1-1-2014

# Validation of the Item-Attribute Matrix in TIMSS-Mathematics Using Multiple Regression and the LSDM

Lin Ma

*University of Denver*

Recommended Citation

Ma, Lin, "Validation of the Item-Attribute Matrix in TIMSS-Mathematics Using Multiple Regression and the LSDM" (2014). *Electronic Theses and Dissertations.* 987.

https://digitalcommons.du.edu/etd/987

# Validation of the Item-Attribute Matrix in TIMSS–Mathematics
# Using Multiple Regression and the LSDM

---

A Dissertation

Presented to

the Faculty of the Morgridge College of Education

University of Denver

---

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Lin Ma

March 2014

Advisor: Dr. Kathy E. Green

Author: Lin Ma
Title: Validation of the Item-Attribute Matrix in TIMSS–Mathematics Using Multiple
Regression and the LSDM
Advisor: Dr. Kathy E. Green
Degree Date: March 2014

## Abstract

For many cognitive diagnostic models, the item-attribute matrix (or Q-matrix) is an essential component which displays the relationship between items and their latent attributes or skills in knowledge and cognitive processes. However, it is a challenge to develop an effective Q-matrix. The purposes of the present study were (1) to validate of the item-attribute matrix using two levels of attributes (Level 1 attributes and Level 2 sub-attributes), and (2) through retrofitting the diagnostic models to the mathematics test of the Trends in International Mathematics and Science Study (TIMSS), to evaluate the construct validity of TIMSS mathematics assessment by comparing the results of two assessment booklets.

Item data were extracted from Booklets 2 and 3 for the 8th grade in TIMSS 2007, which included a total of 49 mathematics items and every student's response to every item (15,654 students and 15,935 students took Booklets 2 and 3, respectively). The study developed three categories of attributes at two levels: *content*, *cognitive process (TIMSS* or *new)*, and *comprehensive cognitive process* (or *IT*) based on the TIMSS assessment framework, cognitive procedures, and item type. At level one, there were 4 content attributes (*number*, *algebra*, *geometry*, and *data and chance*), 3 TIMSS process attributes (*knowing*, *applying*, and *reasoning*), and 4 new process attributes (*identifying*, *computing*, *judging*, and *reasoning*). At level two, the level 1 attributes were further divided into 8 content attributes (b1 ~ b8), 12 TIMSS process attributes (know_a1 ~

reas_a4), 11 new process attributes (d1 ~ d11). There was only one level of IT attributes (*multiple steps/responses*, *complexity*, and *constructed-response*). Twelve Q-matrices (4 originally specified, 4 random, and 4 revised) were investigated with eleven Q-matrix models (QM1 ~ QM11) using multiple regression based on the linear logistic test model (LLTM) and the least squares distance method (LSDM).

Comprehensive analyses indicated that the proposed Q-matrices explained most of the variance in item difficulty (i.e., 64% to 81%). The cognitive process attributes contributed to the item difficulties more than the content attributes, and the IT attributes contributed much more than both the content and process attributes. The new retrofitted process attributes explained the items better than the TIMSS process attributes. Results generated from the level 1 attributes and the level 2 attributes were consistent. Most attributes could be used to recover students' performance, but some attributes' probabilities showed unreasonable patterns. The items were adequately explained by the Q-matrices of QM5 and QM5-2 with the new process attributes. However, the analysis approaches could not demonstrate if the same construct validity was supported across booklets. The proposed attributes and Q-matrices explained the items of Booklet 2 better than the items of Booklet 3. The specified Q-matrices explained the items better than the random Q-matrices.

# Acknowledgements

I have many great people to thank for helping me successfully complete my doctoral study and this project. I sincerely thank my advisor, Dr. Kathy E. Green, for her invaluable guidance and support. Her intelligence, kindness, and overwhelming patience made the difference in my academic career. My special thanks to Dr. Duan Zhang, for her inspiration and generous support. During my graduate life, Dr. Zhang provided insight, advice, and encouragement on my research work and study. I am very thankful to Dr. Antonio Olmos-Gallo for his valuable suggestions on my dissertation. Also, I am grateful to Dr. Donald Bacon for serving as my committee chair. Meanwhile, I extend my gratitude to Dr. Thomas E. Obremski, the other faculty, and the staff at the Morgridge College of Education and at the other DU schools and departments for their great support during my study at DU. Many thanks to my friends and colleagues, Dr. Marybeth Lehto, Turker Toker, and Wei Xie, for their generous help in the matrix development and the statistical computation. I would also like to thank my parents and sisters for their enduring support and encouragement. Finally, thanks to all my friends in Denver, especially Carol Carder and Laura Compton.

## Table of Contents

## List of Tables

# List of Figures

# Chapter One

## Introduction and Literature Review

Assessments play a vital role in society. Probably, all of us have experienced numerous examinations or tests since we began school. Assessments, large or small, not only decide one's immediate educational path, but also impact one's choice of future careers and life, as well as bearing effects on social equality (Hanson, 1993; Miyazaki, 1976; Wilbrink, 1997). Because of the power of assessment in our daily lives, the pursuit of scientific and effective assessment systems is an enduring endeavor for researchers, educators, and policymakers, especially for those in the field of educational assessment. For any nation, human resources are key to development and prosperity both economically and socially. Education is the primary foundation for nurturing knowledgeable and skilled citizens (National Commission on Excellence in Education, 1983), while educational assessments, as evaluation tools of educational outcomes, are core components in the educational system. Nowadays to better appraise the development of children, intellectually, psychologically, or physically, educational assessments are administrated at all levels—international, national, regional, and local (Schmeiser, 2007).

From an historical perspective, the approaches and ends of educational assessment have changed due to progress in education systems and advances in technology and human society. Education is in many places no longer a privilege of the wealthy class. In modern society, all members are entitled to equal educational opportunities to fully develop their talents (National Commission on Excellence in

Education, 1983). For example, the progress in higher education was from being available to the elite to being available to the masses, and then to approaching universal access during the past decades (Trow, 2006). Thus, current large-scale educational assessments are not implemented only for selecting students for college or university education, such as college entry examinations, or a small group of elites for the state's civil services, such as the imperial Chinese civil service examinations or the examinations in Western Europe during the 18th and 19th centuries (Hanson, 1993; Wilbrink, 1997). For educational assessments employed in present K-12 education in U.S. and many countries around the world, the main goals are to display the students’ learning status, to provide feedback for learning and instruction, to collect information for educational policies, and finally, to raise student achievement to higher levels (Linn, 2006, 2010; Nichols, 1994) .

In the 1960s, the National Assessment of Educational Progress (NAEP) was launched to periodically monitor student’s academic achievement (Jones & Thissen, 2007). During the 1970s and 1980s, under educational systems deemed to be unsatisfactory, “both the top-down accountability and the bottom-up instructional perspectives” (p. 4) demanded expanding educational assessments, for both educational policy-making and for classroom instruction (Linn, 1993). As a result, since the 1980s, standardized tests have been employed by more and more states in their educational accountability systems (Linn, 2006, 2010; Nichols, 1994). This trend was strengthened by the passage of the No Child Left Behind Act of 2001 (NCLB: U.S. Department of Education, 2001). To improve the performance of U.S. primary and secondary schools, NCLB required all states to have compulsory achievement tests for multiples subjects at different grades (Pellegrino, 2004). However, examinations are not ultimate goals; they "are supposed to be a means to an end, that end being learning" (Hanson, 1993, p. 218). Increasing assessment does not necessarily lead to the improvement in students' achievement. As Pellegrino (2004) pointed out, "weighing the pig won't cause it to grow—you still have to feed it" (p. 5). Only when educational assessments become integrated components of learning and instruction can they exert significant impact on improving students' academic performance (Pellegrino, 2004). Thus, design and implementation of more reliable and effective assessments are still challenges in the educational assessment community. Educators expect "to develop a more efficient and effective diagnostic testing model that provides technically sound information about student achievement" and that can "facilitate differential instruction to individual students or groups of students" (Schmeiser, 2007, p. 1118).

Technically, modern educational assessments are rooted in the anthropometric testing and intelligence testing of the 1880s (Hanson, 1993; Jones & Thissen, 2007; Wilbrink, 1997). Frances Galton's anthropometric testing and then James McKeen Cattell's mental tests focus more on individual physical differences, while Binet and Simon's intelligence testing was intended to assess higher cognitive abilities such as child's mental age. Following them, numerous psychologists and statisticians contributed to the development and popularity of measurements, and their application with the U.S. Army and with educational systems. As in every scientific and social discipline, the long-term development of educational assessments needs a sturdy theoretical foundation. Louis Leon Thurstone, as a measurement pioneer, explored underlying theories of measurement and the creation of reliable tests in the 1930s (Jones & Thissen, 2007).

Before the 1970s, classical test theory (CTT), which uses raw scores, dominated the field of measurement research. But, the test scores based on the CTT are dependent on test items and examinees, and so are not invariant. Item response theory (IRT) provides invariant estimates for both test item and examinee ability (Hambleton, 1993). In the 1970s and 1980s, substantially increasing numbers of research studies employed IRT with practical assessment issues (Linn, 1993). However, although IRT has more advantages than CTT, only a single ability of an examinee is estimated in a test based on a summative form. IRT models still “have little connection with the concerns of cognitive theory about the processes, strategies, and knowledge structures that underlie item solving” (Embretson, 1993, p. 125). Numerous studies pointed out that assessments lacked cognitive foundations (Snow & Lohman, 1993). Since the 1980s, educational researchers have increasingly explored the integration of cognitive psychology with psychometrics to create better diagnostic testing and measurements (Mislevy, 2010; Nichols, 1994; Sheehan & Mislevy, 1990; Snow & Lohman, 1993). Growing interest in cognitive diagnostic assessment (CDA) concerns validity issues in educational assessment and a way to create efficient diagnostic tools for improving class instruction and students’ learning (Chen, 2006).

Meanwhile, the rapid advances in science and technology not only created more advanced computing tools, but also demanded more precise information for measurement. As studies in the science fields such as chemistry, biology, and physics, research in educational assessment reached to a relative “micro-level” with the aid of advanced technical tools (K. K. Tatsuoka, 2009). Following the relatively coarser-grained scores, such as CTT overall score and IRT score, educational researchers investigated educational assessments that can provide cognitive information at a finer level of grain size, such as mastery of latent attributes or sub-skills, and cognitive processes that affect students' acquisition of knowledge. As a result, cognitive diagnostic models (CDMs) are "designed to measure specific knowledge structures and processing skills in students so as to provide information about their cognitive strengths and weaknesses" (Leighton & Gierl, 2007, p. 3). Also, the study of cognitive diagnostic assessment was impelled by the enactment of NCLB (U.S. Department of Education, 2001), which called for interpretive, descriptive, and diagnostic assessment reports for individual students (U.S. Department of Education, 2003).

According to Fu and Li (Fu, 2005; Fu & Li, 2007), there are over sixty types of CDMs, for binary, polytomous, or continuous response items, for compensatory or non-compensatory attribute structures, and for unidimensional or multidimensional latent traits. For many CDMs, the item-attribute matrix (or attribute matrix, or Q-matrix) is the quintessential component "because it represents the operationalization of the substantive theory that has given rise to the design of the diagnostic assessment" (Rupp, Templin, & Henson, 2010, p. 49). So, the correct specification of the relationship between items and attributes is a crucial step in the analyses with a CDM (Baker, 1993; de la Torre, 2008; Im, 2007; Im & Corter, 2011; Rupp & Templin, 2008; Rupp et al., 2010). However, relatively few studies have investigated validation of the item-attribute matrix. As Liu, Xu, and Ying (2011a) pointed out, "Despite the importance of the Q-matrix in cognitive diagnosis, its estimation problem is largely an unexplored area" (p. 2). Although in the past five years, researchers have conducted some studies on validating the Q-matrix, there are still no agreed-upon standards for examining the attribute matrix. The validation of the attribute matrix needs to be further examined to enable reliable inferences from a CDM. Thus, the present study explored validation of the item-attribute matrix using test item data from the Trends in International Mathematics and Science Study 2007 (TIMSS 2007).

TIMSS is one of a few international large-scale studies. Modern communications make the earth a global village with transparent borders. The world calls for more collaboration than ever before, although global competition still exists among countries. Educators extend their lenses beyond their own countries. They concern themselves with education and human development in the whole world, especially development of young people—the future of the world. Their concerns led to the development of international large-scale assessments, such as the Progress in International Literacy Study (PIRLS), Programme for International Student Assessment (PISA), and TIMSS. The primary goals of these international assessments are to collect solid information on trends in students' achievement in mathematics, science, and reading literacy, or in the essential knowledge and skills needed in adult life, and also comprehensive background information that could affect their performance; to provide guidance for educational policy-making and practices; and ultimately to improve the quality and equity of education in national as well as international contexts (Mullis, Martin, & Foy, 2008; Mullis et al., 2005b; Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009; Organization for Economic Co-operation and Development [OECD], 2010).

TIMSS is the largest international educational study, involving more than 60 countries around the world (Mullis et al., 2005b, 2009). The study is supervised by the International Association for the Evaluation of Educational Achievement (IEA). It is designed to monitor fourth and eighth graders’ achievement in mathematics and science over time. The assessment content is in line with mathematics and science curricula in the participating countries and is intended to provide information important to evaluating the effectiveness of curricular and instructional methods. As the largest international comparative assessment, TIMSS results exert significant influence on the participating countries’ educational policies and educational improvement all around the world. Thus, the validation of the TIMSS assessment is extremely important. The present study focuses on the eighth graders’ mathematics assessment in TIMSS 2007, the latest released assessment.

