.81	0.01	0.31	0.33	0.32
3-10	7	2	3	2	0	0.89	0.79	0.87	0.87	0.19	0.32	0.18	0.18
3-11	7	2	2	2	1	0.88	0.74	0.86	0.95	0.19	0.44	0.18	0.01
3-12	8	4	2	2	0	0.96	0.78	0.80	0.81	0.05	0.35	0.34	0.33
3-13	8	2	4	2	0	0.89	0.79	0.87	0.87	0.19	0.36	0.22	0.23
3-14	8	2	2	2	2	0.88	0.73	0.86	0.95	0.20	0.43	0.23	0.08
3-15	9	3	3	3	0	0.96	0.91	0.94	0.94	0.01	0.00	0.01	0.00

Note: J = test length; J_q = number of items with a certain q-vector; CCR = correct classification rate.

Table 26: NPC results for H3.4

Q	J	J_q				CCR				Pr(tie)			
		$q_{\{100\}}$	$q_{\{010\}}$	$q_{\{110\}}$	$q_{\{111\}}$	$\boldsymbol{\alpha}_{000}$	$\boldsymbol{\alpha}_{100}$	$\boldsymbol{\alpha}_{110}$	$\boldsymbol{\alpha}_{111}$	$\boldsymbol{\alpha}_{000}$	$\boldsymbol{\alpha}_{100}$	$\boldsymbol{\alpha}_{110}$	$\boldsymbol{\alpha}_{111}$
3-1	3	1	1	0	1	0.81	0.76	0.73	0.81	0.02	0.08	0.02	0.17
3-2	4	2	1	0	1	0.81	0.78	0.73	0.84	0.19	0.25	0.17	0.11
3-3	4	1	1	1	1	0.80	0.78	0.76	0.88	0.03	0.15	0.22	0.04
3-4	4	1	1	0	2	0.80	0.81	0.73	0.88	0.01	0.02	0.14	0.16
3-5	5	3	1	0	1	0.87	0.83	0.79	0.85	0.00	0.08	0.01	0.11
3-6	5	1	1	2	1	0.81	0.82	0.86	0.88	0.02	0.10	0.02	0.04
3-7	5	1	1	0	3	0.81	0.80	0.78	0.95	0.00	0.02	0.01	0.04
3-8	6	2	2	0	2	0.81	0.80	0.73	0.88	0.33	0.34	0.45	0.20
3-9	7	3	2	0	2	0.87	0.86	0.80	0.90	0.18	0.19	0.32	0.18
3-10	7	2	2	1	2	0.81	0.86	0.86	0.89	0.32	0.18	0.19	0.17
3-11	7	2	2	0	3	0.81	0.81	0.80	0.97	0.32	0.31	0.31	0.01
3-12	8	4	2	0	2	0.87	0.87	0.79	0.89	0.22	0.24	0.36	0.18
3-13	8	2	2	2	2	0.80	0.86	0.86	0.89	0.33	0.23	0.25	0.19
3-14	8	2	2	0	4	0.81	0.81	0.78	0.96	0.34	0.32	0.36	0.06
3-15	9	3	3	0	3	0.95	0.94	0.92	0.96	0.00	0.01	0.00	0.01

Note: J = test length; J_q = number of items with a certain q-vector; CCR = correct classification rate.

Table 27: NPC results for H4.1

Q	J	CCR					Pr(tie)				
		α_{0000}	α_{1000}	α_{1100}	α_{1110}	α_{1111}	α_{0000}	α_{1000}	α_{1100}	α_{1110}	α_{1111}
4-1	4	0.65	0.65	0.65	0.65	0.66	0.00	0.00	0.00	0.00	0.00
4-2	5	0.66	0.64	0.66	0.65	0.66	0.17	0.18	0.18	0.19	0.18
4-3	5	0.65	0.64	0.68	0.68	0.68	0.03	0.17	0.25	0.24	0.25
4-4	5	0.67	0.67	0.64	0.72	0.70	0.00	0.02	0.14	0.28	0.29
4-5	5	0.66	0.65	0.64	0.64	0.73	0.00	0.00	0.02	0.13	0.33
4-6	6	0.71	0.72	0.71	0.70	0.71	0.00	0.00	0.00	0.00	0.00
4-7	6	0.67	0.68	0.76	0.77	0.77	0.01	0.08	0.02	0.02	0.02
4-8	6	0.66	0.65	0.67	0.84	0.84	0.00	0.02	0.06	0.03	0.03
4-9	6	0.65	0.65	0.65	0.67	0.90	0.00	0.00	0.01	0.06	0.04
4-10	8	0.66	0.66	0.66	0.65	0.66	0.54	0.55	0.54	0.56	0.55
4-11	9	0.71	0.71	0.71	0.70	0.71	0.44	0.45	0.46	0.45	0.45
4-12	9	0.64	0.71	0.77	0.77	0.76	0.55	0.44	0.32	0.33	0.34
4-13	9	0.66	0.64	0.70	0.83	0.84	0.54	0.55	0.45	0.20	0.19
4-14	9	0.66	0.65	0.65	0.69	0.91	0.54	0.54	0.55	0.45	0.03
4-15	10	0.71	0.71	0.72	0.71	0.71	0.46	0.48	0.47	0.47	0.47
4-16	10	0.66	0.70	0.76	0.77	0.77	0.54	0.48	0.38	0.39	0.38
4-17	10	0.64	0.66	0.70	0.83	0.84	0.56	0.55	0.48	0.27	0.27
4-18	10	0.65	0.66	0.65	0.71	0.91	0.54	0.54	0.55	0.47	0.14
4-19	12	0.90	0.89	0.89	0.89	0.89	0.00	0.00	0.00	0.00	0.00

Table 28: NPC results for H4.2

Q	J	CCR					Pr(tie)				
		α_{0000}	α_{1000}	α_{1100}	α_{1110}	α_{1111}	α_{0000}	α_{1000}	α_{1100}	α_{1110}	α_{1111}
4-1	4	0.85	0.76	0.74	0.77	0.86	0.10	0.10	0.16	0.09	0.10
4-2	5	0.89	0.76	0.76	0.77	0.84	0.17	0.23	0.11	0.09	0.09
4-3	5	0.88	0.79	0.76	0.79	0.85	0.03	0.16	0.22	0.04	0.09
4-4	5	0.85	0.79	0.76	0.79	0.88	0.10	0.04	0.22	0.16	0.03
4-5	5	0.85	0.78	0.76	0.78	0.88	0.09	0.10	0.10	0.23	0.18
4-6	6	0.96	0.82	0.77	0.77	0.84	0.03	0.10	0.11	0.09	0.11
4-7	6	0.88	0.86	0.81	0.80	0.85	0.03	0.02	0.11	0.03	0.09
4-8	6	0.86	0.80	0.82	0.87	0.89	0.10	0.04	0.12	0.02	0.03
4-9	6	0.85	0.76	0.76	0.82	0.96	0.09	0.09	0.12	0.09	0.02
4-10	8	0.89	0.80	0.79	0.80	0.89	0.18	0.34	0.34	0.33	0.19
4-11	9	0.97	0.87	0.80	0.82	0.89	0.01	0.17	0.32	0.32	0.19
4-12	9	0.89	0.88	0.86	0.81	0.89	0.17	0.17	0.20	0.32	0.18
4-13	9	0.89	0.81	0.86	0.87	0.90	0.19	0.31	0.19	0.18	0.17
4-14	9	0.89	0.79	0.80	0.87	0.97	0.19	0.33	0.34	0.18	0.01
4-15	10	0.97	0.88	0.80	0.80	0.89	0.05	0.22	0.34	0.33	0.19
4-16	10	0.90	0.87	0.87	0.82	0.89	0.18	0.22	0.23	0.32	0.19
4-17	10	0.89	0.81	0.86	0.87	0.89	0.20	0.33	0.23	0.22	0.19
4-18	10	0.89	0.82	0.80	0.87	0.97	0.19	0.32	0.34	0.23	0.06
4-19	12	0.97	0.95	0.94	0.95	0.97	0.01	0.01	0.02	0.01	0.01

Table 29: NPC results for H4.3

Q	J	CCR					Pr(tie)				
		α_{0000}	α_{1000}	α_{1100}	α_{1110}	α_{1111}	α_{0000}	α_{1000}	α_{1100}	α_{1110}	α_{1111}
4-1	4	0.77	0.70	0.70	0.76	0.76	0.09	0.09	0.10	0.09	0.09
4-2	5	0.79	0.69	0.72	0.76	0.76	0.18	0.24	0.03	0.09	0.10
4-3	5	0.80	0.73	0.73	0.79	0.80	0.03	0.15	0.14	0.03	0.03
4-4	5	0.76	0.71	0.69	0.79	0.79	0.10	0.03	0.23	0.17	0.16
4-5	5	0.77	0.69	0.70	0.76	0.82	0.09	0.09	0.11	0.22	0.24
4-6	6	0.86	0.76	0.73	0.76	0.76	0.02	0.09	0.03	0.09	0.09
4-7	6	0.81	0.77	0.78	0.80	0.79	0.02	0.03	0.03	0.02	0.03
4-8	6	0.75	0.72	0.76	0.87	0.86	0.10	0.03	0.09	0.02	0.03
4-9	6	0.76	0.69	0.69	0.79	0.94	0.10	0.09	0.10	0.15	0.04
4-10	8	0.80	0.72	0.73	0.81	0.81	0.34	0.46	0.45	0.34	0.33
4-11	9	0.87	0.77	0.74	0.81	0.79	0.19	0.34	0.44	0.32	0.35
4-12	9	0.81	0.78	0.80	0.81	0.80	0.33	0.33	0.32	0.31	0.32
4-13	9	0.80	0.72	0.78	0.86	0.87	0.34	0.45	0.33	0.19	0.18
4-14	9	0.80	0.73	0.73	0.85	0.95	0.34	0.46	0.44	0.18	0.01
4-15	10	0.88	0.78	0.73	0.80	0.80	0.22	0.38	0.46	0.34	0.34
4-16	10	0.80	0.78	0.78	0.81	0.80	0.33	0.36	0.37	0.32	0.33
4-17	10	0.80	0.72	0.78	0.87	0.88	0.33	0.46	0.37	0.23	0.22
4-18	10	0.80	0.73	0.72	0.86	0.94	0.33	0.44	0.46	0.23	0.09
4-19	12	0.94	0.91	0.92	0.94	0.95	0.01	0.01	0.01	0.01	0.00