According to the design framework, TIMSS is intended to better understand what knowledge and skills students should learn at school and the degree to which students have mastered that knowledge and those skills, that is, students’ particular strengths and weaknesses. Since the first wave of TIMSS in 1995, students’ overall mathematics achievement and achievement in the content domains (e.g., number, algebra, geometry, and data and chance) have been reported. To meet increasing needs for students’ performance information in cognitive processes, the subscores of three cognitive domains (knowing, applying, and reasoning) were required to be reported in TIMSS 2007. The above review shows that both TIMSS and CDA have similar study purposes, that is, to provide fine-grained diagnostic feedback that can facilitate instruction and learning. Thus, cognitive diagnostic methods were applied in the current study to explore achievement scales in TIMSS mathematics assessment.

## Research Purposes and Research Questions

The purposes of this study were (1) to validate the item-attribute matrix using two levels of attributes (Level 1 attributes and Level 2 sub-attributes), and (2) through retrofitting the diagnostic models to the TIMSS test, to evaluate the construct validity of TIMSS mathematics assessment by comparing the results of two assessment booklets in TIMSS 2007. The matrix of mathematics attributes was cross-validated by two cognitive methods: multiple regression based on Fischer's linear logistic test model (LLTM, 1973) and Dimitrov's least squares distance method (LSDM, 2007). These two approaches are IRT-based methods, and assume that all latent attributes are non-compensatory, that is, all attributes must be mastered to solve an item correctly.

Moreover, the current study examined attributes for both mathematical content or knowledge and cognitive processes involving solving mathematics problems. According to Embretson (1983), an effective "cognitive model should diagnose attributes on knowledge and skills as well as cognitive processes/mechanisms" (p. 180). For eighth-grade mathematics, the content domains are fairly consistent and the content attributes are easily specified. However, the attributes for cognitive processes are relatively difficult to identify. If every step of the cognitive process is specified, there are many parameters which would likely be unstable. If cognitive process attributes are summarized in a very broad way, insufficient feedback would be provided for instruction and learning. TIMSS 2007 reported the subscores for the four content domains and three cognitive domains. Therefore, this study explored cognitive attributes with appropriate grain size for both domains. (***Note*****:** In this study and other CDM studies, cognitive attributes include attributes of both content and cognitive domains—content attributes and cognitive process attributes. In TIMSS, cognitive domains only refer to cognitive processes/behaviors such as knowing, applying, and reasoning.)

The following research questions were addressed in this study.

(1) Which type of attributes contributes more to item difficulty: content, cognitive process, or complex cognitive process (item type)?

(2) Do the new cognitive process attributes provide an explanation of the items? What different results were found between the TIMSS cognitive process attributes and the new cognitive process attributes? (The TIMSS cognitive process attributes (*knowing*, *applying*, and *reasoning*) were developed based on the TIMSS assessment framework, while the new cognitive process attributes (*identifying*, *computing*, *judging*, and *reasoning*) were based on hypothesized cognitive procedures in solving the test items.)

(3) What differences are generated from the level 1 attributes and the level 2 attributes? (The level 2 attributes were the sub-attributes of the level 1 attributes)

(4) Are the attributes of the two levels appropriate for recovering the students' mathematics achievement?

(5) What attributes combined into a Q-matrix can adequately explain the TIMSS mathematics test?

(6) Do the two booklets hold  the same construct validity in mathematics assessment?

The current study contributes to the following aspects of assessment. First, the relationship between items and attributes is the key component for many cognitive diagnostic models. Through different methods and indexes, this study sheds additional light on the process for creating and testing a reliable attribute matrix. Second, this study can provide useful information for constructing ideal test items in mathematics assessments, especially for TIMSS, in terms of knowledge content and cognitive processes, as well as item type (e.g., multiple-choice response or open response, wording, and length of question). As the biggest international series of studies, every wave of TIMSS assessments relies on numerous efforts, collaboration, time, and funding for designing, developing, and administering the tests (Gonzales et al., 2008; Mullis et al., 2005b, 2008, 2009). It is an ambitious effort for TIMSS to offer scientific mathematics tests and feedback for instruction and learning. Third, the study also generates valuable results that help better understand students' strengths and weaknesses in mathematics around the world.

## Literature Review

The following review summarizes literature related to four topics: (1) cognitive diagnostic assessment, (2) the item-attribute matrix, (3) cognitive diagnostic models—LLTM and LSDM, and (4) TIMSS mathematics assessment. The first section introduces the motivations and advantages of developing cognitive diagnostic assessments, common CDMs, and the importance of the Q-matrix in CDMs. The second section addresses definition and selection of attributes and the challenges in building attribute matrices, and reviews studies of the item-attribute matrix, such as misspecification and validation of the Q-matrix. The third section describes the CDMs related to the present study: the LLTM and LSDM. The last section of this literature review summarizes the development of

TIMSS, design of the TIMSS 2007 mathematics assessment, and studies of cognitive constructs of the TIMSS mathematics test.

### Cognitive diagnostic assessment.

The generation of CDA is a product of multiple streams of influence. The idea of mass education concerns progress of our society as a whole; also, it concerns development of individual students. The current educational system not only examines educational outcomes, but also pays attention to learning processes. Educators, parents, and students themselves want to know what knowledge has been mastered; they also want to deeply understand what cognitive skills are necessary for mastery. Moreover, the development of education seeks advanced and more reliable educational assessment, while advancement in modern science and techniques makes it possible to provide more precise measures. In addition, U.S. educational policy, specifically the enactment of NCLB (U.S. Department of Education, 2001), provided a strong impetus for developing more fine-grained assessment. Under these circumstances, CDA has attracted more researchers' and educators' interest during the past thirty years. Researchers retrofit CDMs to existing tests to search for more detailed diagnostic information. Meanwhile, they implement CDA to explore a new generation of educational assessments.

***Merits of cognitive diagnostic assessment.*** Cognitive diagnostic assessment integrates advanced knowledge in cognitive psychology and psychometrics. Compared to unidimensional measurement methods such as CTT and IRT, CDA has some potential advantages. The information derived from CDA is “interpretive and diagnostic, highly informative, and potentially prescriptive” (Pellegrino et al., 1999, p. 335). First, CDA diagnoses students' achievement and learning at a finer level of grain size from multiple dimensions within a given subject, which allows a better understanding of students' knowledge states (Dimitrov, 2007; Gierl & Leighton, 2007; Lee & Sawaki, 2009; Leighton & Gierl, 2007; Pellegrino, Baxter, & Glaser, 1999). Individual students' learning profiles are illustrated by the attributes required by each item, which cover underlying knowledge and skills, as well as cognitive processes (K. K. Tatsuoka, 2009).

Second, the fine-grained feedback provides detailed information for teaching and instructional interventions (Leighton & Gierl, 2007; Nichols, 1994). One of the main purposes of CDA is to classify students according to latent knowledge characteristics (Rupp & Templin, 2008; Rupp et al., 2010). Through students' knowledge profiles, teachers learn about students' cognitive strengths and weaknesses in specific knowledge areas, which assist them in implementing more effective instructional strategies to remedy their weaknesses (Dimitrov, 2007; Lee & Sawaki, 2009).

Third, CDA focuses on structured procedural or knowledge networks (Yang & Embretson, 2007). It holds promise for designing tests with desirable measurement and cognitive characteristics, especially for devising large-scale assessments (Dimitrov, 2007). For large-scale assessments, designing and updating test item pools always takes a great amount of resources in time, finance, and expertise. Use of cognitive diagnostic methods might reduce costs in developing test items. Studies found that CDMs have higher reliability than IRT models for tests with the same length (Templin & Henson, 2009), which indicates that students' latent abilities can be measured with fewer items. Moreover, in general, the required attributes in knowledge and skills are announced openly before a test, while test items are kept in security. In CDA, it would be relatively easier to generate numerous test items through multiple combinations of the required attributes. Motivated by cognitive response processes, some researchers have explored designing tests under CDA frameworks, such as the cognitive design system (Embretson, 1998, 2010), evidence centered design (Almond, Mislevy, Williamson, & Yan, 2011; Mislevy, Almond, & Lukas, 2003; Mislevy & Haertel, 2006; Mislevy, Steinberg, & Almond, 2003), principled test design with the attribute hierarchy method (Gierl, Alves, & Majeau, 2010), and assessment engineering task models (Dallas, Furter, Luo, & Ma, 2012; Luecht, 2002, 2012; Luecht & Dallas, 2010; Masters, 2010).

In addition, CDA brings a promising future for testing and assessment as the educational environment is changing with advanced technology. For instance, some studies tried to integrate cognitive diagnosis into computer adaptive testing (Chang, Boughton, Wang, & Zhang, 2010; Neo, 2011; Neo & Chang, 2012).

***Cognitive diagnostic models with the Q-matrix.*** The advantages of CDA persuade researchers to explore different CDMs that can extract reliable references about students' learning states. Many CDMs are IRT-based models and use a Q-matrix to specify the relationship between items and attributes required by items. The seminal CDMs might track back to Suppes' probabilistic model (1969) and its extension proposed by Spada (Dimitrov, 2007).

Suppes (1969) analyzed task performance and related operations in arithmetic problems using stimulus-response theory of finite automata. Later, Spada (1977) replaced Suppes' attribute probability term with Rasch model probability (Rasch, 1960), which allows variance in probability across persons. Both models assumed that the probability for solving a problem correctly was the product of the probabilities for successfully manipulating the operations or knowledge. Based on the same assumption, a body of *conjunctive* CDMs has been developed, which assume that successful response on an item requires mastery of all necessary knowledge and skills. In the conjunctive models, a low value on one attribute normally cannot be compensated for by a high value on another attribute. The most well-known conjunctive CDMs include the linear logistic test model (Fischer, 1973), the multicomponent and general component latent trait models (Embretson, 1984; Whitely, 1980), K. K. Tatsuoka's rule space model (1983, 1985, 1987, 1990, 1995, 2009), the noncompensatory reparameterized unified model or fusion model (DiBello, Stout, & Roussos, 1995, 2007; Hartz, 2002), the deterministic inputs, noisy ‘‘and’’ gate (DINA) model (de la Torre & Douglas, 2004; de la Torre, 2011; Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1977), the noisy inputs, deterministic ‘‘and’’ gate (NIDA) model (Junker & Sijtsma, 2001; Maris, 1999), the attribute hierarchy method (Gierl & Leighton, 2007; Gierl et al., 2007; Gierl, Zheng, & Cui, 2008; Leighton et al., 2004; Wang & Gierl, 2011), von Davier's general diagnostic model (2008), Dimitrov's least squares distance method (2007, 2010), and Rupp, Templin, and Henson's log-linear CDM (2010). The LLTM and LSDM are introduced in detail in the following section.

Also, researchers explored *disjunctive* CDMs, which assume that successful response to an item only requires mastery of at least one of the necessary knowledge and skill attributes. The disjunctive CDMs encompass the compensatory reparameterized unified model (Hartz, 2002), the deterministic input, noisy-or-gate (DINO) model and the noisy input, deterministic-or-gate (NIDO) model (Templin & Henson, 2006; Rupp et al., 2010), the disjunctive LSDM (Dimitrov, 2012), and the log-linear CDM (Henson, Templin, & Willse, 2009; Rupp et al). For a more comprehensive review and summary of

CDMs, please refer to the studies by Rupp, Templin, and Henson (2010), Fu and Li (Fu, 2005; Fu & Li, 2007), and DiBello, Roussos, and Stout (2007). They summarized CDMs in terms of attribute scale, attribute dimensionality, item type, Q-matrix incompleteness, cognitive strategy, and attribute structure. These models have been applied to or examined with simulated data or real data in the fields of mathematics, English language, computer science, biology, architecture, and gambling behaviors.

For all of the above CDMs, the item-attribute matrix is an integral component. According to Rupp, Templin, and Henson (2010), CDMs "are confirmatory multidimensional latent-variable models. Their loading structure/Q-matrix can be complex to reflect within-item multidimensionality or simple to reflect between-item multidimensionality" (p. 83). To be confirmatory models, CDMs require that the loading structure or Q-matrix be specified before data analyses. In general, cognitive diagnostic analyses involve the following main steps: identifying and defining attributes according to tests and test items, developing the Q-matrix, choosing a model and analyzing data, and providing diagnostic feedback—reporting scores and classifying examinees (Lee & Sawaki, 2009). This analysis procedure reflects the fundamental position of the Q-matrix for many CDMs. The quality of the Q-matrix highly impacts whether CDMs can be implemented appropriately, whether the parameters can be estimated correctly, and whether the results are reliable. However, many studies or analyses with CDMs assume that the Q-matrix is correct or fixed. Also relatively few studies assess the construction and validation of the Q-matrix.

**The item-attribute matrix.**

The attribute or Q-matrix, as the “weight” of attributes in item difficulty, was first introduced in the study of the linear logistic test model by Fischer (1973). DiBello et al. (2007) pointed out “LLTM’s use of a Q matrix as an historically pivotal development that lies mid-way between unidimensional models such as the Rasch model and fully diagnostic models” (p.18). Later, K. K. Tatsuoka (1983, 1985, 1990, 2009) further elaborated attributes and the item-attribute matrix. Researchers from the different perspectives addressed the construction of the item-attribute matrix and its challenges and the consequences of a misspecified Q-matrix, as well as potential validation methods.