Table 30: NPC results for H4.4

Q	J	CCR					Pr(tie)				
		α_{0000}	α_{1000}	α_{1100}	α_{1110}	α_{1111}	α_{0000}	α_{1000}	α_{1100}	α_{1110}	α_{1111}
4-1	4	0.84	0.73	0.69	0.78	0.76	0.10	0.16	0.09	0.09	0.09
4-2	5	0.87	0.74	0.73	0.77	0.76	0.18	0.29	0.03	0.09	0.10
4-3	5	0.86	0.79	0.73	0.80	0.81	0.05	0.14	0.13	0.03	0.03
4-4	5	0.83	0.76	0.70	0.77	0.79	0.10	0.11	0.24	0.25	0.17
4-5	5	0.86	0.73	0.69	0.76	0.75	0.09	0.16	0.11	0.23	0.24
4-6	6	0.96	0.78	0.74	0.79	0.76	0.03	0.17	0.03	0.08	0.09
4-7	6	0.89	0.86	0.78	0.80	0.78	0.03	0.04	0.02	0.02	0.02
4-8	6	0.85	0.76	0.75	0.83	0.86	0.09	0.11	0.09	0.08	0.02
4-9	6	0.85	0.72	0.69	0.79	0.80	0.09	0.16	0.10	0.15	0.15
4-10	8	0.88	0.80	0.73	0.80	0.81	0.19	0.33	0.44	0.34	0.33
4-11	9	0.97	0.85	0.73	0.80	0.81	0.01	0.20	0.44	0.34	0.33
4-12	9	0.90	0.87	0.79	0.81	0.81	0.18	0.19	0.30	0.31	0.32
4-13	9	0.89	0.80	0.78	0.87	0.87	0.18	0.32	0.33	0.19	0.18
4-14	9	0.88	0.79	0.73	0.85	0.86	0.19	0.34	0.44	0.19	0.18
4-15	10	0.97	0.86	0.73	0.80	0.79	0.05	0.23	0.43	0.34	0.34
4-16	10	0.89	0.87	0.79	0.81	0.80	0.19	0.22	0.36	0.33	0.33
4-17	10	0.88	0.81	0.77	0.86	0.88	0.19	0.34	0.37	0.23	0.22
4-18	10	0.89	0.80	0.71	0.86	0.86	0.19	0.33	0.45	0.22	0.23
4-19	12	0.97	0.94	0.92	0.94	0.94	0.01	0.01	0.01	0.01	0.01

Table 31: NPC results for H4.5

Q	J	CCR					Pr(tie)				
		α_{0000}	α_{1000}	α_{1100}	α_{1110}	α_{1111}	α_{0000}	α_{1000}	α_{1100}	α_{1110}	α_{1111}
4-1	4	0.81	0.76	0.77	0.70	0.73	0.03	0.08	0.08	0.09	0.16
4-2	5	0.81	0.79	0.77	0.70	0.75	0.18	0.18	0.26	0.23	0.10
4-3	5	0.81	0.78	0.79	0.72	0.80	0.03	0.16	0.16	0.27	0.04
4-4	5	0.80	0.80	0.80	0.72	0.79	0.01	0.03	0.03	0.14	0.14
4-5	5	0.80	0.76	0.75	0.73	0.73	0.02	0.09	0.10	0.03	0.30
4-6	6	0.88	0.86	0.83	0.75	0.76	0.01	0.02	0.09	0.09	0.11
4-7	6	0.81	0.84	0.83	0.81	0.78	0.02	0.08	0.08	0.11	0.04
4-8	6	0.82	0.80	0.80	0.78	0.85	0.00	0.02	0.02	0.02	0.04
4-9	6	0.80	0.77	0.77	0.73	0.79	0.02	0.09	0.09	0.04	0.16
4-10	8	0.80	0.79	0.80	0.71	0.80	0.33	0.34	0.34	0.46	0.34
4-11	9	0.87	0.87	0.86	0.78	0.80	0.18	0.18	0.19	0.33	0.33
4-12	9	0.81	0.87	0.87	0.84	0.82	0.32	0.18	0.17	0.19	0.31
4-13	9	0.81	0.81	0.80	0.78	0.87	0.33	0.31	0.32	0.32	0.19
4-14	9	0.81	0.80	0.81	0.73	0.85	0.33	0.33	0.34	0.42	0.20
4-15	10	0.87	0.88	0.86	0.77	0.81	0.22	0.22	0.23	0.37	0.33
4-16	10	0.81	0.87	0.86	0.84	0.81	0.34	0.23	0.22	0.26	0.33
4-17	10	0.81	0.80	0.81	0.79	0.87	0.33	0.33	0.33	0.36	0.23
4-18	10	0.81	0.80	0.80	0.73	0.85	0.33	0.35	0.34	0.44	0.23
4-19	12	0.94	0.94	0.94	0.91	0.94	0.00	0.01	0.01	0.01	0.01

Table 32: NPC results for H5.1

Q	J	CCR						Pr(tie)					
		α_{00000}	α_{10000}	α_{11000}	α_{11100}	α_{11110}	α_{11111}	α_{00000}	α_{10000}	α_{11000}	α_{11100}	α_{11110}	α_{11111}
5-1	5	0.58	0.58	0.60	0.59	0.59	0.59	0.00	0.00	0.00	0.00	0.00	0.00
5-2	6	0.58	0.59	0.58	0.59	0.59	0.60	0.18	0.17	0.18	0.18	0.19	0.18
5-3	6	0.60	0.59	0.62	0.60	0.61	0.61	0.03	0.16	0.23	0.24	0.25	0.23
5-4	6	0.58	0.59	0.57	0.64	0.62	0.64	0.00	0.02	0.16	0.29	0.30	0.29
5-5	6	0.60	0.60	0.59	0.58	0.66	0.66	0.00	0.00	0.02	0.13	0.34	0.33
5-6	6	0.59	0.58	0.59	0.59	0.56	0.69	0.00	0.00	0.00	0.02	0.12	0.35
5-7	7	0.63	0.62	0.63	0.64	0.65	0.63	0.00	0.00	0.00	0.00	0.00	0.00
5-8	7	0.60	0.62	0.69	0.70	0.68	0.69	0.01	0.08	0.02	0.02	0.02	0.02
5-9	7	0.58	0.60	0.60	0.76	0.75	0.76	0.00	0.02	0.07	0.03	0.03	0.03
5-10	7	0.59	0.58	0.59	0.60	0.81	0.81	0.00	0.00	0.01	0.06	0.04	0.04
5-11	7	0.58	0.57	0.59	0.59	0.61	0.88	0.00	0.00	0.00	0.01	0.06	0.06
5-12	10	0.59	0.59	0.60	0.59	0.59	0.60	0.63	0.63	0.63	0.63	0.63	0.63
5-13	11	0.65	0.64	0.64	0.63	0.64	0.64	0.54	0.56	0.55	0.54	0.55	0.55
5-14	11	0.59	0.62	0.69	0.68	0.69	0.69	0.63	0.55	0.44	0.46	0.45	0.45
5-15	11	0.58	0.60	0.64	0.76	0.75	0.76	0.64	0.63	0.53	0.33	0.34	0.33
5-16	11	0.60	0.59	0.60	0.62	0.81	0.81	0.64	0.62	0.62	0.56	0.21	0.21
5-17	11	0.58	0.59	0.59	0.59	0.63	0.89	0.64	0.64	0.63	0.63	0.54	0.05
5-18	12	0.63	0.64	0.63	0.64	0.64	0.63	0.58	0.57	0.56	0.57	0.58	0.57
5-19	12	0.59	0.63	0.69	0.69	0.68	0.68	0.64	0.57	0.49	0.50	0.51	0.51
5-20	12	0.59	0.60	0.64	0.75	0.76	0.75	0.63	0.63	0.56	0.40	0.40	0.40
5-21	12	0.59	0.59	0.58	0.63	0.83	0.83	0.63	0.62	0.63	0.57	0.29	0.28
5-22	12	0.58	0.61	0.60	0.59	0.62	0.89	0.64	0.61	0.63	0.63	0.57	0.17
5-23	15	0.87	0.88	0.87	0.86	0.87	0.88	0.00	0.00	0.00	0.00	0.00	0.00

Table 33: NPC results for H5.2

Q	J	CCR						Pr(tie)					
		α_{00000}	α_{10000}	α_{11000}	α_{11100}	α_{11110}	α_{11111}	α_{00000}	α_{10000}	α_{11000}	α_{11100}	α_{11110}	α_{11111}
5-1	5	0.85	0.76	0.72	0.73	0.76	0.85	0.10	0.10	0.17	0.16	0.10	0.11
5-2	6	0.89	0.77	0.75	0.72	0.76	0.84	0.17	0.24	0.11	0.17	0.09	0.10
5-3	6	0.87	0.79	0.77	0.76	0.76	0.85	0.04	0.16	0.22	0.11	0.10	0.09
5-4	6	0.84	0.79	0.75	0.76	0.78	0.85	0.11	0.04	0.22	0.22	0.04	0.10
5-5	6	0.85	0.76	0.76	0.76	0.78	0.88	0.09	0.10	0.12	0.23	0.16	0.04
5-6	6	0.84	0.76	0.72	0.76	0.76	0.89	0.11	0.09	0.17	0.11	0.25	0.17
5-7	7	0.96	0.82	0.77	0.74	0.77	0.84	0.03	0.11	0.11	0.16	0.09	0.10
5-8	7	0.89	0.86	0.82	0.76	0.76	0.85	0.03	0.02	0.10	0.11	0.09	0.09
5-9	7	0.84	0.80	0.82	0.82	0.80	0.85	0.09	0.03	0.12	0.11	0.03	0.09
5-10	7	0.85	0.77	0.76	0.83	0.86	0.89	0.09	0.09	0.11	0.11	0.02	0.02
5-11	7	0.85	0.76	0.73	0.77	0.83	0.96	0.09	0.09	0.16	0.10	0.09	0.03
5-12	10	0.89	0.79	0.79	0.80	0.80	0.89	0.19	0.34	0.35	0.35	0.33	0.19
5-13	11	0.97	0.87	0.80	0.79	0.81	0.89	0.01	0.18	0.33	0.33	0.32	0.19
5-14	11	0.89	0.87	0.86	0.80	0.80	0.89	0.17	0.19	0.19	0.33	0.34	0.20
5-15	11	0.89	0.82	0.87	0.86	0.81	0.88	0.18	0.32	0.19	0.19	0.32	0.20
5-16	11	0.88	0.81	0.80	0.86	0.87	0.89	0.20	0.32	0.32	0.19	0.18	0.17
5-17	11	0.89	0.80	0.79	0.80	0.86	0.97	0.19	0.33	0.34	0.33	0.19	0.01
5-18	12	0.97	0.86	0.80	0.79	0.80	0.89	0.05	0.23	0.33	0.36	0.33	0.19
5-19	12	0.90	0.86	0.86	0.81	0.80	0.89	0.19	0.23	0.23	0.33	0.32	0.19
5-20	12	0.89	0.81	0.86	0.87	0.80	0.89	0.18	0.33	0.24	0.22	0.35	0.19
5-21	12	0.88	0.80	0.79	0.86	0.88	0.90	0.20	0.33	0.34	0.23	0.22	0.18
5-22	12	0.89	0.79	0.80	0.79	0.86	0.97	0.18	0.34	0.33	0.34	0.23	0.05
5-23	15	0.97	0.95	0.93	0.94	0.94	0.97	0.01	0.01	0.01	0.02	0.01	0.01