***Definition of attributes and the item-attribute matrix*.** K. K. Tatsuoka (1983, 1990, 2009), as one of most influential figures in CDA, illuminated attributes and the item-attribute matrix in detail in studies of the rule space method. *Attributes* (or *cognitive attributes*) refer to underlying knowledge and cognitive processing skills that are required to solve problems in a specific content area. Sometimes the terms “attributes” and “skills” are used interchangeably by researchers. Attribute patterns that exhibit mastery or nonmastery of attributes are defined as *knowledge states or knowledge structures*. Attributes are latent variables that can be expressed by the indicators—observed item scores. Compared to item-level responses, responses to attributes can be treated as micro-level responses. *Item-attribute matrix* (*Q-matrix* or *attribute matrix*) is an incidence matrix that displays the hypothesized relationship between items and necessary attributes. Generally, the rows of a Q-matrix represent items and the columns represent attributes. For example, a Q-matrix ($j \times k$) indicates that there are $j$ items related to $k$ attributes. The entries of a Q-matrix are 1s or 0s. For an element $q_{jk}$, “1” means that a correct response to item *j* requires attribute *k*; otherwise, it is “0”. The item-attribute matrices represent the design of assessments (Close, 2012).

The development of attributes and Q-matrices with high quality is always a challenge. It often needs the collaboration of experts in the content area fields of the test subject, psychology, educational measurement, statistics, and education practice (K. K. Tatsuoka, 2009). The experts derive a list of attributes through analyzing goals of assessments, test specifications, and test items, and also via studying students’ test-taking protocols and talking out loud when students are answering items. First, the selection and identification of attributes is based on the degree or level of information one wants to diagnose or report (Lee & Sawaki, 2009; Rupp et al., 2010; K. K. Tatsuoka, 2009), which means that researchers need to choose an appropriate grain size. In the studies of addition and fraction subtraction questions, K. K. Tatsuoka specified every source of erroneous rules of operations as an attribute. However, for the studies of large-scale assessments such as TIMSS and NAEP, she pointed out it is not appropriate to choose attributes at such a small grain size, which would result in unstable parameters and overly complex diagnostic feedback to teachers and students. Hartz (2002) also pointed out a tradeoff between the complexity of the Q-matrix and the accuracy of parameter estimation must be evaluated on a case by case basis. Therefore, a balance between the richness of diagnostic information and the stability of estimation is necessary. The rule space method can handle eighteen or fewer attributes for a test with 60 items (K. K. Tatsuoka, 2009); most CDMs use four to eight attributes (Rupp & Templin, 2008).

Second, the attribute definition should reflect major knowledge, skills, and thinking processes that underlie the responses on tasks. Some CDA studies focus on the test content or knowledge, such as mathematics attributes in number, algebra, geometry, and probability, or English language attributes in listening, speaking, writing, and reading. Some studies focus on cognitive processes, such as K. K. Tatsuoka's addition and fraction subtraction questions. In the analysis of patterns of students' sub-skill achievement in the Third International Math and Science Study-Revised (TIMSS-Revised) in 1999, K. K. Tatsuoka and her associates developed a set of comprehensive attributes classified into three categories: content knowledge variables, cognitive process variables, and special skill variables unique to item types (K. K. Tatsuoka, Corter, & C. Tatsuoka, 2004). Using the same attribute classifications, researchers conducted cognitive analyses on mathematics of Turkish university entrance exam and a national assessment on Turkish eighth-grade (Dogan, 2006; Ma, Çetin, & Green, 2009). For a mathematics test, the content domains are normally known, while the appropriate cognitive domains are not as easily identified as the content domains. So, one of the purposes of this study was to identify cognitive attributes with acceptable grain size, which can provide enough but not too much information about students' cognitive abilities.

***Challenges in developing the item-attribute matrix.*** "A successful skills diagnosis critically depends on high quality Q matrix development" (DiBello et al., 2007, p. 6). Through simulation analyses or studies using real data sets, researchers found that the misspecification of the item-attribute matrix can cause serious consequences such as incorrect estimation of item parameters and attribute distribution, poor model fit, misclassification of respondents, and wrong diagnostic inferences (Baker, 1993; de la Torre, 2008; de la Torre & Chiu, 2009; Fall & Templin, 2009; Henson & Templin, 2009; Im & Corter, 2011; Liu, Xu, & Ying, 2011a, 2011b; Rupp & Templin, 2008). Meanwhile, it is a difficult effort to construct a Q-matrix to reflect relatively complex cognitive processes in solving problems. “In fact, little consensus exists as to how to identify attributes on a test, how to create a good Q-matrix, and how to validate or detect the misspecification of the Q-matrix” (Lee & Sawaki, 2009, p. 184). Although some studies of the Q-matrix have been conducted, generally speaking, selection and validation of the Q-matrix is still not well understood and needs further exploration.

According to Lee and Sawaki (2009), one challenge in developing the Q-matrix is that there is a limited conceptual foundation for many commonly used Q-matrix constructions. Generally, domain or content experts identify the relationship between the attributes and test items according to a list of predefined attributes as well as their knowledge and intuitions. To a certain degree, creating a Q-matrix is a subjective process (Baker, 1993; Chen, de la Torre, & Zhang, 2012; Lee & Sawaki, 2009); and “a proposed Q-matrix by content experts’ specification may not be identical to the ‘true’ Q-matrix” (Im, 2007, p. 2). Sometimes, multiple rater methods and item statistical parameters can assist in adjusting and reducing the subjectivity of judgment. However, sometimes it is difficult to reach a consensus due to disagreements among the raters and the complexity of assessments.

Challenges in developing the Q-matrix also come from examinees’ inconsistent response behaviors as pointed out by DiBello, Stout, and Roussos (1995, 2007), which result in different response patterns from those predicted by the Q-matrices. First are the issues of *strategy* and *completeness* in the Q-matrix. Many Q-matrices are constructed based on experts’ views about students’ responses to the test items. In practice, students may use different cognitive processes or strategies from those presumed by experts. For example, Rho’s (2010) results indicated that the use of different solution strategies affected cognitive skill diagnosis. Under this situation, studying students’ protocols in test taking is a good way to validate a Q-matrix based on experts’ views (K. K. Tatsuoka, 2009). Also, some skills or cognitive processes affecting performance may not be included in the list of proposed attributes. However, only the unified model and its extensive models acknowledge that a Q-matrix does not have to include all attributes using the completeness parameter. The second challenge is lack of perfect *positivity*, which shows that a respondent who has mastered all necessary attributes fails to answer an item correctly due to slipping, while a respondent who has not mastered all necessary attributes answers an item correctly through guessing. Some cognitive diagnostic studies incorporate this factor into the models. For example, the DINA and DINO models add slip and guessing parameters at the item level; and the NIDO and NIDA models restrict slip and guessing parameters at the attribute level. The third factor is the *random error*, such as when a respondent accidentally selects a wrong response category. However, many CDMs do not consider this factor.

The study of Q-matrices and CDMs is at a very beginning stage. Ongoing debates exist about the Q-matrix of the well-known rule space model. K. K. Tatsuoka and her associates’ rule space model has been applied to hundreds of studies, including large-scale assessments such as TIMSS, NAEP, and SAT. However, some researchers pointed out flaws in K. K. Tatsuoka’s Q-matrix theory, and the Boolean descriptive function is not appropriate for estimating examinees’ knowledge state using the Q-matrix (Ding, Zhu, Lin, & Cai, 2009). Moreover, in studies of the rule space model, K. K. Tatsuoka and her associates used addition and subtraction problems to explore classification of examinees and students' cognitive processes and misconceptions involved in solving arithmetic problems (Klein, Birenbaum, Standiford, & K. K. Tatsuoka, 1981; C. Tatsuoka, 2002, 2005; K. K. Tatsuoka, 1983, 1984, 1985, 1987, 1990, 1996, 2009; K. K. Tatsuoka, Birenbaum, & Arnold, 1989; K. K. Tatsuoka, Linn, M. M. Tatsuoka, & Yamamoto, 1988; K. K. Tatsuoka & M. M. Tatsuoka, 1983, 1997). Among them, one fraction subtraction data set, which comprises the responses to 20 test items from 536 middle school students (K. K. Tatsuoka, 1984, 1990), was used by many researchers as a typical case for CDM research. K. K. Tatsuoka originally extracted seven attributes. Then, de la Torre and Douglas (2004) added one more attribute. But, besides this added attribute, their rest of the Q-matrix is not exactly the same as K. K. Tatsuoka's Q-matrix. Based on a subset or the whole of this data set and one of the two above Q-matrices, researchers evaluated the exploratory technique in finding a Q-matrix, Q-matrix validation, relative and absolute model fit, consequences of misspecification of the Q-matrix and the cognitive model, and classification consistency and accuracy, or explored various cognitive models such as higher-order latent trait models, partially ordered latent classification models, the DINA and NIDA models, general DINA model framework, a model for multiple strategies of problem solving, and the log-linear CDM (Chen et al., 2012; Close, 2012; Close, Davison, & Davenport, 2012; Cui, Gierl, & Chang, 2012; de la Torre, 2008, 2009, 2011; de la Torre & Douglas, 2004, 2008; de la Torre & Lee, 2010; DeCarlo, 2011; Henson et al., 2009; Im, 2007; Im & Corter, 2011; Rho, 2010; C. Tatsuoka, 2002; Wang, Ding, Song, & Liu, 2012). However, using a logistic version of the DINA model, DeCarlo (2011) found that the structure of de la Torre and Douglas' Q-matrix resulted in some neglected problems in the classification of examinees: for example, estimates of latent class size were too large, close to unity; and examinees who incorrectly answered all items were classified as mastering most of the skills. The disagreements about the widely known CDM and Q-matrices indicate that there still exist many unsettled issues in building a reliable Q-matrix.

***Studies of the item-attribute matrix*.** Although during the past thirty years, numerous CDA studies have been conducted and CDMs have been explored, “CDA is still in its infancy” (Leighton & Gierl, 2007, p. 3). Among CDA research projects, studies of the item-attribute matrix were relatively few in number. Fortunately, during the recent five years and especially in the past couple of years, more and more researchers realized the importance as well as the difficulty of developing a reliable Q-matrix. A summary of a thorough literature review of the studies relative to the item-attribute matrix is found in Appendix A. Studies of the item-attribute matrix focus on three aspects: impact of Q-matrix misspecification on the estimation of item and person parameters and classification of respondents; validation of the Q-matrix; and exploratory methods of developing a Q-matrix. Most of the studies investigated the consequences of Q-matrix misspecification, while a few studies explored empirical approaches to validating cognitive attributes.

*Impact of Q-matrix misspecification on model parameters and classification of respondents*. In general, studies explored three types of Q-matrix misspecification: over-specification/fit, under-specification/fit, and mixed misfit with both over- and under-specification. The over-specified Q-matrix refers to unnecessary attributes being included (the entries of 0 are incorrectly specified as 1s), which leads to noise in the parameter estimation (Kunina-Habenicht, Rupp, & Wilhelm, 2012). The under-specified Q-matrix refers to the required attributes being omitted (some entries of 1 are incorrectly replaced by 0s), which results in fewer item parameters to be estimated. Most studies assumed independence of attributes. Some studies assumed dependency between attributes (Rupp & Templin, 2008) or ordered relationships among attributes (DeCarlo, 2011; Im, 2007; Im & Corter, 2011).

Baker (1993) first conducted a systematic study of the misspecification of the Q-matrix using simulation. He investigated sensitivity of the linear logistic test model (Fischer, 1973) to misspecification of the Q-matrix, sample size, and density of the Q-matrix. The LLTM assumes that the difficulty of items is a linear combination of a smaller set of latent cognitive attributes, and the entries of the Q-matrix were first introduced as the "weights" of cognitive operations in items. Thus, the Q-matrix is crucial to accurate parameter estimation in the LLTM. Baker set up six levels of random misspecification, which were defined by the percentage of misspecified elements in an error-free Q-matrix, from 1% to 10%. This study found that a small degree of Q-matrix misspecification (1% to 3%) could lead to a large impact on the item difficulty parameters and basic parameters, while a higher misspecification (5% to 10%) seriously degraded the parameter estimation. The impact of Q-matrix misspecification was smaller with a dense Q-matrix (a matrix with more 1s, that is, solving items required more attributes) than with a sparse Q-matrix (a matrix with more 0s). Compared to the Q-matrix misspecification, the effect of sample size was quite small.

In the study of a Bayesian framework for the reparameterized unified model (RUM), Hartz (2002) evaluated the effectiveness and robustness of RUM parameter estimation and examinee classification under the conditions that the user supplied Q-matrix was less than ideal or not correct. The RUM is a diagnostic model with parameters for both the relationship between items and attributes and examinee attribute profiles. The big difference between the RUM and other CDMs, which may be the important advantage of the RUM, is that the RUM encompasses attributes that are not included in the Q-matrix. In the study, Hartz examined the Q-matrix with a weaker than ideal relationship between the attributes and the items (i.e., higher item discrimination for the relative attributes $r^*$). She found that although most of the parameters were well estimated, the estimation of the attribute completeness index ($c$) was less accurate and the power of the examinees classification was reduced. Normally, a more complex Q-matrix resulted in a less accurate estimation of the parameters. After increasing the average number of attributes per item from two to three, the parameter estimation of the 40 items was still very good; and the examinee classification rates were slightly reduced. Results indicated that for both the Q-matrices with two or three attributes per item, the RUM was an effective diagnostic model. However, when the entries of the Q-matrix became more complex, the accuracy of parameters would be impacted greatly.

Also, Hartz (2002) investigated the performance of RUM parameter estimation with three types of Q-matrix misspecification: inclusion of an extra attribute, exclusion of a required attribute, and an item specified with severely incorrect attributes. The first two types of misspecification involved change in the total number of attributes, which were different from the misspecification in some studies with only a few entries over-specified or under-specified while the total number of attributes was fixed. The results showed that inclusion of an extra attribute did not affect the item parameter estimation and examinee attribute classification; and the extra attribute was excluded from the parameter analysis.