Table 34: NPC results for H5.3

Q	J	CCR						Pr(tie)					
		α_{00000}	α_{10000}	α_{11000}	α_{11100}	α_{11110}	α_{11111}	α_{00000}	α_{10000}	α_{11000}	α_{11100}	α_{11110}	α_{11111}
5-1	5	0.84	0.68	0.62	0.70	0.73	0.73	0.10	0.20	0.09	0.07	0.02	0.01
5-2	6	0.86	0.69	0.64	0.69	0.73	0.73	0.18	0.34	0.04	0.09	0.02	0.00
5-3	6	0.87	0.77	0.65	0.71	0.72	0.71	0.06	0.15	0.13	0.03	0.01	0.00
5-4	6	0.84	0.72	0.62	0.72	0.74	0.74	0.11	0.16	0.23	0.17	0.17	0.17
5-5	6	0.83	0.68	0.63	0.68	0.75	0.76	0.11	0.22	0.11	0.23	0.25	0.24
5-6	6	0.84	0.68	0.62	0.69	0.71	0.78	0.11	0.21	0.09	0.11	0.17	0.31
5-7	7	0.95	0.74	0.64	0.70	0.73	0.72	0.03	0.22	0.04	0.07	0.02	0.00
5-8	7	0.88	0.84	0.70	0.72	0.72	0.72	0.03	0.06	0.02	0.02	0.01	0.00
5-9	7	0.84	0.71	0.67	0.79	0.79	0.79	0.10	0.18	0.08	0.02	0.01	0.00
5-10	7	0.83	0.68	0.62	0.72	0.86	0.86	0.11	0.22	0.10	0.14	0.02	0.02
5-11	7	0.83	0.69	0.62	0.69	0.74	0.93	0.12	0.22	0.10	0.09	0.09	0.03
5-12	10	0.88	0.80	0.64	0.72	0.73	0.73	0.19	0.34	0.55	0.45	0.45	0.44
5-13	11	0.97	0.84	0.66	0.71	0.73	0.73	0.01	0.20	0.53	0.46	0.45	0.45
5-14	11	0.90	0.87	0.70	0.73	0.72	0.74	0.18	0.18	0.45	0.45	0.46	0.43
5-15	11	0.89	0.78	0.70	0.78	0.79	0.79	0.18	0.34	0.46	0.34	0.33	0.33
5-16	11	0.89	0.77	0.66	0.77	0.86	0.85	0.19	0.36	0.55	0.34	0.19	0.19
5-17	11	0.88	0.78	0.66	0.71	0.78	0.93	0.20	0.35	0.54	0.46	0.31	0.02
5-18	12	0.97	0.85	0.65	0.72	0.72	0.73	0.05	0.24	0.55	0.47	0.46	0.45
5-19	12	0.90	0.87	0.71	0.73	0.73	0.73	0.18	0.23	0.47	0.46	0.44	0.45
5-20	12	0.89	0.78	0.69	0.78	0.79	0.80	0.19	0.35	0.49	0.37	0.37	0.35
5-21	12	0.88	0.78	0.67	0.77	0.85	0.85	0.19	0.35	0.54	0.38	0.25	0.24
5-22	12	0.88	0.79	0.66	0.72	0.78	0.93	0.19	0.34	0.54	0.46	0.37	0.11
5-23	15	0.97	0.93	0.89	0.91	0.91	0.92	0.00	0.02	0.01	0.01	0.00	0.00

Table 35: NPC results for H5.4

Q	J	CCR						Pr(tie)					
		α_{00000}	α_{10000}	α_{11000}	α_{11100}	α_{11110}	α_{11111}	α_{00000}	α_{10000}	α_{11000}	α_{11100}	α_{11110}	α_{11111}
5-1	5	0.80	0.66	0.61	0.66	0.69	0.74	0.17	0.15	0.09	0.03	0.10	0.02
5-2	6	0.86	0.66	0.64	0.64	0.70	0.72	0.17	0.27	0.04	0.03	0.08	0.02
5-3	6	0.83	0.72	0.65	0.66	0.72	0.73	0.12	0.15	0.14	0.14	0.02	0.01
5-4	6	0.80	0.66	0.61	0.66	0.71	0.74	0.18	0.17	0.24	0.23	0.19	0.19
5-5	6	0.79	0.65	0.62	0.64	0.75	0.74	0.17	0.16	0.11	0.16	0.23	0.25
5-6	6	0.80	0.65	0.63	0.66	0.66	0.79	0.17	0.16	0.08	0.05	0.23	0.29
5-7	7	0.94	0.70	0.65	0.66	0.69	0.73	0.05	0.16	0.03	0.02	0.08	0.02
5-8	7	0.83	0.76	0.70	0.70	0.72	0.72	0.12	0.05	0.03	0.01	0.02	0.00
5-9	7	0.80	0.64	0.64	0.76	0.78	0.78	0.18	0.17	0.14	0.02	0.02	0.01
5-10	7	0.79	0.66	0.61	0.68	0.86	0.85	0.18	0.16	0.10	0.10	0.03	0.02
5-11	7	0.80	0.65	0.63	0.65	0.70	0.93	0.17	0.16	0.08	0.05	0.14	0.04
5-12	10	0.88	0.72	0.65	0.66	0.73	0.72	0.19	0.46	0.56	0.54	0.44	0.45
5-13	11	0.97	0.77	0.67	0.67	0.71	0.73	0.01	0.33	0.53	0.54	0.46	0.44
5-14	11	0.89	0.79	0.72	0.71	0.72	0.73	0.18	0.33	0.44	0.45	0.44	0.45
5-15	11	0.87	0.71	0.71	0.77	0.78	0.79	0.21	0.45	0.44	0.32	0.32	0.32
5-16	11	0.88	0.72	0.66	0.71	0.85	0.85	0.19	0.45	0.55	0.43	0.20	0.19
5-17	11	0.87	0.71	0.65	0.66	0.77	0.92	0.20	0.46	0.53	0.54	0.33	0.02
5-18	12	0.97	0.77	0.66	0.66	0.72	0.73	0.06	0.37	0.54	0.54	0.46	0.44
5-19	12	0.89	0.78	0.69	0.71	0.72	0.73	0.19	0.37	0.49	0.48	0.45	0.45
5-20	12	0.87	0.73	0.69	0.78	0.79	0.79	0.21	0.44	0.48	0.36	0.36	0.36
5-21	12	0.88	0.72	0.65	0.72	0.86	0.86	0.20	0.45	0.54	0.46	0.25	0.24
5-22	12	0.87	0.71	0.65	0.67	0.77	0.93	0.20	0.46	0.55	0.53	0.37	0.11
5-23	15	0.96	0.91	0.89	0.90	0.92	0.93	0.01	0.02	0.01	0.00	0.01	0.00

Table 36: NPC results for H5.5

Q	J	CCR						Pr(tie)					
		α_{00000}	α_{10000}	α_{11000}	α_{11100}	α_{11110}	α_{11111}	α_{00000}	α_{10000}	α_{11000}	α_{11100}	α_{11110}	α_{11111}
5-1	5	0.73	0.73	0.70	0.62	0.68	0.83	0.00	0.02	0.08	0.09	0.20	0.11
5-2	6	0.73	0.73	0.69	0.63	0.72	0.84	0.19	0.19	0.24	0.24	0.18	0.11
5-3	6	0.73	0.71	0.71	0.65	0.76	0.84	0.03	0.16	0.30	0.30	0.12	0.11
5-4	6	0.72	0.73	0.69	0.67	0.78	0.83	0.01	0.03	0.15	0.33	0.06	0.11
5-5	6	0.73	0.73	0.72	0.66	0.78	0.87	0.00	0.01	0.03	0.13	0.15	0.06
5-6	6	0.71	0.73	0.69	0.64	0.70	0.88	0.00	0.02	0.08	0.04	0.32	0.17
5-7	7	0.79	0.78	0.74	0.67	0.72	0.84	0.00	0.03	0.08	0.09	0.17	0.11
5-8	7	0.73	0.75	0.81	0.74	0.74	0.84	0.02	0.07	0.10	0.10	0.12	0.10
5-9	7	0.72	0.72	0.73	0.79	0.78	0.85	0.00	0.02	0.08	0.11	0.06	0.09
5-10	7	0.73	0.73	0.72	0.69	0.85	0.88	0.00	0.01	0.02	0.03	0.05	0.03
5-11	7	0.73	0.73	0.69	0.64	0.74	0.95	0.00	0.02	0.08	0.04	0.22	0.03
5-12	10	0.73	0.74	0.71	0.65	0.79	0.88	0.44	0.44	0.46	0.55	0.34	0.20
5-13	11	0.79	0.79	0.77	0.70	0.79	0.89	0.32	0.32	0.34	0.45	0.34	0.19
5-14	11	0.74	0.79	0.84	0.76	0.79	0.89	0.44	0.32	0.19	0.33	0.32	0.19
5-15	11	0.73	0.73	0.78	0.82	0.82	0.89	0.44	0.44	0.32	0.19	0.31	0.18
5-16	11	0.73	0.72	0.72	0.71	0.87	0.89	0.44	0.44	0.44	0.44	0.19	0.18
5-17	11	0.73	0.74	0.73	0.66	0.85	0.97	0.44	0.44	0.44	0.54	0.20	0.01
5-18	12	0.79	0.79	0.77	0.70	0.80	0.88	0.35	0.35	0.37	0.46	0.32	0.20
5-19	12	0.73	0.79	0.85	0.77	0.80	0.89	0.44	0.36	0.26	0.37	0.33	0.18
5-20	12	0.73	0.73	0.77	0.82	0.80	0.89	0.45	0.45	0.36	0.28	0.34	0.19
5-21	12	0.73	0.72	0.73	0.70	0.86	0.90	0.44	0.45	0.45	0.48	0.22	0.18
5-22	12	0.73	0.72	0.72	0.67	0.85	0.97	0.45	0.45	0.46	0.53	0.25	0.05
5-23	15	0.92	0.93	0.92	0.89	0.93	0.97	0.00	0.00	0.01	0.01	0.02	0.01

Table 37: NPC results for H5.6

Q	J	CCR						Pr(tie)					
		α_{00000}	α_{10000}	α_{11000}	α_{11100}	α_{11110}	α_{11111}	α_{00000}	α_{10000}	α_{11000}	α_{11100}	α_{11110}	α_{11111}
5-1	5	0.72	0.69	0.65	0.69	0.65	0.79	0.02	0.08	0.03	0.23	0.17	0.17
5-2	6	0.73	0.69	0.65	0.73	0.67	0.79	0.18	0.25	0.20	0.17	0.11	0.18
5-3	6	0.72	0.71	0.68	0.75	0.71	0.79	0.03	0.15	0.23	0.12	0.05	0.18
5-4	6	0.73	0.72	0.66	0.75	0.71	0.83	0.01	0.02	0.15	0.22	0.15	0.13
5-5	6	0.73	0.69	0.66	0.71	0.75	0.86	0.02	0.08	0.04	0.26	0.20	0.06
5-6	6	0.73	0.69	0.65	0.73	0.66	0.86	0.02	0.08	0.03	0.17	0.28	0.16
5-7	7	0.79	0.76	0.71	0.73	0.69	0.80	0.01	0.09	0.02	0.18	0.11	0.17
5-8	7	0.73	0.75	0.78	0.75	0.70	0.81	0.02	0.09	0.03	0.12	0.05	0.17
5-9	7	0.74	0.72	0.72	0.82	0.78	0.84	0.01	0.03	0.01	0.13	0.05	0.12
5-10	7	0.72	0.70	0.66	0.74	0.84	0.87	0.01	0.07	0.03	0.22	0.05	0.04
5-11	7	0.73	0.69	0.65	0.72	0.69	0.94	0.02	0.08	0.02	0.18	0.17	0.05
5-12	10	0.72	0.72	0.66	0.78	0.71	0.87	0.46	0.46	0.53	0.34	0.47	0.20
5-13	11	0.79	0.78	0.70	0.80	0.72	0.88	0.33	0.33	0.46	0.34	0.45	0.20
5-14	11	0.73	0.77	0.77	0.80	0.73	0.88	0.45	0.33	0.32	0.32	0.44	0.20
5-15	11	0.74	0.71	0.71	0.85	0.79	0.89	0.44	0.45	0.45	0.21	0.33	0.18
5-16	11	0.72	0.72	0.66	0.84	0.85	0.89	0.46	0.46	0.54	0.20	0.17	0.18
5-17	11	0.73	0.72	0.65	0.80	0.77	0.97	0.46	0.47	0.55	0.34	0.33	0.01
5-18	12	0.79	0.76	0.71	0.79	0.72	0.87	0.35	0.38	0.46	0.34	0.44	0.21
5-19	12	0.74	0.77	0.76	0.79	0.72	0.89	0.44	0.38	0.38	0.35	0.47	0.19
5-20	12	0.73	0.72	0.72	0.85	0.80	0.88	0.46	0.46	0.47	0.24	0.34	0.20
5-21	12	0.73	0.73	0.66	0.85	0.86	0.89	0.44	0.44	0.54	0.23	0.25	0.19
5-22	12	0.73	0.72	0.67	0.78	0.76	0.97	0.43	0.46	0.54	0.35	0.37	0.06
5-23	15	0.92	0.91	0.90	0.93	0.91	0.96	0.00	0.01	0.00	0.02	0.01	0.02

5.5 Discussion

Nonparametric classifications could play an important role in formative classroom assessment. Tests developed by the teachers constitute a large part of classroom assessments. With the guidance of psychometric theory, teachers may be able to extract more formative feedback. Nonparametric classifications based on CDMs offer solutions to both test construction and result interpretations. The teachers may develop the items under the guidance of CDM-based assessment (Rupp et al., 2010). However, it is not likely to collect enough response data in the classroom setting for model estimation (including calibration and classification). Besides, there are concerns about the invariance properties of model parameters. In response to these limitations, researchers have proposed different nonparametric classification methods to produce student results without having to estimate item parameters (Chiu & Douglas, 2013; Chiu, Sun, & Bian, 2018; Wang & Douglas, 2015). This study adds to the literature by providing insights into how to construct such a test.

Q-matrix design is at the center of test construction for both parametric and nonparametric CDM-based tests. Test construction involves practical questions, including how long the test should be and how many items are needed from each type. Note that what we discuss in Chapter 3 about equivalent q-vectors and different types of Q-matrices also applies to the nonparametric situation. Generally, Q-matrix designs that work well for MLE classifications also work well for nonparametric classifications. The ties in the hamming distance are parallel to equal or similar likelihoods between attribute profiles.

The simulation study compared Q-matrix designs with K to $3 \times K$ items. Longer tests were not considered because the situation is teacher-developed classroom assessment. It is important to include the single-attribute items for nonparametric classifications. Adding an odd number of

multiple-attribute items can increase the CCR of a subset of $\boldsymbol{\alpha}$s while adding an odd number of single-attribute items leads to an increased CCR for every $\boldsymbol{\alpha}$.

It is recommended that a Q-matrix has an odd number of items in each q-vector. A test with an even number of items in a certain q-vector is generally not substantially better than a test with one less item in this q-vector. This is especially true when the item quality is homogeneous.

An important implication for teachers is that more items do not necessarily mean more accurate classifications. A single-attribute item is generally more useful than a multiple-attribute one. However, if the classification of certain $\boldsymbol{\alpha}$s, say $\boldsymbol{\alpha}_{110}$, is of particular interest, then including the appropriate multiple-attribute item (in this case, $\boldsymbol{q}_{\{110\}}$) in the Q-matrix becomes meaningful in terms of CCR.

A classroom assessment network can be built where teachers develop their items based on CDMs with q-vectors and the corresponding curriculum identified. Such items can be collected from teachers and form various item pools, which can later be used for CD-CAT or nonparametric CD-CAT. At last, this study assumes the DINA model as the underlying CDM. Future research could explore different Q-matrix designs for NPC with other underlying CDMs.

Chapter 6 Item pool design for CD-CAT

6.1 Introduction

Item pool design is an important but often neglected area for CD-CAT. Since the item pool design for CD-CAT has not been addressed in the literature, we draw from studies on item pool design for CAT based on IRT models (e.g., Reckase, 2010; Thissen, Reeve, Bjorner, & Chang, 2007; Veldkamp & van der Linden, 2000). The findings for IRT-based CAT can be informative because CD-CATs are the same sequential optimization problems using CDMs instead of IRT models as the item response model. However, the categorical nature of the latent constructs in CDM decides that new studies are needed for the CD-CAT context.

Besides, CD-CAT has different priorities from those of IRT-based CAT. Classroom formative assessments are generally low-stakes tests, so test security issues are not of primary concern. It is acceptable that tests overlap between students. What is of more importance is to assign new items to a student each time he or she takes the test during the instructional period. Therefore, different requirements are imposed on item pool design for classroom formative assessments as compared to high-stakes standardized tests.

When a series of formative assessments are needed for one school year's teaching and learning, multiple item pools should be constructed. For example, each unit addresses different attributes, so a new item pool may be needed for each unit to support the formative assessment when learning a unit. Considering a large number of item pools required for one school year and the high cost in item development, it is important to know the minimal size of an item pool that satisfies the purposes of a test.

This study aims to propose an item pool design method for CD-CAT so that the item pool can fully support a test. The proposed method will be applied to explore the number of items and

item types needed for an item pool for classroom formative assessments under various conditions. The item pools obtained will be evaluated in terms of their performances using with a CD-CAT algorithm.

6.2 Method for CD-CAT item pool design

The proposed method for item pool design borrowed the ideas from Veldkamp and van der Linden (2000) and Reckase (2010) for the item pool design of IRT-based CAT. The core of the method is computer simulations.

6.2.1 The minimum optimal pool

The minimum optimal pool is defined to be the smallest item pool that can provide the ideal item at each item-selection step, given the CD-CAT algorithm and test constraints. The potential item pool in the case of IRT-based CAT has an infinite number of items. A CDM-based item pool, however, has a limited number of item types defined by the q-vectors. For example, an item pool for three independent attributes (H3.1) can have seven item types. For three attribute hierarchies—H3.2, H3.3, and H3.4—there are three, four, and four item types, respectively, under the DINA model, which are listed in Table 7-Table 9.

Items within an item type only differ in item parameters. The output of the item pool design process would be the number of items needed for each item type. In the item writing process, it is difficult, if not impossible, to control the level of item parameters, which is especially true for complicated item response models. Therefore, we start with an ideal situation in item pool design, assuming all items have equally high or low quality — a high-quality condition and a low-quality condition yield a range of item numbers. The proposed method can be used for any CD-CAT algorithm and test requirement.

Below is a brief illustration of the proposed method when applied to a variable-length CD-CAT. Suppose an examinee with the true attribute profile $\boldsymbol{\alpha}_1 = (0\ 0\ 0)$ is taking a CAT measuring three linear attributes. The items are calibrated using the DINA model. We further assume that for all items the probability of the correct response interval for examinees who have mastered none of the required attributes on an item is $P(X = 1 \mid \boldsymbol{\alpha} = 0) = 0.1$ and the probability of the correct response interval for examinees who have mastered all the required attributes on an item is $P(X = 1|\boldsymbol{\alpha} = 1) = 0.9$. The first item is fixed to be $\boldsymbol{q} = (1\ 0\ 0)$. A simulation of the CAT process using the KL algorithm to select items leads to the administration of 2 items when the test terminates when the desired accuracy level is achieved, that is the largest $\pi^{(t)}(\boldsymbol{\alpha}_l) > .85$. The items administered to this examinee are summarized by item type in Table 38. Suppose anther examinee with the true attribute profile $\boldsymbol{\alpha}_2 = (1\ 0\ 0)$ takes the test, and the items used are also summarized in Table 8. Since two examinees can use the same items, a union of the two sets of items leads to an item pool for two such examinees. In other words, the maximum number of items from each item type among the examinees constitutes the number of items required for two such examinees. If a third examinee is to be simulated, the union or maximization can be taken between the set of items for the new examinee and the union obtained earlier in Table 38.

Table 38: Item distribution for two hypothetical examinees with true attribute profiles of $\boldsymbol{\alpha}_1 = (0\ 0\ 0)$ and $\boldsymbol{\alpha}_2 = (1\ 0\ 0)$ and the union of the two sets of items

Item type	$\boldsymbol{\alpha}_1 = (0\ 0\ 0)$	$\boldsymbol{\alpha}_2 = (1\ 0\ 0)$	Union/Maximum
$\boldsymbol{q}_{\{100\}}$	2	1	2
$\boldsymbol{q}_{\{010\}}$	0	4	4
$\boldsymbol{q}_{\{001\}}$	0	3	3

6.2.2 The minimum p-optimal pool

After the test is administered to more examinees, the maximum number of items selected from each item type among all examinees will eventually become stable except for a few outliers. The test is extremely long in these extreme cases.

Suppose an item pool is designed for measuring three linear attributes given a certain CD-CAT algorithm. We further assume that all candidate items are of low quality, that is, $P(X = 1|\boldsymbol{\alpha} = \mathbf{0}) = 1 - P(X = 1|\boldsymbol{\alpha} = \mathbf{1}) = 0.3$. The simulations of 1,000 examinees per attribute profile produced a distribution of item numbers for each item type. The distribution for $q_{\{100\}}$ is shown in Figure 26.