Second, eliminating a necessary attribute from the whole Q-matrix significantly impacted the completeness indices (*c*) of the items that required the missing attribute, all of which were present in the final model and were significantly under-estimated. Because of the inclusion of the *c* parameter and the examinee ability parameter ($\theta$), the estimation of the item parameters to the other attributes were not influenced, and the attribute classification rates were slightly reduced, but acceptable. Third, when an item was completely misspecified with attributes which did not generate the item's response, only the parameters of the misspecified item were affected. Although this study indicated the robustness of RUM to the less-ideal cognitive structure and the misspecified Q-matrix, the misspecified Q-matrix, especially missing important attributes, generated less satisfactory parameters.

Im and Corter investigated the statistical consequences of attribute misspecification in the rule space method (Im, 2007; Im & Corter, 2011). They examined the impact of two types of attribute misspecification (exclusion of an essential attribute and inclusion of a superfluous attribute) and the ordered relations of an excluded/included attribute with the other attributes. The essential attributes refer to the attributes required to solve the problems, whereas the superfluous attributes mean those unnecessary attributes. A simulation study displayed the opposite impact of the two kinds of attribute misspecification on examinees' attribute mastery probabilities. Excluding an essential attribute tended to lead to underestimation of mastery probabilities, while including a superfluous attribute generally led to overestimation of mastery probabilities. The impact of attribute exclusion was higher than attribute inclusion on examinees' mastery probabilities and the classification of attribute mastery. Attribute exclusion caused lower classification consistencies of examinees' attribute mastery. Meanwhile, the impact of attribute misspecifications was mediated by the order relations among attributes. When an essential attribute in the *subset* of some remaining attributes was excluded, examinees' mastery probabilities were underestimated or remained the same; whereas when an essential attribute in the *superset* of some remaining attributes was excluded, examinees' mastery probabilities were overestimated or remained the same.

Rupp and Templin (2008) investigated the effects of Q-matrix misspecification on item parameter estimates and respondent classification accuracy for a simple but restricted CDM—the DINA model. A simulation study examined four types of Q-matrix misspecification: an underfitting, an overfitting, a balanced misfit (exchanging 0s and 1s while the overall numbers of 0s and 1s were held constant), and attribute misspecification under incorrect assumptions about attribute dependencies. The results showed that when the required attributes were incorrectly deleted, the slipping parameter was overestimated most strongly; on the other hand, when unnecessary attributes were added, the guessing parameter was overestimated most strongly. For those items that did not involve misspecified attributes, the item parameters were not affected. Also, the misspecifications in the Q-matrix resulted in misclassifying the respondents. Under the context of balanced misfit, a large number of misclassifications existed. The four kinds of misspecification indicated that omitting certain attribute combinations led to completely misclassifying the respondents with such attribute combinations. However, the degrees of misclassification for different attribute patterns were not the same. The attribute class with all attributes present or absent did not show large elevations in the misclassification rates.

Following this, Henson and Templin (2009) explored the impact of a misspecified Q-matrix on the item and examinee parameters of a complex model—the RUM (Hartz, 2002). The RUM involves five parameters: probability of a correct item response ($\pi$*) when all required attributes are mastered; reduced proportion of item response probability (or item discrimination for the relative attributes, *r**) when the $k^{th}$ attribute is not mastered; attribute completeness index (*c*); proportion of examinees mastering each attributes ($p_k$); correlation of attribute pairs ($\rho$), which lead to the correct classification rate. This study analyzed the three types of Q-matrix misspecifications: over-fit, under-fit, and a combination of both. The variation of the estimates was tested by computing the mean deviation, mean absolute deviation, and the root mean squared error of all the parameters. Also, the study examined the correct classification rates. The results showed that robustness of the RUM statistical inference to the misspecified Q-matrix heavily depended on the type of misspecifications. Specifically, when the Q-matrix was over-specified and the parameter c was included, the Q-matrix misspecification had little impact on parameter estimation; whereas the under-specified Q-matrix affected seriously most parameter estimates such as: underestimating *c*, and $p_k$, and overestimating *r**, $\rho$, and the correct classification rate. When both the over- and under- misspecifications existed in the Q-matrices, the parameter estimates were affected moderately or strongly. The results also indicated that when the attribute completeness index *c* was excluded, the errors of estimation caused by the Q-matrix misspecification were worse. Based on the findings, the researchers suggested that one should identify those required, or possibly required, attributes for each item, but being careful not to assign an attribute when much uncertainty exists.

Within a Bayesian framework, Zhang and Rupp (2009) explored how prior informative specification for item parameters compensated for a potential misspecification of the Q-matrix, especially for relatively small respondent sample size (n = 250 and 500). The study involved two types of Q-matrix misspecification: (a) items measuring two attributes were replaced by items assessing one attribute; and (b) balanced attribute misspecification (exchanging 0s and 1s while the overall numbers of 0s and 1s were held constant). The researchers compared the item and person estimates of two Bayesian DINA models, one with and another without informative priors—reasonable true values for each item parameter. The simulation analyses showed that the Q-matrix misspecification resulted in strong bias in the estimation of the relative item parameters (slipping and guessing) and person parameters (the latent class distribution, the correct classification rate across attribute patterns, the marginal correct classification rate for each attribute, and the tetrachoric correlations between attributes). More respondents were incorrectly classified into the group lacking attribute mastery.

Shu, Henson, and Willse (2010) implemented Q3 (Yen, 1984) to detect misspecification of skills/attributes in the Q-matrix within the DINA and DINO models. Q3 is a measure of local item dependence in IRT analysis. It is defined as the correlation of residuals of the pairwise items, Q3 = $r_{d1.d2}$ (d1 and d2 are the deviations of two items). “Theoretically, the covariance in the response data should be fully explained by the underlying attributes and thus Q3 is expected to be zero if the underlying attributes are correctly specified in the Q-matrix” (Shu et al., 2010, p. 9). A non-zero Q3 indicated that the attributes of the items did not explain the items well; and a large value of Q3 suggested that local item dependence was violated. In this simulation study, Shu et al.

investigated misspecification of the Q-matrix under different conditions: different response strategies (the non-compensatory DINA model and the compensatory DINO model), attribute correlation, and misspecification types (omitting or adding an attribute). The Q3 values of the misspecified Q matrices were compared with that of the random and true Q-matrices. The results showed that Q3 was a sufficient index to detect omitting an attribute but not to adding an attribute. For both the DINA and DINO models, deleting a necessary attribute led to violation of local item independence (mainly for the affected items) and deteriorated the estimation of slipping and guessing parameters; adding an unnecessary attribute had no impact on local item independence and the parameter estimates. In addition, attribute correlation was not an obvious factor mediating the relationship between the Q-matrix and Q3 and item parameter estimation. One limitation of this study is that it is still hard to evaluate the magnitude of Q3 and the degree of violation of local item independence.

Choi, Templin, and Cohen (2010) analyzed the effect of both Q-matrix misspecification and model misspecification on item and structural parameters and classification accuracy of attribute mastery with different sample sizes, from 100 to 4,000 examinees. The model misspecification refers to incorrectly constrained main or interaction effects in the DINA model, DINO model, and compensatory RUM. The Q-matrix misspecification involved underfit and overfit within the framework of log-linear cognitive diagnosis model (LCDM). Researchers used bias and root mean square error to evaluate parameter estimates, used marginal proportions of mastery profiles, Cohen’s k, and correct classification rate to assess accuracy of mastery classification. The simulation study of the Q-matrix misspecification showed that with respect to the AIC and BIC, the model with a correct Q-matrix fit better than both models with an overfitted or an underfitted Q-matrix, and the model with an overfitted Q-matrix fit much better than the model with an underfitted Q-matrix. For the item and structural parameters and mastery classification, the impact of excluding necessary attributes was much greater than that of including unnecessary attributes, while the impact of including unnecessary attributes was negligible. The analyses of the effect of sample size indicated that the larger the sample size, the better the parameter estimates and mastery classification (but not for the model misspecification). When the sample size was only 100, the attribute mastery was accurately classified in a correct model or even in a model with an overfitted Q-matrix. However, a sample size of 200 or over was necessary for selecting a correct model; and a sample size of 500 or over was needed for reasonably estimating parameters.

Using a complex simulation study with 32 data-generation conditions, Kunina-Habenicht, Rupp, and Wilhelm (2012) investigated the impact of Q-matrix misspecification (under- and over-specification, and a combination of both) and interaction effect misspecification on item and respondent parameter recovery, classification accuracy, and sensitivity of selected fit measures in log-linear diagnostic classification models. The results showed that the misspecification of interaction effects had little impact on classification accuracy, but the misspecification of the Q-matrix notably decreased classification accuracy. The misspecification of the Q-matrix also had a dramatic effect on parameter recovery of latent class distributions, correlations and attribute proportions. The item-fit indexes—mean absolute difference (MAD) and root mean square error of approximation (RMSEA) were more sensitive to the over-specified Q-matrix than to the under-specified Q-matrix. The relative model fit indexes, Akaike's information criterion (AIC: Akaike, 1974) and Bayesian information criterion (BIC: Schwarzer, 1976), were sensitive to both over- and under-specification, which is in line with the studies by Choi, Templin, and Cohen (2010) and Chen et al. (2012). In this study, the researchers also found the distributions of MAD and RMSEA values supported similar model choices as the AIC and BIC.

Within the framework of generalized DINA (G-DINA) model and six types of CDM, Chen, de la Torre, and Zhang (2012) investigated the sensitivity and usefulness of six model fit statistics under different CDM settings, with either Q-matrix misspecification, or model misspecification, or both. The model fit indices were of absolute or relative fit, including: -2 log likelihood, AIC, BIC, and the residuals based on the proportion correct of individual items, correlation of item pairs, and log-odds ratio of paired item responses. The six CDMs were the saturated/unconstrained G-DINA model, DINA model, DINO model, additive CDM, linear logistic model, and reduced reparameterized unified model. Using simulated data and real data, they found that the AIC and BIC were useful for identifying the misspecification of the Q-matrix and CDM, while the BIC performed better. Under the context of Q-matrix misspecification, the saturated G-DINA model could serve as the true model to compare models across the Q-matrices. The absolute fit statistics—the residual of the correlation of item pairs with the Fischer transformation and the residual of log-odds ratios of pair-wise item responses—were sensitive to most conditions. However, these two fit indices were not sensitive to the over-specified Q-matrices unless in a highly constrained CDM, such as DINA model.

One of goals of the CDMs is to diagnose students' strengths and weaknesses in knowledge and skills and to classify students according to their knowledge states. So, the performance of classification is important for a CDM (de la Torre, 2008; Dogan & K. K. Tatsuoka, 2008; K. K. Tatsuoka, 2009). DeCarlo (2011) found, analytically and via simulations, that some neglected problems were largely associated with the structure of the Q-matrix. He used a widely analyzed data set, the fraction subtraction data of K. K. Tatsuoka (1984, 1990), and the popular DINA model to explore the relationship between the Q-matrix and classification of examinees. Although K. K. Tatsuoka's data only included 20 simple items, DeCarlo pointed out that "it is still debatable (after 20 years) as to the correct specification of the Q-matrix" (p. 21). He employed a partly Bayesian approach to estimation, posterior mode estimation. This approach provides more reliable parameter estimates and standard errors than either maximum-likelihood estimation or parametric bootstrapping (Galindo-Garre & Vermunt, 2006), and can better deal with boundary problems in latent class analysis, such as large or indeterminate parameter estimates, or a latent class size of zero or one with large or indeterminate standard errors.

DeCarlo (2011) examined three types of the DINA model: the reparameterized DINA (RDINA) model, which is a logistic regression model with latent classes; the higher order DINA (HO-RDINA) model (de la Torre & Douglas, 2004), which assumes a hierarchical structure among the cognitive attributes; and a restricted version of the HO-RDINA (RHO-RDINA) model, with an equal discrimination parameter across all cognitive attributes. DeCarlo found some potential problems in terms of the classification of examinees and the estimated latent class sizes. For instance, the examinees that received a zero score for every item were classified as mastering seven of eight attributes; the latent class size of the higher level attribute was higher than that of the lower level attribute. The researcher argued that classification problems might also exist in other

CDMs, which would limit the utility of CDMs, because incorrect classification of respondents supplied little useful or wrong feedback for further teaching and learning. The analyses also showed that inclusion of an irrelevant skill had little effect on estimates of the class size of other attributes in the Q-matrix. However, the RDINA model (and to some extent the HO-RDINA model) clearly picked up the irrelevant attributes based on the latent class size estimates close to unity, whereas the RHO-RDINA model did not. This study provided a useful way to indentify potential misspecification in the Q-matrix.

The above studies examined the impact of the misspecified Q-matrix on the item and person parameters and classification of respondents based on the various CDMs such as the LLTM, rule space method, DINA, DINO, G-DINA, RDINA, HO-RDINA, RHO-RDINA, RUM, and log-linear diagnostic models. It is hard to generate a unified result because of the different parameters in the different models. But all studies indicated the negative impact of the misspecified Q-matrix on parameter estimation; exclusion of the required attributes had more serious impact than inclusion of unnecessary attributes.

*Validation of the Q-matrix*. Although the serious consequences of a misspecified Q-matrix in the CDMs have been realized by more and more researchers, relatively few studies addressed the validation of the Q-matrix. In early studies, researchers used the relationship between item difficulty and attributes to validate the attributes and Q-matrix (Hartz, 2002; K. K. Tatsuoka, Corter, & C. Tatsuoka, 2004). Dimitrov (2007) used the least squares distance method (LSDM) to validate the cognitive structures for a test. More validation studies of the Q-matrix were presented in the recent years. Some researchers proposed empirical approaches to testing a Q-matrix within the framework of the DINA model and generalized DINA model (de la Torre, 2008; de la Torre & Chiu 2009; Tu, Cai,

& Dai, 2012). Some studies were based on the RUM and the fusion model (Chen & Zhang, 2012; Feng & Habing, 2012). Liu, Xu, and Ying's studies (2011a, 2011b) built on rigorous mathematical analyses. Although these studies provided valuable perspectives or approaches to validating a Q-matrix, no uniform standards have been achieved for a reliable Q-matrix. However, researchers agreed that the construction of a Q-matrix should not be based only on these statistical methods. As K. K. Tatsuoka (2009) pointed out, "it is dangerous to rely on a single optimization technique for attribute selection" (p. 273). These approaches, as well as assessment theory, experts' views, and analysis of test items and examinees' protocols, should be implemented together (Close, 2012; de la Torre, 2008; Dimitrov, 2007; Hartz, 2002; K. K. Tatsuoka, 2009).

As an important bridge connecting IRT and diagnostic models, Fischer's LLTM (1973) decomposes item difficulty as a combination of the component subtasks, in which the Q-matrix was the weight of attributes for item. Based on the concept of the LLTM or a similar idea, some researchers employed the relationship between item difficulty and attributes to validate the Q-matrix, such as Hartz (2002), K. K. Tatsuoka (2009), K. K. Tatsuoka, Corter, and C. Tatsuoka (2004), Dogan and K. K. Tatsuoka (2008), and Rho (2010). In the analysis, Hartz implemented the method slightly different from K. K. Tatsuoka and her colleagues.

In developing the RUM, Hartz (2002) proposed an approach to constructing a reliable and valid Q-matrix with both substantive meaning and statistical ties to the data. Based on the LLTM, Hartz pointed out that the derived attributes should have both homogeneous item content, because of cognitive content and process similarity, and homogeneous statistical properties such as consistent difficulty. The hard homogeneous cognitive attributes were the attributes frequently occurring in the hard items (item difficulty *p*-value < 0.4); and the easy homogeneous attributes frequently occurred in the easy items (*p*-value > 0.6). The hard items must be related to at least one hard attribute, while the easy items cannot involve any hard attributes. Different from most of the attributes based on math content, Hartz specified "speededness" as one attribute affecting the examinees' performance. With respect to the relationship between the eight attributes and the item difficulty of 60 ACT math items, three attributes were identified as hard attributes, and the remaining five attributes as easy attributes. Following this, Hartz further refined the Q-matrix to maximize the power to classify the examinees through a stepwise parameter reduction algorithm. The item parameters that were not statistically significant or determining were deleted. The items without the parameter $r^*$ (item discrimination for the relative attributes) and the attributes required by less than three items were excluded from the Q-matrix. Finally, 21 of the 100 entries (1s) were eliminated from the Q-matrix. Using the refined Q-matrix, the well-estimated probability of attribute mastery and item parameters indicated strong substantive information regarding the cognitive structure; and the expected performance order indicated good classification between the masters, high non-masters, and low non-masters.

In the study of student's mathematics performance at the attribute level in the 1999 TIMSS-Revised, K. K. Tatsuoka, Corter, and C. Tatsuoka (2004) validated the Q-matrix for 23 attributes and 163 items through linear multiple regression. As was done by Scheiblechner in 1972 (as cited in K. K. Tatsuoka, 1990), researchers implemented a multiple regression of the attributes on the item difficulties; and they examined the variance ($R^2$ and adjusted $R^2$) in item difficulties (mean proportion correct values across all respondents) explained by the entries of the Q-matrix. According to K. K. Tatsuoka (2009), item difficulty “should be a function of the cognitive demands that it poses to the examinee; therefore, a valid Q matrix should explain a significant portion of the variance in item difficulties” (p. 272). A higher adjusted $R^2$ indicated that the proposed attributes well predicted the item difficulties. K. K. Tatsuoka also used a multiple regression of the attribute mastery probabilities on the total scores or the IRT scores to validate the Q-matrix. Moreover, study of the rule space method indicated that a high-quality Q-matrix should explain students’ knowledge states and generate a high classification rate of the examinees (Dogan & K. K. Tatsuoka, 2008; K. K. Tatsuoka, 2009). Later, Rho (2010) extended this multiple regression analysis to the mixed effects logistic regression analysis, in which the item difficulties were predicted by both students (random effects) and the Q matrices (fixed effects).

Dimitrov (2007) investigated the validation of cognitive attributes using the LSDM for binary items. The unique features of the LSDM were that (1) score information was not required, as long as item parameters were available through IRT item calibration; and (2) it displayed the correct probability of attributes across ability levels and individual items. With known IRT estimates of item response probability and the Q-matrix, the attribute probability was estimated, as an “intact unit” (p. 371), through the minimization of matrix norms using the Euclidean least squares distance. For each item, the probability of correct item response recovered by the LSDM was equal to the product of the probabilities for attributes, which represented the approximation of the item characteristic curve (ICC). Compared to the ICC estimated with IRT, the ICC recovery with the LSDM displayed how well the specified attributes accounted for the items across ability levels. Also, the attribute probability curves exhibit a logical pattern (The LSDM is addressed in detail in the next section).

As Dimitrov and Atanasov (2011) pointed out, using the LSDM psychometric information of both items and their attributes can be illustrated on the IRT logit scale. The LSDM can be used for identifying potential Q-matrix misspecifications, validation screening of cognitive attributes for IRT bank items before test administration, providing additional information on validation of prior tests, and comparing group performance at the sub-skill level.

de la Torre (2008) developed an empirical approach to validating a Q-matrix for the DINA model. The selection of Q-matrix was based on the discrimination index *delta* ($\delta_j$). According to de la Torre, a correctly specified Q-matrix should maximize the difference of probabilities of correct response to item $j$ ($\delta_j$) between examinees who have all the required attributes and those who do not. Because $\delta_j = 1 - s_j - g_j$, maximizing $\delta_j$ was equivalent to minimizing the sum of the slip and guessing parameters, $s_j$ and $g_j$. The items with a high $\delta_j$ differentiate highly between the examinees, while those items with a low $\delta_j$ do not. The study found that omitting required attributes dramatically increases the slip parameter and including unnecessary attributes increases the guessing parameter. de la Torre elaborated the exhaustive search algorithm and sequential search algorithm. However, these two algorithms involve demanding computations. They were not practical for a large number of attributes and real data without a clear cut-off of $\delta_j$.

Then, de la Torre (2008) proposed the sequential expectation–maximization (EM) $\delta$-method for validating the Q-matrix, which set a cut-off point ($\varepsilon$) for the minimum increment in the delta index $\delta_j$ resulting from an additional attribute. The liberal criterion, a small $\varepsilon$ (e.g., $\varepsilon = 0$), allowed more attributes to be included, which might cause an over-specified Q-matrix. A stringent criterion (e.g., $\varepsilon = 0.20$) might result in an under-specified Q-matrix. In the simulation study, the researcher investigated the proposed method using five cut-off points of $\varepsilon$ (.00, .01, .05, .10, and .20). In the analyses using real data, the proposed approach was employed to K. K. Tatsuoka's (1984, 1990) fraction-subtraction problems and the NAEP 8th grade mathematics assessment. The results showed that the sequential EM-based $\delta$-method can identify and correctly replace inappropriate Q vectors while retaining appropriately specified Q vectors. For the real data studies, the proposed method yielded useful statistical information for constructing or repudiating Q-vector specifications. The studies indicated that the proposed approach was potentially viable for validating and optimizing a Q-matrix. This study is an initial, but significant, step in examining the Q-matrix (Close, 2012).

Following this study, de la Torre and Chiu (2009) extended the sequential EM-based delta method to the generalized DINA (G-DINA) model. The G-DINA model uses a more flexible parameterization. Based on different parameterizations, the G-DINA model can be converted to a class of reduced CDMs such as DINA and DINO models, additive CDMs, linear logistic models, and reparameterized unified models. In this study, a generalization of the discrimination index $\delta_j$ was proposed, represented by index $\varsigma_j^2$. The researchers pointed out that "a correct q-vector will yield homogeneous latent groups in terms of the probability of success [within-group probabilities] and therefore will result in groups with the highest variability of probabilities of success given a parsimonious subset of attributes" (p.10). Based on the sequential search algorithm and five CDMs, a simulation study displayed that all misspecified Q-vectors were accurately identified and replaced while the correct Q-vectors were retained. The findings indicated the viability of the general index $\varsigma_j^2$ for validating the Q-matrix. However, for both indices $\delta_j$ and $\varsigma_j^2$, the researchers also pointed out that these statistical methods cannot be used in isolation, whereas they should be implemented with other methods such as information about the items, or expert knowledge, to create a more integrative framework for selecting and validating a Q-matrix.

Based on the DINA model, Tu, Cai, and Dai (2012) developed another method ($\gamma$ method) to validate the Q-matrix. The indexes for validating the Q-matrix were the slip and guessing parameters and the score differences between the groups mastering and without mastering the attributes. Specifically, when (1) the guessing value of an item was too big, greater than the critical guessing values, and (2) the item score of the group mastering attribute *k* was not significantly different from that of the group without mastering attribute *k*, that is, the effect size was less than .20, then it suggested that attribute *k* probably was a unnecessary attribute for the item, and the element “1” of the Q-matrix was changed to “0”. On the other hand, when (1) the slip value of an item was too big, greater than the critical values, and (2) the item score of the group mastering attribute *k* was significantly different from the non-mastering group, with a effect size of .20 or greater, then attribute *k* probably was a necessary attribute for the item, and the element “0” of the Q-matrix was changed to “1”. In the third situation, when both the guessing and slip were too big, two of more attribute entries for an item needed to be changed.

As the study of Q-matrix misspecification by Baker (1993), Tu et al. (2012) designed six Q-matrices with different percents of misspecified elements, from zero, 5%, to 25%. The misspecified entries of the Q-matrices based on the simulated data were randomly selected and incorrectely re-specified. The researchers supposed that there were three levels of the observed response error ratio, which were 5%, 10%, and 15% of the expected examinee response patterns. They also set up five critical values for the slip and guessing parameters, from .10 to .30. Through comparing the original Q-matrices with the modified Q-matrices, Tu et al. found that (1) there was not any modification for the error-free Q-matrix; (2) when the critical values for the slip and guessing were .20, .25, and .30, the γ method effectively improved the Q-matrix specifications; and (3) the γ method was sensitive to the wrong Q-matrices when the critical values of the slip and guessing were small, while it was not sensitive when the critical values were high: the lower critical values of the slip and guessing led to more incorrect modifications of the Q-matrix, while the higher critical values of the slip and guessing resulted in less or no modifications of the wrong Q-matrices. Also, Tu et al. (2012) compared the γ method with de la Torre's (2008) EM-based *δ*-method using the same Q-matrix. They generated the very similar results. Considering the *δ*-method was based on the complex and sequential EM computation, the γ method was relatively simpler. In addition, the γ method raised the correct ratios of cognitive diagnosis, which were measured by the marginal match ratio and pattern match ratio. The study indicated that the γ method was proved to be an effective method for validating a Q-matrix. However, the researchers pointed out that the γ method should be used with the experts' views together; the slip and guessing greater than the critical values did not mean the Q-matrix must be wrong.

Two studies of the Q-matrix validation (Chen & Zhang, 2012; Feng & Habing, 2012) were within the framework of the RUM. Feng and Habing (2012) implemented a sequential search method with the BIC to validate a Q-matrix based on the reduced RUM (Hartz, 2002), which eliminates the response probability ( $P_{C_I}(\theta_i)$) from the model. $P_{C_I}(\theta_i)$ is the correct response probability associated with the examinee's ability and the skills that were not included the Q-matrix. The researchers pointed out that a correct Q-matrix maximized the difference (δ) between the response probabilities between the examinees who mastered all the required attributes and those who did not. They first compared the δ for all single-attribute patterns and chose the attribute with the largest $\delta^{(1)}$. Then, they found the largest $\delta^{(2)}$ for all two-attribute patterns. If $\delta^{(2)} > \delta^{(1)}$, the second attribute was identified and the process was continued to search the three-attribute patterns. If $\delta^{(2)} < \delta^{(1)}$, the search process was stopped. Finally, the reduced RUMs with all possible Q-vectors were compared using the BIC. This method can find the wrong Q-vectors, but it had a low rate for correcting the misspecified Q-matrix.

Chen and Zhang (2012) investigated an iterative framework of the Q-matrix optimization based on the fusion model (Hartz, 2002; Roussos et al., 2007), which is a RUM in a hierarchical Bayesian framework. The refined process of the Q-matrix built on the estimation of item parameters such as conditional item difficulty ($\pi_i$*), item discrimination ($r_{ik}$*), and the Q-matrix completeness index ($c_i$). The elements of the Q-matrix were changed, or the ineffective items were deleted, if the items were very difficult for the assigned attributes ($\pi_i$* < .50), or if an item had a low discrimination between the groups mastering and not mastering a attribute ($r_{ik}$* > .90), or if a necessary attribute was not included in the Q-matrix ($0 \leq c_i \leq 1.50$). After multiple times of revising the Q-matrix and the items, the model fit with the final optimized Q-matrix was evaluated by comparing the observed statistics with the model-predicted statistics.