An examinee used 44 items of $\boldsymbol{q}_{\{100\}}$ in an extreme case but 95% of the simulated examinees only needed 12 items of $\boldsymbol{q}_{\{100\}}$ or fewer. The maximum numbers of items for $\boldsymbol{q}_{\{110\}}$ and $\boldsymbol{q}_{\{111\}}$ were 54 and 44, respectively. Therefore, the minimum optimal pool as defined earlier would consist of 44 items of $\boldsymbol{q}_{\{100\}}$, 54 items of $\boldsymbol{q}_{\{110\}}$, and 44 items of $\boldsymbol{q}_{\{111\}}$. However, considering the need to construct a large number of item pools and the high cost of item development, an optimal item pool becomes impractical. If we instead take the p^{th} percentile of the distribution instead of the maximum, the size of the item pool will be substantially smaller. Such an item pool is called the minimum p-optimal pool.

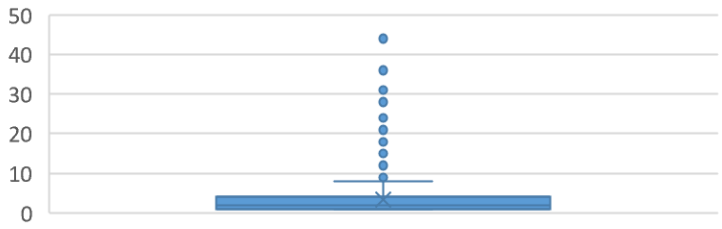


Figure 26: Distribution of the number of items for $\boldsymbol{q}_{\{100\}}$ in an example

6.3 Simulation design

Two sets of simulations will be conducted. The first set of simulations apply the proposed item pool design method to construct minimal 95-optimal pools. The second set of simulations evaluate the performances of the item pools. We consider item pools involving three attributes. The attribute hierarchies in Figure 6 are used.

Item pools are designed for the following variable-length CD-CAT. All items are calibrated by the DINA model. Following the termination rule in Hsu et al. (2013), the variable-length test is terminated at stage t when the largest $\pi^{(t)}(\boldsymbol{\alpha}_c)$ is greater than or equal to 0.90. The item selection criterion is the posterior-weighted KL index (PWKL) proposed by Cheng (2009). The index of PWKL was chosen because of its popularity and high attribute profile recovery rate (Xu, Wang & Shang, 2016). The first item in a test was fixed to be $q_{\{100\}}$ or randomly selected from the subset of q-vector for each attribute hierarchy as shown in Table 39.

Table 39: Q-vectors for the first item

Hierarchy	First item
H3.1	$\boldsymbol{q}_{\{100\}}$, $\boldsymbol{q}_{\{010\}}$, $\boldsymbol{q}_{\{001\}}$
H3.2	$\boldsymbol{q}_{\{100\}}$
H3.3	$\boldsymbol{q}_{\{100\}}$
H3.4	$\boldsymbol{q}_{\{100\}}$, $\boldsymbol{q}_{\{010\}}$

In the simulations for item pool design, item quality was held constant for the entire item pool. Two item quality levels were simulated. An item pool of high quality has $P(X = 1|\boldsymbol{\alpha} = \mathbf{0}) = 1 - P(X = 1|\boldsymbol{\alpha} = \mathbf{1}) = 0.1$. An item pool of low quality has $P(X = 1|\boldsymbol{\alpha} = \mathbf{0}) = 1 - P(X = 1|\boldsymbol{\alpha} = \mathbf{1}) = 0.3$. The minimal 95-optimal pools were constructed for both item quality levels.

For both sets of simulations, a total of 1,000 examinees were simulated for each true attribute profile. The CD-CAT algorithm described above was used on each simulated examinee.

Item responses were generated based on the DINA. A random $U(0,1)$ variable u is generated. The correct response probability $P(X_{ij}=1|\boldsymbol{\alpha})$ is compared with u to decide the response of examinee i to item j:

$$Y_{ij}=\begin{cases}1 & if\ u\le P(Y_{ij}=1|\boldsymbol{\alpha})\\ 0 & \text{otherwise}\end{cases}. \tag{49}$$

To evaluate the performance of the item pool design method, we constructed ten minimal 95-optimal pools for each hierarchy, assuming low item quality. Under each attribute hierarchy, ten designed item pools were compared with ten random item pools in terms of test length, the percent of times that the precision criterion is met, and CCR. The random item pools have the same size as the corresponding designed pool, but the Q-matrix was randomly selected from all the available q-vectors. For both designed and random item pools, item parameters $\phi_0=P(X=1|\boldsymbol{\alpha}=\mathbf{0})$ and $P(X=1|\boldsymbol{\alpha}=1)$ were generated from the uniform distribution $U(0.1,0.4)$ and $U(0.6,0.9)$, respectively.

6.4 Simulation results

The number of items needed for the minimal 95-optimal pools is shown in Table 39 for two item quality levels. The total column presents the size of the item pools. The first row of Table 40Table 40 describes the item pool designed for three independent attributes (H3.1) assuming low item quality. For example, fifteen items of $\boldsymbol{q}_{\{100\}}$ are required. The second row shows that only four items of $\boldsymbol{q}_{\{100\}}$ are required if the item quality is high.

To test the performance of the proposed item-pool design method, the designed item pools were compared with the random pools, and the statistics are summarized in Table 41. The designed pool for low item quality was used in the comparison because item parameters for this set of simulations were generated from a uniform distribution with the low item quality as a lower bound.

Table 40: The minimum 95-optimal pools

Item quality	H	$q_{\{100\}}$	$q_{\{010\}}$	$q_{\{001\}}$	$q_{\{110\}}$	$q_{\{101\}}$	$q_{\{011\}}$	$q_{\{111\}}$	Total
Low	3.1	15	15	15	10	10	10	9	84
High	3.1	4	4	4	2	2	2	2	20
Low	3.2	12			18			16	46
High	3.2	4			4			4	12
Low	3.3	13			16	17		10	56
High	3.3	4			4	4		2	14
Low	3.4	15	15		11			14	55
High	3.4	4	4		2			4	14

Table 41: Comparison between the random and designed item pools

		Test length	Modified test length	% criterion met	CCR					
Pool	H				$\boldsymbol{\alpha}_{000}$	$\boldsymbol{\alpha}_{100}$	$\boldsymbol{\alpha}_{010}$	$\boldsymbol{\alpha}_{110}$	$\boldsymbol{\alpha}_{101}$	$\boldsymbol{\alpha}_{111}$
Random	3.1	12.05	9.60	96.65	0.88	0.91	0.91	0.91	0.91	0.91
Designed	3.1	9.92	9.24	99.10	0.90	0.92	0.89	0.91	0.93	0.92
Random	3.2	6.40	5.96	98.89	0.95	0.92		0.92		0.92
Designed	3.2	6.27	5.87	99.01	0.94	0.91		0.91		0.91
Random	3.3	8.06	7.03	97.88	0.96	0.89		0.92	0.91	0.92
Designed	3.3	7.52	7.07	99.11	0.94	0.92		0.90	0.92	0.91
Random	3.4	8.02	7.11	98.07	0.91	0.92	0.92	0.91		0.91
Designed	3.4	7.45	6.97	99.00	0.93	0.92	0.93	0.90		0.91

Note: CCRs for $\boldsymbol{\alpha}_{001}$ and $\boldsymbol{\alpha}_{011}$ with H3.1 are not presented for brevity.

Take H3.1 for an example. The average test length using the random item pools was 12.05, longer than the average test length using the designed pools, which was 9.92. The difference in test length is partly due to the percent of times that the precision criterion is met. With random pools, the precision criterion was met at an average of 96.65% of the repetitions, which means 3.35% of the examinees would have to take all the items in the pool. The precision criterion was met for 99.10% of the cases on average when using the designed pools. The modified test length was calculated by excluding the cases where the precision criterion was never met. The designed pools were associated with slightly shorter tests than the random tests after excluding the extreme cases. With random or designed item pools, the average CCR for each attribute profile was close

to or higher than 0.90, which was the precision criterion. The same conclusion can be drawn for other attribute hierarchies except for H3.3, where the modified test length by using designed pools was not lower than that by using random pools.

6.5 Discussion

An important practical question is how many items are needed for a CD-CAT item pool. This type of questions belongs to the research area of the item pool design. Although numerous item selection methods have been proposed, the item pool design has been given limited attention. This study aims to guide practitioners when CD-CAT is involved. The method for item pool design is based on simulations. As Dr. Reckase noted, "there is no correct answer to the question 'How big should a CAT item pool be?'" The proposed method leads to an item pool designed for a specific CD-CAT program.

The concept of the minimum optimal pool is introduced but is deemed impractical. The minimum p-optimal pool is defined to be a practical item pool design for a formative assessment system. We then demonstrate the construction of minimum p-optimal pools for variable-length CD-CAT with two item quality levels and four attribute hierarchies. With designed item pools, the precision criterion is supposed to be met with shorter tests compared to with random item pools, which was supported by the simulation results.

Future research may consider the item pool design for fixed-length CD-CAT. Another situation worth explored is when a student would take the test multiple times (M = 1, 2, 3, 4) during an instruction period (a couple of weeks), and each time new items should be administered to a student.

The p in the minimum p-optimal pool take the value of 0.95 in this study but it could also take other values. Another variable that can be manipulated is the item quality. Currently, we

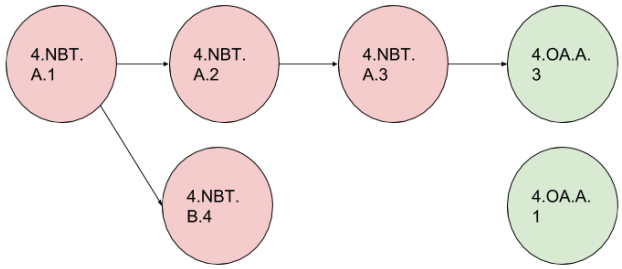
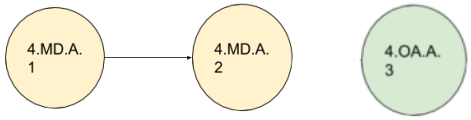
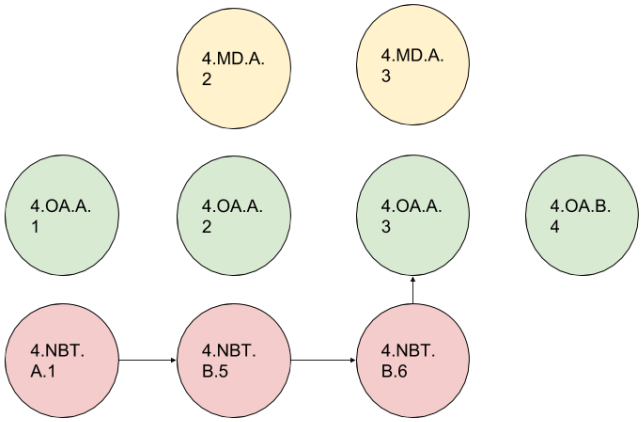
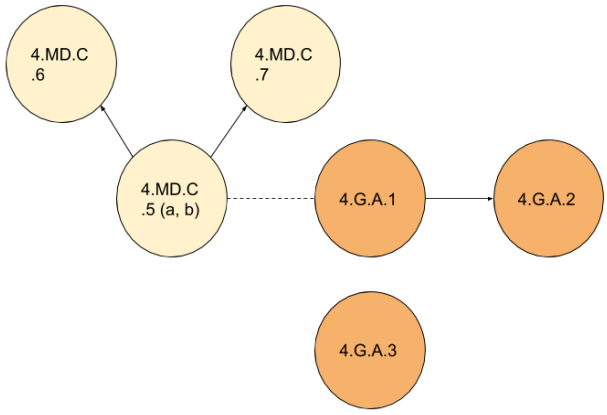
assume homogeneous item quality between item types, which is a common setting in simulation studies. However, it is possible that single-attribute items and multiple-attribute items tend to have different levels of item quality, or items involving a certain attribute have lower or higher item quality than others. Future research may take heterogeneous item quality into consideration and some practical evidence is needed regarding the item quality of different item types.