Different from the other studies of the Q-matrix, Liu, Xu, and Ying (2011a, 2011b) conducted complex and rigorous mathematical analyses of the identification of the underlying Q-matrix within the commonly used DINA model. In the study *Theory of Self-learning Q-Matrix*, Liu et al. (2011a) used a set of mathematical expressions to display basic theoretical properties of estimating the Q-matrix in the DINA mode with known slipping and guessing parameters. The researchers developed sufficient conditions for the Q-matrix to be identifiable up to an explicitly defined equivalence relation, which is regarded as a natural partition of the space of Q-matrices. They also showed the corresponding consistent estimators. In the continuous study, Liu et al. (2011b) explored the estimation of the Q-matrix when both the slipping and guessing parameters were unknown in the DINA model; then, they extended the analyses to the DINO model. They proposed a principled estimation procedure for the Q-matrix, relative parameters, and validation of a Q-matrix. Based on the rigorous theoretical proofs, Liu et al. pointed out that the proposed estimation procedure can be applied to a large class of CDMs, and it potentially serves as a principled inference tool for the Q-matrix. Liu, Xu, and Ying's studies provided a new prospective to investigate the estimation of the Q-matrix. However, it seems to be a challenge for many people to completely understand their complex mathematical theorems and methods and apply them into practices.

*Exploratory approach to developing the Q-matrix*. In the study of exploratory Q-matrix discovery procedures, Fall (2009) investigated a probabilistic estimation procedure that allowed for uncertainty in the construction of the Q-matrix. The study focused on the skills (six attributes) involved in reading comprehension. The items were taken from the subtests of three literacy assessments: the Comprehensive Adult Student

Assessment System (CASAS), the National NAEP, and the Graduate Equivalency Degree (GED). This study compared two types of Q-matrices: a *deterministic Q-matrix* and a *probabilistic Q-matrix*. The traditional deterministic Q-matrix was developed from experts' ratings the relationship of items and attributes, using "1" or "0". The probabilistic Q-matrix was created using an MCMC estimation algorithm. The prior probabilities were equal to the ratios of the number of experts endorsing an attribute on an item and the total number of experts (six); then, the elements of the probabilistic Q-matrix were estimated using a Bayesian algorithm; at each iteration, the elements with a value below .50 were specified as zero and those with a value of .50 or over were specified as one. Both types of Q-matrices were examined using the conjunctive DINA model and the disjunctive DINO model. This study analyzed the model fit, item parameters—slipping and guessing, and classification of respondents. The results from the fourteen DINA or DINO models showed that most of the models with a probabilistic Q-matrix fit better than those with a deterministic Q-matrix; and some models with a probabilistic Q-matrix reduced the slipping and guessing parameters. However, overall there was limited evidence to support that the CDMs using a probabilistic Q-matrix generated more stable parameters and more accurate classification rates. As Fall pointed out, the undesirable results partially resulted from using data that did not provide sufficient input for cognitive diagnosis of reading comprehension, such as overuse of a single attribute (even a Q-matrix with a single attribute) or over-endorsement of a single item. So, this study is a useful step in investigating construction of the Q-matrix. Further studies of the probabilistic estimation procedure are needed to provide more rigorous results, such as using diverse data and inviting more experts to rate skills.

Different from the studies of the Q-matrix based on a known number of attributes, Close, Davison, and Davenport (2012; also Close, 2012) proposed a potential exploratory technique that can be used to supplement theory in deriving the Q-matrix with an unknown number of attributes. The researchers reparameterized the skills or attributes of the DINA model as components. Then, they used a principal components analysis to search for an appropriate Q-matrix. Following a simulation analysis, they evaluated the Q-matrix exploratory method using real data, including the widely analyzed fraction subtraction data from K. K. Tatsuoka (1984, 1990) and the NAEP 2003 grade eight mathematics data. The analyses showed that when a skill is sufficient for some items, which means that solving some items need a single skill, the skill in the Q-matrix corresponds to a component in the component analysis; whereas if a skill is insufficient, which means that the skill must accompany other skills to answer a question, the skill in the Q-matrix does not correspond to a component in the component analysis. The components analysis method was suggested to be a viable approach to augmenting theory in developing a Q-matrix. The study also found that the components analysis method was useful when the test items assess narrow content and each skill set was measured by multiple items. It cannot be implemented with assessments measuring broad content domains. Like many statistical methods, this method should not be used alone, instead together with content/domain theory and item task analysis.

Based on formal concept analysis (FCA: Wille, 1984), Wang, Ding, Song, and Liu (2012) proposed an exploratory method to identify cognitive attributes. FCA, also called Galois lattice or concept lattice, is an unsupervised learning technique using formal contexts for clustering concepts and discovering knowledge (Wikipedia, n.d.). The relation between concepts and attributes in a concept lattice is like the relation between items and attributes in a Q-matrix. Wang et al. used the isomorphism relation of lattice to reveal a concise structure of items and attributes. This study assessed the effectiveness of FCA in identifying cognitive attributes for the DINA model. They implemented the FCA method to the simulated data with different sample sizes and with different levels of guessing and slipping. However, the sample sizes were too small, only 10, 20, and 30. The results of the simulation study showed that on average, the proportion of correctly identified attribute vectors was 75.69%. The lower guessing and slipping parameters and the smaller sample size, the higher proportion of correctly identified attribute vectors. The impact of noise (guessing and slipping) was relatively higher than that of sample size. Using the real data—K. K. Tatsuoka's (1990) fraction subtraction data, the study showed that on average, only 50% of the attribute vectors identified by FCA were in line with those by the experts—de la Torre and Douglas (2004). The low proportion was caused partially by the fact that some attribute vectors were indistinguishable in terms of attribute patterns. Moreover, the appropriate Q-matrix of these subtraction data was still in question by researchers (DeCarlo, 2011). This study indicated that to some degree, FCA can be applied for aiding identifying the Q-matrix. Meanwhile, it needs additional studies of the application of FCA in developing a Q-matrix.

Using the DINA and DINO models, the above three studies investigated using exploratory approaches to develop a Q-matrix, such as a probabilistic estimation procedure, a principal components analysis, and a formal concept analysis. These studies provide additional perspectives on exploring the construction of a Q-matrix. However, these tentative methods need to be further examined with more items.

**Cognitive diagnostic models.**

In the present study, validation of the Q-matrix and the TIMSS mathematics test was based on Fischer's LLTM (1973) and Dimitrov's LSDM (2007). Both the LLTM and the LSDM are conjunctive models, assuming that successful response to an item requires all relative attributes to be mastered. With respect to the theory of the LLTM, a linear multiple regression model was implemented to investigate the relationship of the Q-matrix and item difficulty. Then, using the LSDM, the probability of correct response to each item was recovered by the product of the attribute probabilities.

***The linear logistic test model.*** Fischer (1973) introduced the LLTM to illustrate the relationship of item difficulty and the components (attributes) of test items. He pointed out that "the psychological complexity of compound problems can be defined by means of a small number of basic components (= cognitive operations) in the reasoning process" (p. 361). Thus, item difficulty ($\delta_j$) can be decomposed into a weighted sum of basic parameters ($\eta_k$) and a normalization constant c, with the Q-matrix as the weight of attributes in item difficulty as in Equation 1:

$$\delta_j = \sum_{k=1}^{K} q_{jk}\eta_k + c \qquad (1)$$

The LLTM bridges cognitive diagnostic models and IRT psychometric models. It has been implemented to test the validation of the Q-matrix and the construct validity of a test. According to this theory, the present study used a set of multiple regression models to investigate how item difficulty was explained by the attributes of the Q-matrices. If a higher proportion of variance in item difficulty can be explained by a Q-matrix, that Q-matrix is more reliable.

***The least squares distance method.*** Dimitrov (2007) proposed the LSDM to estimate the probability of correct performance on each attribute using the IRT parameters and then to recover the probability of correct response to each item across ability levels. It assumes that all cognitive attributes covered by an item need to be mastered to successfully complete an item, and that performance on one attribute does not affect performance on another attribute, that is, local statistical independence for attribute performance. Thus, the correct item response probability is assumed to be equal to the product of the probabilities of correct performance on all relative attributes, as modeled in Equation 2.

$$P_{ij} = \prod_{k=1}^{K} \left[ P\left(\alpha_k = 1 \middle| \theta_i\right) \right]^{q_{jk}} \tag{2}$$

where $P_{ij}$ —probability of correct response on item $j$ at ability level $\theta_i$ (*item probability*);

$P(\alpha_k = 1 | \theta_i)$ —probability of correct performance on attribute $k$ at ability level $\theta_i$ (*attribute probability*); and,

$q_{jk}$ —entry of the Q-matrix for item $j$ and attribute $k$. $\alpha_k = 1$, indicates attribute $k$ is required by an item; otherwise, $\alpha_k = 0$.

Equation 3 is generated by taking the natural logarithm of both sides of Equation 2:

$$\ln P_{ij} = \sum_{k=1}^{K} q_{jk} \ln P\left(\alpha_k = 1 \middle| \theta_i\right) \tag{3}$$

Then,  Equation 3 is simplified to:

$$\mathbf{L} = \mathbf{QX} \tag{4}$$

where **L** is the vector with known elements $\ln P_{ij}$ ($P_{ij}$ is estimated with the IRT parameters); **Q** is the Q-matrix; and **X** is the vector with unknown elements $\ln P(\alpha_k = 1| \theta_i)$.

By minimizing the Euclidean norm of the vector $\|\mathbf{QX} - \mathbf{L}\|$, the unknown vector **X** and the least squares distance (LSD) are estimated. Then, the attribute probabilities are calculated with $P(\alpha_k = 1 | \theta_i) = \exp(\mathbf{X}_k)$. The attribute probability curves (APCs) display the locations of the probability of performance on all attributes across ability levels in a diagram. Following this, the item response probability is recovered by the product of the attribute probabilities at all selected ability levels, which represents the LSDM approximation of the item characteristic curve (ICC). The mean absolute difference (MAD) between the recovered ICC and the ICC estimated with the IRT parameters indicates the degree that the specified attributes can explain the items. The smaller MADs suggest the better the attributes account for the items.

Dimitrov presented heuristic criteria for evaluating cognitive attributes:

- The smaller the LSD in minimizing the norm ||QX−L|| at a given ability level, the better the cognitive attributes hold together (jointly for all items) at this ability level.
- The attribute probability curves (APCs) should exhibit logical and substantively meaningful behavior in terms of monotonicity, relative difficulty, and discrimination. (p. 372)
- The better the ICC recovery for an item, the better the required attributes explain the item. (p. 373)

With respect to the MAD, Dimitrov's studies suggested using the following standards to validate the recovery degree: "(a) very good ($0.00 \leq \text{MAD} < 0.02$), (b) good ($0.02 \leq \text{MAD} < 0.05$), (c) somewhat good ($0.05 \leq \text{MAD} < 0.10$), (d) somewhat poor ($0.10 \leq \text{MAD} < 0.15$), (e) poor ($0.15 \leq \text{MAD} < 0.20$), and (f) very poor ($\text{MAD} \geq 0.20$)" (p. 373). These standards were applied for this study.

### TIMSS mathematics assessment

***Overview of TIMSS.*** TIMSS is a series of international assessments that examine trends in student achievement in mathematics and science, and collect comprehensive background information that affect teaching and learning, such as educational system, schools, curricula, instruction, lessons, home contexts, and students' behaviors and attitudes (Mullis, Martin, & Foy, 2005, 2005b, 2008, 2009; National Center for Education Statistics, 2011). The projects are dedicated to providing high quality data for policy makers and educational reform, and eventually to enhancing student performance in mathematics and science through ongoing cross-national comparisons. Since the first study in 1995, TIMSS assessments have been administrated every four years by the IEA, an international cooperative of national educational research institutions and governmental research agencies that has been conducting internationally comparative studies in a wide array of subjects since 1959. So far, five waves of assessments (1995, 1999, 2003, 2007, and 2011) have been conducted. The latest released assessment data is TIMSS 2007. Both fourth and eighth graders took the assessments in mathematics and science. Also, the students, the school teachers and principals, and the National Research Coordinators of the participating countries took part in the studies of backgrounds for learning and teaching mathematics and science. Now TIMSS is the most comprehensive study of educational achievement around the world. Approximately 425,000 students from fifty-nine countries and eight benchmarking jurisdictions participated in TIMSS 2007 (Gonzales et al., 2008; Mullis & Martin, 2008; Mullis et al., 2008).

TIMSS is a collaborative product of many experts, professional groups, and institutions. To ensure reliability and validity, high quality standards were maintained in the whole process from project planning, assessment design, item development, creation of background questionnaires, sampling, test administration, data collection, scoring, data analysis, to reporting; besides, comparative validity was assessed for this international study (Mullis & Martin, 2008; Mullis et al., 2005a, 2005b, 2008, 2009). One distinct feature of TIMSS is that the assessments focus on what students have learned at school. TIMSS tests tried to cover an array of topics which align broadly with the mathematics and science curricula in the participating countries. Thus, the design of assessments is based on the TIMSS curriculum model, which consists of three components: *the intended curriculum*, *the implemented curriculum,* and *the achieved curriculum* (Mullis et al., 2005b, 2009). These broadly-defined curriculums reflect three levels of learning contexts and final educational outcomes. The intended curriculum illustrates, at the national context, what knowledge and skills students are expected to learn and how the education systems facilitate students' learning. The implemented curriculum addresses, at the school and classroom context, what is actually taught and what types of schooling and instructions directly affect student learning. The achieved curriculum describes student attitudes and what students have learned.