Most previous studies are built upon item pools that are calibrated using a single CDM. This study uses the DINA model. However, it is likely to observe that different items require different processes in practice, which suggests that the item pool may be made up by various CDMs (Kaplan, de la Torre, & Barrada, 2015). Recent progress in item-level model selection indices provides a theoretical basis for such item pools (Liu, Andersson, Xin, Zhang, & Wang, 2018; Ma et al., 2015). Suppose multiple-attribute items calibrated by ACDM are also included as candidate items. Item selection methods based on KL information, such as PWKL index, would always prefer a single-attribute item to a multiple-attribute item under the ACDM. The current item pool design method, therefore, would produce an item pool without any ACMD based multiple-attribute items. The optimal pool needs to be redefined with mixed models.

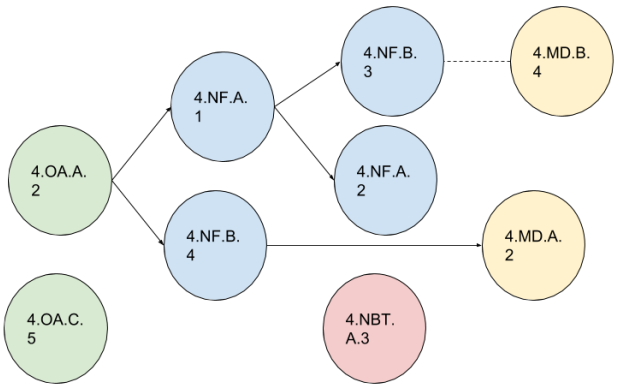
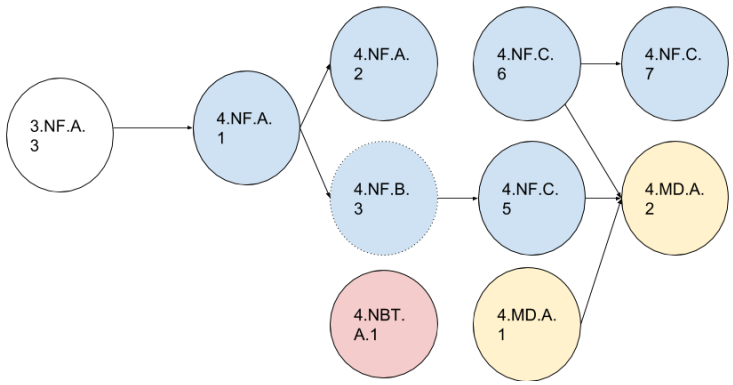
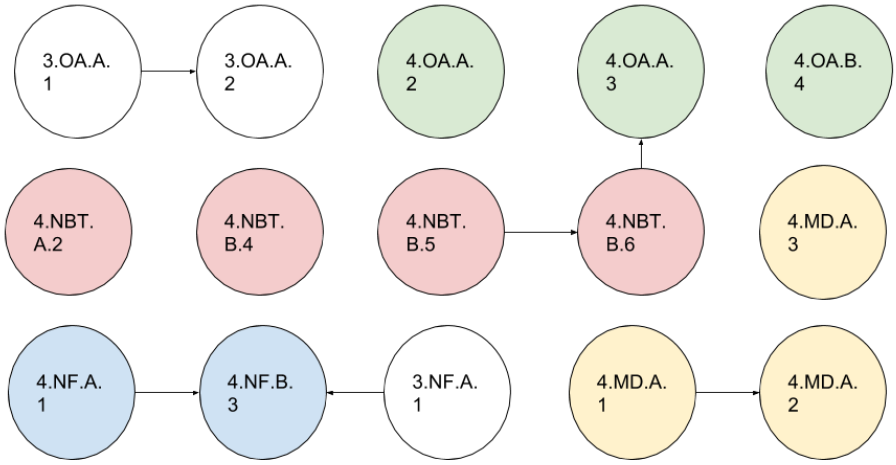
APPENDIX

APPENDIX *Hierarchies in Two Textbooks*

Eureka Math Grade 4 (2015)

Unit 1 (4 weeks) 4.OA.A.1[1], 4.NBT.A.1, 4.NBT.A.2, 4.NBT.A.3, 4.NBT.B.4, 4.OA.A.3	
Unit 2 (1 week) 4.MD.A.1, 4.MD.A.2, 4.OA.A.3	
Unit 3 (8 weeks) 4.MD.A.2, 4.MD.A.3, 4.NBT.A.1, 4.NBT.B.5, 4.NBT.B.6, 4.OA.A.1, 4.OA.A.2, 4.OA.A.3, 4.OA.A.4	
Unit 4 (3.3 weeks) 4.G.A.1, 4.G.A.2, 4.G.A.3, 4.MD.C.5, 4.MD.C.6, 4.MD.C.7	
Unit 5 (8.4 weeks)	

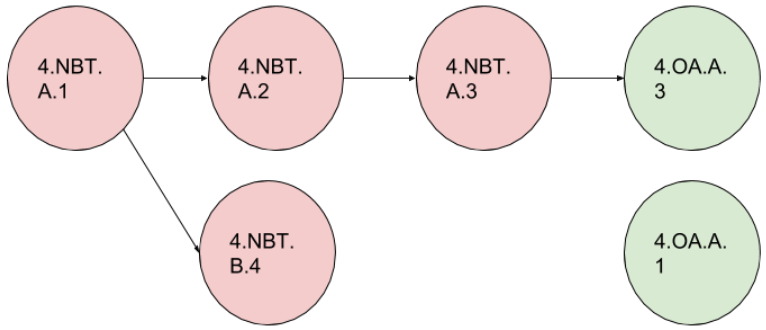
[1] 4.OA.1 is not connected with any other Grade 4 standards in the Coherence Map.

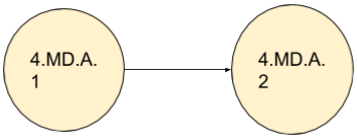
4.MD.A.2 4.MD.B.4, 4.NBT.A.3, 4.NF.A.1, 4.NF.A.2, 4.NF.B.3, 4.NF.B.4, 4.OA.A.2, 4.OA.C.5[2]	
Unit 6 (3.3 weeks) 3.NF.A.3, 4.NF.A.1, 4.NF.A.2, 4.NF.C.5, 4.NF.C.6, 4.NF.C.7, 4.MD.A.1, 4.MD.A.2, 4.NBT.A.1	
Unit 7 (3.8 weeks) 3.NF.A.1, 4.NF.A.1, 4.NF.B.3, 3.OA.A.1, 3.OA.A.2, 4.OA.A.2, 4.OA.A.3, 4.OA.B.4, 4.MD.A.1, 4.MD.A.2, 4.MD.A.3[3], 4.NBT.A.2, 4.NBT.B.4, 4.NBT.B.5, 4.NBT.B.6	

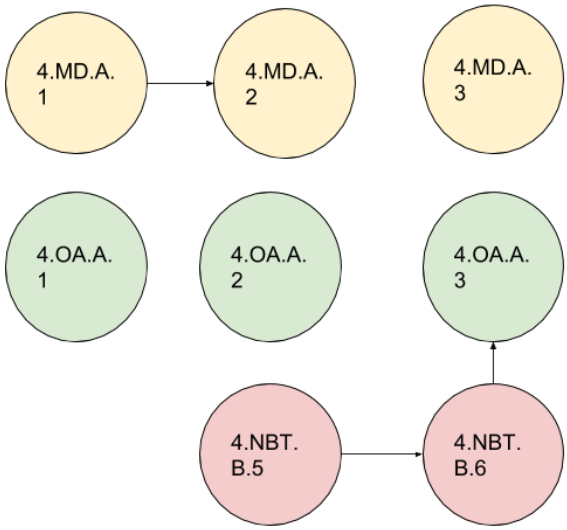
[2] 4.OA.C.5 is not connected with any other Grade 4 standards in the Coherence Map.
[3] 4.MD.A.3 is not connected with any other Grade 4 standards in the Coherence Map.