***TIMSS 2007 mathematics assessment for the eighth grade*.** The mathematics test of TIMSS 2007 assessed students' performance from two dimensions: *content dimension* and *cognitive dimension* (Mullis et al., 2005b). The content dimension examined the mathematical domains and subjects matter at eighth grade. There are four content domains: number, algebra, geometry, and data and chance. For each content domain, the target percentages of testing time are 30% for number, 30% for algebra, 20% for geometry, and 20% for data and chance. The content domains are further classified into several topic areas. Each topic represents a list of objectives required in the mathematics curriculum in the majority of participating countries. For example, the algebra content domain includes three major topic areas: recognizing patterns and generalizing pattern relationships, using and evaluating algebraic expressions, and generating formulas and solving linear equations (Mullis et al., 2005b, pp. 26-27). The curricula of the participating countries show that the content domains are very consistent (Mullis et al, 2005a). The content domain scales, as well as the overall mathematics scale, have been developed since the first wave of TIMSS in 1995. However, no scales were built for the cognitive domains until TIMSS 2003 was administered, while students' performance in the *performance expectations* relative to cognitive behaviors was reported using an average percentage score.

To meet the growing needs for information about students' cognitive skills and abilities in solving mathematics and science problems, TIMSS experts started to create the cognitive domain scales using the mathematics data of TIMSS 2003 (Mullis et al., 2005a, 2005b). In TIMSS 2007, the cognitive dimension examined students' abilities to know facts, procedures, and concepts, to apply knowledge learned, and to apply intuitive and inductive reasoning to solve complex and non-routine problems; that is, there are three cognitive domains: *knowing*, *applying*, and *reasoning*. Each domain consists of a set of cognitive behaviors or processes involved in responding problems correctly. For example, the *reasoning* domain involves the processing behaviors such as generalizing results in more general and more widely applicable terms, synthesizing results to produce a further result, justifying a statement, and solving non-routine problems. For each cognitive domain, the target percentages of testing time are 35% for *knowing*, 40% for *applying*, and 25% for *reasoning*.

In the cycle of TIMSS 2007, the mathematics items are well designed to balance the assessment for both content and cognitive dimensions. Each domain includes a substantial number of items. In total, the test includes 215 items and 238 score points. The items were presented in two formats, multiple-choice and constructed-response. Each type of item makes up about half of the total score points. Using IRT scaling, the psychometric characteristics of each item were evaluated in terms of item difficult and discrimination. Then, all mathematics items are assembled into 14 blocks with a good balance of assessment domain and item format. Each student completed an assessment booklet with two mathematics blocks. To measure the trends in students' performance, only 6 of 14 assessment blocks, for a total of 89 items, have been released into the public domain for use in publications, research, and teaching, while the remaining blocks were retained securely for the future assessments (Mullis et al., 2005b). In the next cycle of TIMSS, new items will be developed to take the released items' place.

***Research of cognitive constructs of the TIMSS mathematics test.*** The content domains and cognitive domains in the TIMSS mathematics assessment are in line with the attributes or sub-skills in the CDM, which indicates that the test experts tried to design the assessment based on the idea of cognitive diagnostic assessment. In creating scales for the content domains and cognitive domains, the experts conducted a comprehensive study to make sure the scales are reliable and valid (Mullis et al., 2005a). Although four content domains and three cognitive domains have been generated, as one of the most important international assessments, the cognitive structures of the TIMSS mathematics test need to be further explored in order to supply better diagnostic feedback for teaching and learning. However, few studies investigated the cognitive constructs in the TIMSS mathematics tests.

Literature review shows that besides the TIMSS assessment experts, Corter and K. K. Tatsuoka (2002) explored the attributes involving in the mathematics test items in the TIMSS-Revised 1999 for the eighth-grade. According to the TIMSS-Revised assessment framework, the test items were developed based on the *content areas* and *performance expectations* (behaviors that might be expected of students in school mathematics). Corter and K. K. Tatsuoka extended the test framework and identified 27 attributes for the 163 test items. The attributes were classified into three categories: content knowledge attributes, cognitive process attributes, and special skill attributes unique to item types. The content knowledge attributes are similar to the content domains, while the cognitive process attributes are similar to the topics of the cognitive domains in the TIMSS 2007 test framework. Different from the content and cognitive domains that are mutually exclusive, there are overlaps among the 27 attributes. Using the rule space method with these attributes, K. K. Tatsuoka and her associates compared the mathematics sub-skill achievement of eighth-graders from the U.S. Japan, Israel, and Singapore or across 20 countries (Birenbaum, C. Tatsuoka, & Xin, 2005; Birenbaum, C. Tatsuoka, & Yamada, 2004; K. K. Tatsuoka, Corter, & C. Tatsuoka, 2004); or researchers examined students' attribute mastery profile of the students from Taiwanese and Turkish (Y.-H. Chen, 2006; Dogan & K. K. Tatsuoka, 2008). The findings of students' knowledge states from these studies are helpful for providing better remedial instructions.

As the one of the largest international assessments, the TIMSS assessments play a very important role in understanding global students' perform in mathematics and science and then informing better instruction and policies to improve achievement. The reliability and validity of the test items are essential in the design of the TIMSS assessments. However, studies of the test items using a cognitive diagnostic analysis are relatively limited. Thus, this study implemented multiple regression and the LSDM to explore the knowledge and sub-skills measured by the TIMSS mathematics assessment and validity of the test across assessment booklets.

# Chapter Two
# Method

In this study, construct validity of the item-attribute matrix and the mathematics assessment in TIMSS 2007 were explored through a series of diagnostic analyses. The data were extracted from the large-scale dataset TIMSS 2007. Over 214,000 eighth-graders from 50 countries and 7 benchmarking entities of the U.S., Canada, United Arab Emirates, and Spain participated in the mathematics assessment. Each student took one of the 14 assessment booklets, which included two blocks of mathematics items and two blocks of science items in each booklet. In total, there were 14 blocks of mathematics items that were assembled rotationally in the 14 booklets. A rotated block design can maximize assessment coverage as specified in the assessment framework; also, it ensures the assessment generates reliable information about students' achievement using sufficient items while keeping student assessment burden to a minimum (Mullis et al., 2005b). Because of test security and the reuse of some items for the future waves of assessments, only six blocks of mathematics items (M01 to M05 and M07) were released to the public for use in publications, research, and teaching. Among them, five blocks of items (M01 to M05) were contained in four assessment booklets: Booklets 1 to 4, in which no secure items were included. Booklets 2 and 3 include more mathematics items than Booklets 1 and 4. So, the present study investigated only the students' responses to the mathematic items in Booklets 2 to 3. As a result, there are two sets of data for the analysis.

## Participants

To ensure the samples efficiently represented national eighth-grade populations in the participating countries, TIMSS implemented a two-stage stratified cluster design (Joncas, 2008). The first stage consisted of sampling schools selected using probability proportionate to the estimated number of eighth-graders. The second stage consisted of sampling intact classrooms within sampled schools using random sampling. However, Russia had a preliminary sampling stage—sampling regions before sampling schools because of its large population. Singapore also sampled students within sampled classrooms. According to the requirement, the sample sizes of all participating countries should be more than 4,000 students. Typically, the participating countries sampled 150 schools and one or two intact classrooms in each school. The participation rates of most countries reached acceptable levels: 85% of both the schools and students, or a combined rate of 75% (the product of schools' and students' participation rates).

The numbers of students who took Booklets 2 and 3 were 17,717 and 17,769, respectively. Only those without missing any mathematics item were included in the study. As a result, the current study used the mathematics achievement data of 15,654 students in Booklet 2 and 15,935 students in Booklet 3. The students' demographic information is shown in Table 1. With respect to the students' gender, age, and grade, the participants who took the two assessment booklets had similar characteristics. On average, half of the participants were boys and half were girls. Their average age was 14.33 years old (*SD* = .78, range from 9.75 to 18.92 years). More than 96.78% of them were eighth-graders, while the remaining 3% were ninth-graders.

Table 1

*Characteristics of the Participants*

| | | Booklet 2 | Booklet 3 |
|---|---|---|---|
| Number of Students | | 15,654 (100%) | 15,935 (100%) |
| Gender | Girl | 7,909 (50.52%) | 7,902 (49.59%) |
| | Boy | 7,745 (49.48%) | 8,033 (50.41%) |
| Grade | Grade 8 | 15,150 (96.78%) | 15,424 (96.79%) |
| | Grade 9 | 504 (3.22%) | 511 (3.21%) |
| Age (years) | Mean | 14.33 | 14.32 |
| | *SD* | 0.78 | 0.78 |
| | Range | 9.83 ~ 18.92 | 9.75 ~ 18.92 |

## Measures

### Mathematics test items.

In total, the present study investigated 49 mathematics items and their attributes. Booklets 2 and 3 comprised 31and 33 mathematics items, respectively. Students were allowed 45 minutes to complete the mathematics assessment in each booklet (Mullis et al., 2005b). Table 2 presents the number of items by item type and domain in the two booklets. In each assessment booklet, there were two blocks of mathematics items. Using a rotated block design, Blocks M02 and M03 were included in Booklet 2, and Blocks M03 and M04 in Booklet 3. Each block of items were designed with a balance of item difficulty, discrimination, item format, content domain, and cognitive domain (Ruddock et al., 2008). Because Item M042304D in Block M04 includes two questions with one score per question, it was divided into two items. Thus, there were 49 items for the attribute analyses.

There were two types of items: multiple-choice items and constructed-response items. For the multiple-choice items, students needed to select a correct answer from four or five response options. A correct response to a multiple-choice item received one score point. For the constructed-response items, students were required to construct a written response. A correct response to a constructed-response item received one or two score points. Students' overall mathematics achievement and achievement on the main content and cognitive domains were measured using IRT scale scores–plausible values (mean = 500, standard deviation = 100, ranged from 0 to 1,000). The CDMs were used in the present study to analyze binary test items with correct or incorrect responses. Thus, the one-score-point items, either multiple choice or constructed response, were automatically dichotomously scored. The constructed-response items with two-score-point items were dichotomized by treating response with partial credit as incorrect and response with full credit as correct.

Table 2

*Number of Items by Booklet*

| | Booklet 2 | Booklet 3 |
|---|---|---|
| Assessment Block | M02 & M03 | M03 & M04 |
| Total Number of Items | 31 | 33 |
| Maximum Score Points | 32 | 35 |
| Type of Items | | |
| Multiple-choice | 20 | 21 |
| Constructed-response | 11 | 12 |
| Content Domains | | |
| Number | 14 | 16 |
| Algebra | 5 | 5 |
| Geometry | 7 | 7 |
| Data and chance | 5 | 5 |
| Cognitive Domains | | |
| Knowing | 12 | 13 |
| Applying | 15 | 16 |
| Reasoning | 4 | 4 |

Table 3 displays the number of items by item type and domain in each assessment block. In total, the three assessment blocks (M02, M03, and M04) consisted of 49 items. For Block M02, one item (M042273) were excluded because no responses were reported to this item. Block M02 included 16 items, nine multiple-choice items and seven constructed-response items. Only one constructed-response item had a maximum score of two points. All 15 items in Block M03 were one-score-point items, including 11 multiple-choice items and 4 constructed-response items. Block M04 comprised 18 items, 10 multiple-choice items and 8 constructed-response items. The maximum score of two items was two. For a detailed description of all of the items, please refer to the released eighth-grade mathematics items published by the TIMSS and PIRLS International Study Center (Foy & Olson, 2009). Two examples of mathematics items are presented in Appendix B.

Students' mathematics performance was assessed from two dimensions: content dimension and cognitive dimension (Mullis et al., 2005b). The mathematics items covered four content domains (number, algebra, geometry, and data and chance) and three cognitive domains (knowing, applying, and reasoning). According to the report about the released items (Foy & Olson, 2009), the number of items that related to each domain is exhibited by booklet (Table 2) and by block (Table 3). In each assessment booklet, every domain was measured by 4 to 16 items. The items of Blocks M02 and M04, which were newly developed for TIMSS 2007, had a better balance of content and cognitive domains. These seven domains were treated as the first level of cognitive attributes.

Table 3

*Number of Items by Assessment Block*

| Assessment Block | Block M02 | Block M03 | Block M04 | Total |
|---|---|---|---|---|
| Number of Items | 16 | 15 | 18 | 49 |
| Maximum Score Points | 17 | 15 | 20 | – |
| Type of Items | | | | |
| Multiple-choice | 9 | 11 | 10 | 30 |
| Constructed-response | 7 | 4 | 8 | 19 |
| Content Domains | | | | |
| Number | 5 | 9 | 7 | 21 |
| Algebra | 5 | 0 | 5 | 10 |
| Geometry | 3 | 4 | 3 | 10 |
| Data and chance | 3 | 2 | 3 | 8 |
| Cognitive Domains | | | | |
| Knowing | 6 | 6 | 7 | 19 |
| Applying | 6 | 9 | 7 | 22 |
| Reasoning | 4 | 0 | 4 | 8 |

#### Attributes developed according to the TIMSS assessment framework.

First, the attributes were constructed with respect to the TIMSS assessment domains. Through a series of deeply analyzing 49 test items, the researcher developed 7 attributes at level one and 20 attributes at level two (see Table 4).

The seven attributes at level one (a1, a2, a3, and a4—level 1 content attributes (contL1); and a5, a6, and a7—level 1 cognitive process attributes (cogL1)) were defined with respect to the four content domains (number, algebra, geometry, and data and chance) and three cognitive domains (knowing, applying, and reasoning) listed in the TIMSS assessment framework (Mullis et al., 2005b, pp. 23-38). Each content domain includes several topics that are represented by a list of study objectives. Each cognitive domain encompasses a set of expected cognitive processes or behaviors in solving mathematics items.