Engage NY Grade 4 (2014)

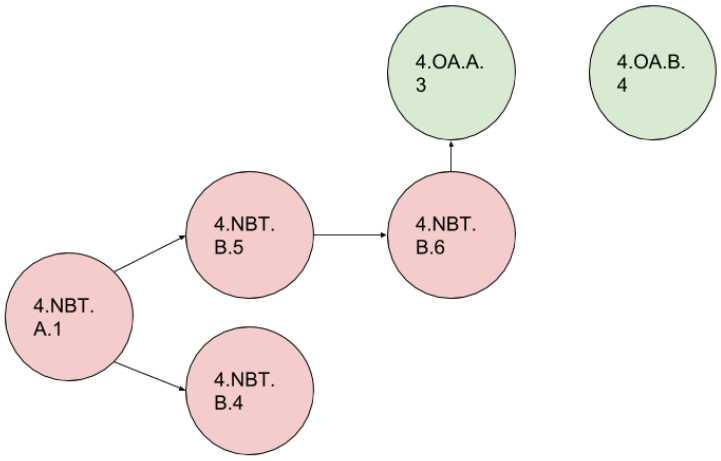
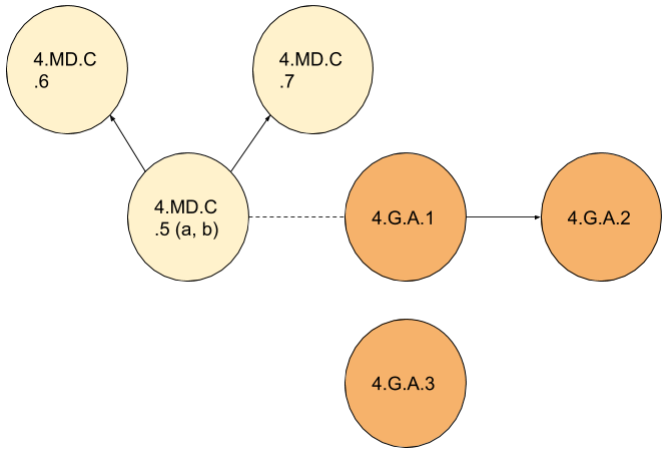
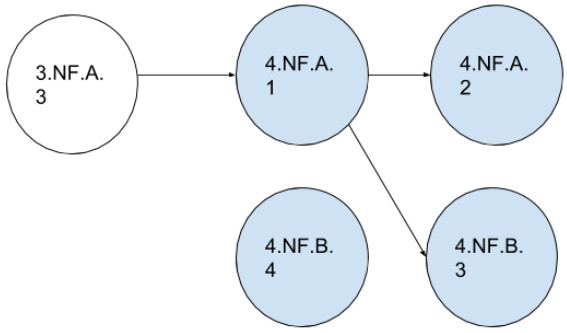
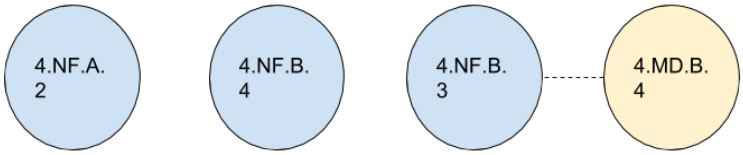
Unit 1 (4 days) 4.OA.A.1 [4] , 4.NBT.A.1, 4.NBT.A.2 Unit 3 (4 days) 4.NBT.A.3 Unit 5 (4 days) 4.NBT.A.2, 4.NBT.B.4, 4.OA.A.3	Unit 2 (2 days) 4.NBT.A.2 Unit 4 (2 days) 4.NBT.B.4, 4.OA.A.3 Unit 6 (3 days) 4.NBT.A.1, 4.NBT.A.2, 4.NBT.B.4, 4.OA.A.3
Unit 7 (3 days) 4.MD.A.1, 4.MD.A.2	Unit 8 (2 days) 4.MD.A.1, 4.MD.A.2
Unit 9 (3 days) 4.MD.A.3, 4.OA.A.1, 4.OA.A.2, 4.NBT.B.5 Unit 11 (5 days) 4.NBT.B.5	Unit 10 (3 days) 4.NBT.B.5 Unit 12 (2 days) 4.NBT.B.5, 4.OA.A.1, 4.OA.A.2, 4.OA.A.3

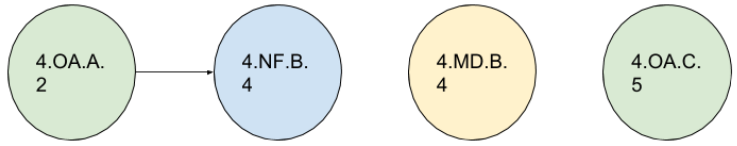
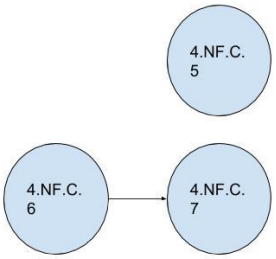
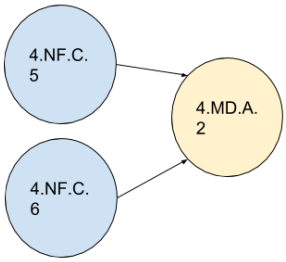
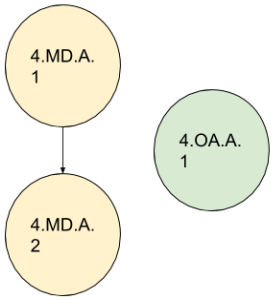
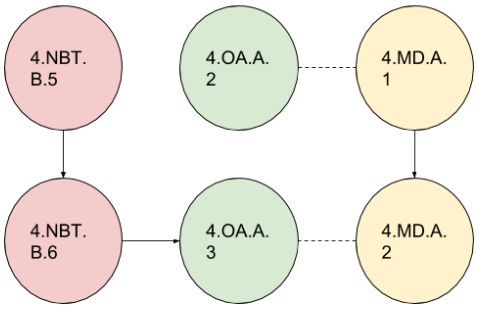






[4] 4.OA.1 is not connected with any other Grade 4 standards in the Coherence Map.

Unit 13 (9 days) 4.NBT.B.6, 4.OA.A.3 Unit 15 (9 days) 4.NBT.B.6, 4.OA.A.3, 4.NBT.B.4, 4.NBT.B.6, 4.NBT.A.1	Unit 14 (4 days) 4.OA.A.4 Unit 16 (5 days) 4.NBT.B.5	
Unit 17 (4 days) 4.G.A.1 Unit 19 (3 days) 4.MD.C.7	Unit 18 (4 days) 4.MD.C.5, 4.MD.C.6 Unit 20 (5 days) 4.G.A.1, 4.G.A.2, 4.G.A.3	
Unit 21 (6 days) 3.NF.A.3, 4.NF.B.4 Unit 23 (4 days) 4.NF.A.2	Unit 22 (5 days) 4.NF.A.1 Unit 24 (6 days) 4.NF.B.3	
Unit 25 (8 days) 4.NF.B.3, 4.NF.B.4, 4.NF.A.2, 4.MD.B.4 Unit 26 (6 days) 4.NF.B.3		

Unit 27 (6 days) 4.NF.B.4, 4.OA.A.2, 4.MD.B.4 Unit 28 (1 day) 4.OA.C.5[5]	
Unit 29 (3 days) 4.NF.C.6 Unit 30 (5 days) 4.NF.C.5, 4.NF.C.6 Unit 31 (3 days) 4.NF.C.7	
Unit 32 (3 days) 4.NF.C.5, 4.NF.C.6 Unit 33 (2 days) 4.MD.A.2	
Unit 34 (5 days) 4.MD.A.1, 4.OA.A.1, 4.MD.A.2	
Unit 35 (3 days) 4.MD.A.2, 4.OA.A.2, 4.MD.A.1, 4.NBT.B.5, 4.NBT.B.6, 4.OA.A.3	

[5] 4.OA.1 is not connected with any other Grade 4 standards in the Coherence Map.

REFERENCES

REFERENCES

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

Ayers, E., Nugent, R., & Dean, N. (2008). Skill set profile clustering based on student capability vectors computed from online tutoring data. In R. S. J. de Baker, T. Barnes, & J. E. Beck (Eds.), *Educational data mining 2008: Proceedings of the 1st international conference on educational data mining, Montreal, Quebec, Canada* (pp. 210–217). Retrieved from http://www.educationaldatamining.org/EDM2008/uploads/proc/full%20proceedings.pdf

Barnes, T. (2010). Novel derivation and application of skill matrices: The q-matrix method. In Ramero, C., Vemtora, S., Pechemizkiy, M., de Baker, R. S. J. (Eds.), *Handbook of educational data mining* (pp. 159-172). Boca Raton, FL: Chapman & Hall.

Beatty, I. D., & Gerace, W. J. (2009). Technology-Enhanced Formative Assessment: A Research-Based Pedagogy for Teaching Science with Classroom Response Technology. *Journal of Science Education and Technology, 18(*2), 146-162.

Belov, D. I., & Armstrong, R. D. (2009). Direct and inverse problems of item pool design for computerized adaptive testing. *Educational and Psychological Measurement, 69*(4), 533-547.

Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25.

Bennett, R. E. (2015). The Changing Nature of Educational Assessment. *Review of Research in Education, 39*(1), 370-407.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1-16). Westport, CT: American Council on Education and Praeger.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Black, P., Wilson, M., & Yao, S. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspectives, 9*, 71–123.

Bloom, B. S. (1968). *Learning for Mastery. Instruction and Curriculum*. Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1. Evaluation comment, 1(2), n2.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971*). Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement, 34*(3), 197-211.

Brennan, R. L. (1981). *Some statistical procedures for domain-referenced testing: a handbook for practitioners*. Iowa City, Iowa: Research and Development Division, American College Testing Program. Retrieved from https://searchworks.stanford.edu/view/1312930

Campbell, C. (2013). Research on teacher competence in classroom assessment. In J.H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 71-84). SAGE, Los Angeles.

Center for K-12 Assessment and Performance Management at ETS. (2014, March). *Coming together to raise achievement: New assessments for the common core state standards*. Retrieved from http://www.k12center.org

Chang, H.-H. (2012). Making computerized adaptive testing diagnostic tools for schools. In R. W. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 195-226). Charlotte, NC: Information Age Publishing.

Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika, 80*, 1-20.

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*(3), 213-229.

Chen, Y., Li, X., Liu, J., & Ying, Z. (2018). Recommendation System for Adaptive Learning. *Applied Psychological Measurement, 42*(1), 24-41.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*(4), 619-632.

Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: The modified maximum global discrimination index method. *Educational and Psychological Measurement, 70*(6), 902-913.

Chiu, C.Y., & Köhn, H.F. (2015), Consistency of Cluster Analysis for Cognitive Diagnosis: The DINO Model and the DINA Model Revisited. *Applied Psychological Measurement, 39*, 465-479.

Chiu, C. Y., & Douglas, J. (2013). A Nonparametric Approach to Cognitive Diagnosis by Proximity to Ideal Response Patterns. *Journal of Classification, 30*(2), 225-250.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*, 633-665.

Chiu, C. Y., Sun, Y., & Bian, Y. (2018). Cognitive Diagnosis for Small Educational Programs: The General Nonparametric Classification Method. *Psychometrika, 83*, 355-375.

Clark, I. (2016). Formative assessment: assessment is for self-regulated learning. *Educational Psychology Review, 24*(2), 205-249.

Conley, T. D. (2018). *The Promise and Practice of Next Generation Assessment*. Cambridge, MA: Harvard Education Press.

Copp, D. T. (2018). Teaching to the test: a mixed methods study of instructional change from large-scale testing in Canadian schools. *Assessment in Education: Principles, Policy & Practice, 25*(5), 468-487.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179-199.

de La Torre, J., & Karelitz, T. M. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure: A simulation study. *Journal of Educational Measurement, 46*(4), 450-469.

Ding, S. L., Luo, F., Cai, Y., Lin, H. J., & Wang, X. B. (2008). Complement to Tatsuoka's Q matrix theory. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.*), New Trends in Psychometrics* (pp. 417-423). Tokyo: Universal Academy.

Embretson, S. E. (1995). Developments toward a cognitive design system for psychological testing. In D. Lupinsky & R. Dawis (Eds.), *Assessing individual differences in human behavior* (pp. 17-48). Palo Alto, CA: Davies-Black Publishing

Embretson, S. E. (2003). *The Second Century of Ability Testing: Some Predictions and Speculations.* Princeton, NJ: Educational Testing Service. Retrievable at http:// www.ets.org/Media/Research/pdf/PICANG7.pdf.

Furtak, E. M., Circi, R., & Heredia, S. C. (2018). Exploring alignment among learning progressions, teacher-designed formative assessment tasks, and student growth: Results of a four-year study. *Applied Measurement in Education, 31*(2), 143-156.

Fyfe, E. R., & Rittle-johnson, B. (2015). Feedback Both Helps and Hinders Learning: The Causal Role of Prior Knowledge Feedback. *Journal of Educational Psychology, 108*(1), 82-97.

George, A. C., & Robitzsch, A. (2018). Focusing on Interactions Between Content and Cognition : A New Perspective on Gender Differences in Mathematical Sub-Competencies. *Applied Measurement in Education, 31*(1), 79-97.

Gierl, M.J., Leighton, J.P., & Hunka, S. (2000). Exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice, 19*, 34-44.

Gierl, M. J., & Lai, H. (2012). The role of item models in automatic item generation. *International journal of testing, 12*(3), 273-298.

Gray, R. M. (2011). *Entropy and information theory* (6th ed.). New York: Springer.

Gorin, J. S., & Mislevy, R. J. (2013). *Inherent Measurement Challenges in the Next Generation Science Standards for Both Formative and Summative Assessment.* Invitational Assessment Symposium, (September), 2-39. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.800.5350&rep=rep1&type=pdf

Gotwals, A. W. (2018). Where are we now? Learning progressions and formative assessment. *Applied Measurement in Education, 31(*2), 157-164.

Haberman, S. J. (2008). When Can Subscores Have Value? *Journal of Educational and Behavioral Statistics, 33*(2), 204–229.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333-352.

Hanna, G. S., & Dettmer, P. (2004). *Assessment for effective teaching: Using context-adaptive planning*. Boston: Pearson A and B.

Harks, B., Klieme, E., Hartig, J., & Leiss, D. (2014). Separating Cognitive and Content Domains in Mathematical Competence. *Educational Assessment, 19*(4), 243-266.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Hefling, K. (January 7, 2015). *Do students take too many tests? Congress to weigh question. Associated Press*. Retrieved from http://www.pbs.org/newshour/rundown/congressdecide-testing-schools

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*, 262–277.

Henson, R., Roussos, L., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement, 32*(4), 275–288.

Henson, R., DiBello, L., & Stout, B. (2018). A Generalized Approach to Defining Item Discrimination for DCMs. *Measurement: Interdisciplinary Research and Perspectives, 16*(1), 18-29.

Heritage, M. (2010). *Formative assessment and next-generation assessment systems: Are we losing an opportunity?* National Center for Research on Evaluation, Standards, and Student Testing (CRESST) and the Council of Chief State School Officers (CCSSO). CCSSO: Washington.

Hively, W. (1974). *Introduction to Domain-referenced Testing. In W. Hively (Ed.), Domain-referenced testing* (pp. 16-30). Englewood Cliffs, N.J.: Educational Technology Publications.

Houang, R. T. (1980). *Estimation of parameters for a latent class model applied to the study of achievement test items* (Unpublished doctoral dissertation). University of California, Santa Barbara, CA.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258-272.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New Item Selection Methods for Cognitive Diagnosis Computerized Adaptive Testing. *Applied Psychological Measurement, 39*(3), 167-188.

Kingsbury, C. G., & Zara, A. R. (1991). A Comparison of Procedures for Content-Sensitive Item Selection in Computerized Adaptive Tests. *Applied Measurement in Education, 4*(3), 241-261.

Köhn, H.-F., & Chiu, C.-Y. (2018). How to Build a Complete Q-Matrix for a Cognitively Diagnostic Test. *Journal of Classification, 35*(2), 273-299.

Kuo, B. C., Pai, H. S., & de la Torre, J. (2016). Modified Cognitive Diagnostic Index and Modified Attribute-Level Discrimination Index for Test Construction. *Applied Psychological Measurement, 40*(5), 315-330.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The Attribute Hierarchy Method for Cognitive Assessment: A Variation on Tatsuoka's Rule-Space Approach. *Journal of Educational Measurement, 41*(3), 205-237.

Luecht, R. M. (2013). Test Specifications under Assessment Engineering. *Journal of Applied Testing Technology, 14*, 1-38.

Liu, J., Xu, G., & Ying, Z. (2012). Data-Driven Learning of Q-Matrix. *Applied Psychological Measurement, 36*(7), 548-564.

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). *Assessing critical thinking in higher education: Current state and directions for next-generation assessment. RR-14-10*. Princeton, NJ: Educational Testing Service.

Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2017). The Impact of Q-Matrix Designs on Diagnostic Classification Accuracy in the Presence of Attribute Hierarchies. *Educational and Psychological Measurement, 77*(2), 220-240.

Liu, Y., Andersson, B., Xin, T., Zhang, H., & Wang, L. (2018). Improved Wald Statistics for Item-Level Model Comparison in Diagnostic Classification Models. *Applied Psychological Measurement*. https://doi.org/10.1177/0146621618798664

Ma, W., Iaconangelo, C., & de la Torre, J. (2015). Model Similarity, Model Selection, and Attribute Classification. *Applied Psychological Measurement, 40*(3), 200-217.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 33*, 379-416.

Mislevy, R. J. (2016). How Developments in Psychology and Technology Challenge Validity Argumentation. *Journal of Educational Measurement, 53*(3), 265-292.

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science, 32*, 99–113.

Nitko, A.J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, NJ: Merrill.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Reckase, M. D. (2010). Designing Item Pools to Optimize the Functioning of a Computerized Adaptive Test. *Psychological Test and Assessment Modeling, 52*(2), 127-141.

Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Schmidt, W.H., & McKnight, C.C. (1995). Surveying educational opportunity in mathematics and science: An international perspective. *Educational Evaluation and Policy Analysis, 17*(3), 337-353.

Schmidt, W., Jorde, D., Cogan, L., Barrier, E., Gonzalo, I., Moser, U., Shimizu, K., Sawada, T., Valverde, G., McKnight, C., Prawat, R., Wiley, D., Raizen, S., Britton, E. & Wolfe, R. (1996). *Characterizing pedagogical flow*. Boston MA: Kluwer Academic Publishers.

Schmidt, W.H., McKnight, C.C., Valverde, G.A., Houang, R.T., & Wiley, D.E. (1996). *Many visions, many aims: A cross-national investigation of curricular intentions in school mathematics*. Boston: Kluwer Academic.

Schmidt, W. H., McKnight, C. C., Valverde, G. A., Houang, R. T. and Wiley, D. E. (1997). *Many Visions, Many Aims: A Cross-National Investigation of Curricular Intentions in School Mathematics* (Dordrecht, The Netherlands: Kluwer).

Schutz, P. A., & Pekrun, R. (Eds.). (2007). *Emotion in education*. Burlington, MA: Academic Press.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (Vol. 1, pp. 39–83). Chicago, IL: Rand McNally

Shavelson, R.J. (2008). Guest editor's introduction. *Applied Measurement in Education, 21*(4), 293-294.

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 623-646). Westport, CT: ACE/Praeger.

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Classroom Assessment Principles to Support Learning and Avoid the Harms of Testing. *Educational Measurement: Issues and Practice, 37*(1), 52-57.

Swanson, L. & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*, 151-166.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.

Templin, J., & Bradshaw, L. (2014). Hierarchical Diagnostic Classification Models: A Family of Models for Estimating and Testing Attribute Hierarchies. *Psychometrika, 79*(2), 317-339.

Thissen, D., Reeve, B. B., Bjorner, J. B., & Chang, C. H. (2007). Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research, 16*(SUPPL. 1), 109-119.

Tu, D., Wang, S., Cai, Y., Douglas, J., & Chang, H. (2018). Cognitive Diagnostic Models With Attribute Hierarchies: Model Estimation With a Restricted Q-Matrix Design. Applied Psychological Measurement. https://doi.org/10.1177/0146621618765721

U.S. Department of Education. (2014). *Secretary's final supplemental priorities and definitions for discretionary grant programs*. Retrieved from https://www.federalregister.gov/articles/2014/ 12/10/2014-28911/secretarys-final-supplemental-priorities-and-definitions-for-discretionary- grant-programs#h-28.

U.S. Department of Education. (2015). *Fact Sheet: Testing Action Plan*, Washington, D.C.

van Der Linden, W. J. (2005a). A Comparison of Item-Selection Methods for Adaptive Tests with Content Constraints. *Journal of Educational Measurement, 42*(3), 283-302.

van der Linden, W. J. (2005b). *Linear models for optimal test design*. New York: Springer.

van der Linden, W. J., & Diao, Q. (2014). Using a universal shadow-test assembler with multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 101-118). New York, NY: CRC Press.

van der Linden, W. J., & Reese, L. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*, 259-270.

von Davier, M. (2005). *A general diagnostic model applied to language testing data, ETS Research Report RR-05-16*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets. org/Media/Research/pdf/RR-05-16.pdf

Walsh, B. (November 3, 2017). *When Testing Takes Over: An expert's lens on the failure of high-stakes accountability tests — and what we can do to change course*. Usable Knowledge. Retrieved from https://www.gse.harvard.edu/news/uk/17/11/when-testing-takes-over

Wang, S., & Douglas, J. (2015). Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika, 80*(1), 85-100.

Wang, W., Song, L., Ding, S., Meng, Y., Cao, C., & Jie, Y. (2018). An EM-Based Method for Q-Matrix Validation. Applied Psychological Measurement, 42(6), 446–459.

Way, WD., Steffen, M., & Anderson, G.S. (1998). *Developing, maintaining, and renewing the item inventory to support computer-based testing*. Paper presented at the colloquium, Computer-Based Testing: Building the Foundation of Our Future Assessments, Philadelphia, PA, September 25-26, 1998.

Willse, J., Henson, R., & Templin, J. (2007). *Using sum scores or IRT in place of cognitive diagnosis models: Can existing or more familiar models do the job?* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Wilson, M. (2018). Making measurement important for education: The crucial role of classroom assessment. *Educational Measurement, 37*(1), 1–37.

Xu, G., & Zhang, S. (2016). Identifiability of Diagnostic Classification Models. *Psychometrika, 81*(3), 625-649.

Zimba, J. (2011). *Examples of structure in the Common Core State Standards' standards for mathematical content*. Retrieved from http://commoncoretools.me/wpcontent/uploads/2011/07/ccssatlas_2011_07_06_0956_p1p2.pdf

Zimba, J. (2015, October 29). *Coherence Map*. Retrieved from www.achievethecore.org/coherence-map