To generate more detailed information about students' mathematics performance, the attributes were further identified at a finer grain size than the first level of attributes. There were 8 level 2 content attributes (contL2: b1 to b8) and 12 level 2 cognitive process attributes (cogL2: know_a1 to reas_a4). The level 2 content attributes were constructed based on the major topic areas. The level 2 cognitive process attributes were developed by classifying the three level 1 cognitive process attributes according to the four content domains.

The seven attributes at level one are described below.

*Number* (a1): Understand numbers, ways of representing numbers, relationships among numbers, and number systems. This domain includes knowledge related to whole numbers, integers, fractions, decimals, ratio, proportion, and percent.

*Algebra* (a2): Recognize and extend patterns, use algebraic symbols to represent mathematical situations, and develop equivalent expressions and solve linear equations. This domain comprises three major topic areas: patterns, algebraic expressions, and equations/formulas and functions.

*Geometry* (a3): Understand the properties and relationships of two and three-dimensional geometric figures, use measuring instruments accurately, and select and use formulas for perimeters, areas, and volumes. This domain encompasses three geometric areas: geometric shapes, geometric measurement, and location and movement.

*Data and Chance* (a4): Organize data collected, display data in graphs and charts, describe and compare characteristics of data (shape, spread, and central tendency), and understand elementary probability. This domain also covers three topic areas: data organization and representation, data interpretation, and chance.

*Knowing* (a5): Need to know the mathematical facts and properties, procedures, and concepts. This domain embraces six types of cognitive processes: (1) recall definitions, terminology, number properties, geometric properties, and notation; (2) recognize mathematical objects, shapes, numbers, and expressions; (3) carry out algorithmic procedures for add, subtraction, multiplication, and division; (4) retrieve information from graphs, tables or other sources, and read simple scales; (5) use measuring instruments and units of measurement, and estimate measures; (6) classify/group objects, shapes, numbers and expressions according to common properties, and order numbers and objects by attributes.

*Applying* (a6): Apply mathematical knowledge and conceptual understanding to create representations and solve routine problems. This domain contains five types of behaviors: (1) select an efficient/appropriate operation, method, or strategy; (2) display mathematical information and data in diagrams, tables, charts, or graphs; (3) model the problems appropriately with and equation or diagram; (4) follow and execute a set of mathematical instructions, or draw figures and shapes; (5) solve routine problems.

*Reasoning* (a7): Through logical or systematic thinking, solve multi-steps problems or non-routine problems within unfamiliar situations or complex contexts. This domain involves five types of behaviors: (1) analyze the information and questions; (2) generalize the solving processes in more general and more widely applicable terms; (3) synthesize/integrate mathematical knowledge and procedures to establish results and to produce a further result; (4) justify a statement by reference to mathematical results or properties; (5) solve non-routine problems.

Table 4

*Attributes Based on the TIMSS Assessment Framework and Item Type*

| | Content or Cognitive Domains (Level 1 Attributes) | Content Topic or Cognitive Process (Level 2 Attributes) |
|---|---|---|
| Content dimension (content attributes) | Number (a1) | Whole numbers and integers (b1) |
| | | Fractions, decimals, ratio proportion, and percent (b2) |
| | Algebra (a2) | Patterns (b3) |
| | | Algebraic expressions and equations/formulas functions (b4) |
| | Geometry (a3) | Geometric shapes (b5) |
| | | Geometric measurement and location movement (b6) |
| | Data and chance (a4) | Data organization and representation (b7) |
| | | Data interpretation and chance (b8) |
| Cognitive dimension (cognitive process attributes) | Knowing (a5) | Knowing_a1number (know_a1) |
| | | Knowing_a2algebra (know_a2) |
| | | Knowing_a3geometry (know_a3) |
| | | Knowing_a4data and chance (know_a4) |
| | Applying (a6) | Applying_a1number (appl_a1) |
| | | Applying_a2algebra (appl_a2) |
| | | Applying_a3geometry (appl_a3) |
| | | Applying_a4data and chance (appl_a4) |
| | Reasoning (a7) | Reasoning_a1number (reas_a1) |
| | | Reasoning_a2algebra (reas_a2) |
| | | Reasoning_a3geometry (reas_a3) |
| | | Reasoning_a4data and chance (reas_a4) |
| Comprehensive cognitive process (Item type—IT attributes) | Multiple steps and/or responses (IT1) | |
| | Complexity (IT2) | |
| | Constructed-response (IT3) | |

### Attributes developed according to item type.

With the above attributes, some item difficulties were found to be significantly different although these items required the same or similar content attributes and cognitive process attributes. Then, the researcher identified three types of attributes that were related to item type (IT), named *the comprehensive cognitive process attributes*, or *the IT attributes* (Table 4). These attributes represented more *comprehensive cognitive processing*, which could not be specified according to the TIMSS assessment framework or the following cognitive behaviors defined by the researcher. Specifically, these three attributes were:

*Multiple steps and/or responses* (IT1): Items involve three or more computing steps, or three or more responses.

*Complexity* (IT2): Items have relatively complex wording (such as including more number of words, “more/less than”, “higher/lower”, and unit conversion), or require higher logical reasoning; deal with relatively complex data (such as including more digits and relationship of data).

*Constructed-response* (IT3): Items are constructed-response items or not.

### The second classification of cognitive process attributes—the new cognitive process attributes.

When specifying the attributes of each item, the researcher found the limitations of using the cognitive process attributes based on the TIMSS assessment domains. Through further analyzing the procedures required to solve the 49 mathematics items, the researcher proposed another classification of the cognitive process attributes. The new cognitive process attributes consisted of 4 attributes at level one and 11 attributes at level two (Table 5).

*Identifying* (c1): Recognize and compare simple numbers, data, and figures. It includes two cognitive behaviors: compare the size of numbers and/or order numbers (d1-comparing numbers); recognize a number, data in plane axis, tables or graphs, and shape of a graph/figure (d2-recognizing).

*Computing* (or *Computational application*: c2): Apply basic computational knowledge in arithmetic and algebra. This includes three cognitive behaviors: model the problems with arithmetic expressions or algebraic equations according to question descriptions (d3-formulating); solve computational questions with numbers (d4-computing_number); apply algebraic knowledge to solve computational questions (d5-computing_algebra).

Table 5

*The New Cognitive Process Attributes*

| Level 1 Attributes | Level 2 Attributes |
|---|---|
| Identifying (c1) | Comparing numbers (d1)<br>Recognizing (d2) |
| Computing (c2) | Formulating (d3)<br>Computing_number (d4)<br>Computing_algebra (d5) |
| Judging (c3) | Judging_number (d6)<br>Judging_operationRule (d7)<br>Judging_geometry (d8) |
| Reasoning (c4) | Reasoning_number (d9)<br>Reasoning_algebra (d10)<br>Reasoning_geometry (d11) |

*Judging* (or *Judgmental application*: c3): Need to make judgments in applying knowledge in arithmetic, algebra and geometry. It comprises three cognitive behaviors: judge the relationship of numbers, such as ratio, “more than”, “times”, “constant speed”, data in diagrams (d6-judging_number); judge operation rules in solving arithmetic and algebraic equations (d7-judging_operationRule); judge conception, properties, and rules in geometry questions (d8-judging_geometry).

*Reasoning* (c4): Generalize the solving processes or results with logical or systematic thinking. This attribute is the same as the attribute “reasoning” based on the TIMSS assessment framework, partially because only a few items involve reasoning. At the second level, according to the reasoning knowledge in number, algebra, or geometry, there are three types of reasoning (d9-reasoning_number, d10-reasoning_algebra, and d11-reasoning_geometry). There is no the attribute “reasoning in data and chance”, which is different from the level 2 reasoning based on the TIMSS assessment framework.

### Procedures for Developing the Attributes and the Q-matrices

The procedures of developing the attributes and the Q-matrices were exhibited in Figure 1. Based on a comprehensive literature review, first, the researcher decided to develop an attribute pool according to the TIMSS assessment domains. The four content domains and three cognitive domains were used to construct the level 1attributes. The list of knowledge topics and cognitive behaviors were used to build the level 2 attributes. However, after reviewing each item and the required attributes in Booklets 2 and 3, the researcher found that some attributes at level 2 were not related to any item, especially the listed cognitive behaviors. Then, the attribute pool was reduced by combining some attributes (knowledge topics) and reclassifying the level 1 cognitive process attributes with respect to the four content domains.

Using the proposed attributes, the researcher developed a pilot Q-matrix for the 49 items and conducted trial analyses based on the linear logistic test model (Fischer, 1973). This was a time consuming analysis, involving iterative processes. With multiple regression, the relationship between the above attributes (content and cognitive process) and the item difficulties were investigated for both Booklets 2 and 3. According to the LLTM, item difficulty was a linear combination of content and sub-skills; items with similar cognitive content and processes should have homogeneous statistical properties such as consistent difficulty (Hartz, 2002). The pilot analysis found that a small proportion of the variance in item difficulty was explained by the proposed attributes. Also, some items requiring the same content and cognitive process attributes had significantly different item difficulties, indicating that other attributes must account for the variances in item difficulty. Through closely comparing those items, the researcher identified three types of attributes relative to item type: multiple steps and/or responses (IT1), complexity (IT2), and constructed-response (IT3). In the opinion of the researcher, these attributes are involved in integrated cognitive behaviors, representing more comprehensive cognitive processing.

Through the above analyses, the final attribute pool consisted of 7 level 1 attributes (4 content and 3 cognitive process), 20 level 2 attributes (8 content and 12 cognitive process), and 3 comprehensive cognitive process attributes based on item type. With these attributes, the relationship between the mathematics items and the required attributes were identified to build the Q-matrix of the 49 items. To ensure the Q-matrix was reliable, three experts in educational assessment and mathematics teaching were invited to develop the Q-matrix, which was approved by the University of Denver's Institutional Review Board (Appendix C).The first expert was a professor in quantitative research methods, who had taught course in statistics, psychometrics, and educational measurement for about 30 years at graduate level and who had no K-12 teaching experience. The second expert had a Ph.D. degree in quantitative research methods; she had taught mathematics for Grades 6 to 12 in USA for 27 years. The third expert, with a Bachelor degree in elementary education, was a doctoral student in research methods and statistics; he had teaching experience for Grades 1 to 5 (including mathematics) in Turkey for three years. The three experts and the researcher were from three counties, from whom students participated in the TIMSS assessments. So, the Q-matrices specified by the experts and the researcher would show their views of this international mathematics test.

First, the three experts independently specified the Q-matrix according to the descriptions of the proposed attributes. Comparing their Q-matrices to the researcher's matrix, the researcher found that the agreement was very low. Among the 49 items, the numbers of the items with the same attributes specified by the three experts as those by the researcher were: 15, 11, and 2, respectively for the three experts. Then, the researcher discussed the discrepancies with the experts separately. The first expert revised her Q-matrix after reviewing the different attribute entries and the number of agreed items became 21. Later, we discussed each of the 28 items with different attributes; the number of items with agreed-upon attributes reached to 42 items. The only difference was those items with the attributes "complexity" (IT2). She thought that all test items were simple and there were no complex items. However, the other two experts had different opinions about "complexity" based on their teaching experience for children in elementary and middle school.

In the panel discussion with these two experts, the researcher found that (1) they were more likely to agree with each other, while the first expert and the researcher (both with teaching experience in college/university) had similar opinions about the required attributes; also, they likely identified more attributes required by the test items; (2) the students' background in the U.S. class was more diverse and differences in students' mathematical ability were relatively bigger than for Turkish students and for Chinese students, which would result in different cognitive behaviors in solving mathematics items and different teaching approaches; (3) different study and curriculum requirements existed among the participating countries. All of these led to our different opinions on the relationship between the items and measured attributes and the difficulty in developing the Q-matrices. Cumulatively, we took about six hours to discuss and debate each item until agreement was reached. However, we still could not agree on two entries of the Q-matrices, which were decided using a multiple regression analysis. The final Q-matrix of the 49 items based on the TIMSS assessment framework is shown in Appendix D.

The classification of mathematics content is very stable and widely accepted. However, no commonly recognized attributes for cognitive processing exist. Thus, the cognitive process attributes are more difficult to identify. When specifying the attributes for each item, the researcher found limitations in using the cognitive process attributes (knowing, applying, and reasoning) based on the TIMSS assessment framework. First, at the first attribute level, all items required the attribute "knowing". A student must master "knowing" attributes to implement "applying" and "reasoning" attributes. This indicates that the attribute "knowing" cannot be used to discriminate the items and the students' ability. Second, at the second attribute level, the four knowing attributes (know_a1 to know_a4) are identical to the four level 1 content attributes (a1 to a4). As a result, there are actually only two types of cognitive process attributes (applying and reasoning). So, the researcher reclassified the cognitive process attributes (4 level 1 attributes and 11 level 2 attributes) based on the procedures required to solve the 49 mathematics items; also, the researcher referred to the classification of mathematical attributes identified by K. K. Tatsuoka, Corter, and C. Tatsuoka (2004). The Q-matrix with the new cognitive process attributes is reported in Appendix E.

To validate the Q-matrices, a random Q-matrix was generated for each specified Q-matrix using the Matlab package (MathWorks, 2005). To avoid random Q-matrices were totally artificial, the entries of the Q-matrices specified by the experts and researcher were randomly reordered by the type and level of attributes for each booklet. Specifically, first, an identified Q-matrix was divided into several blocks, such as QM-content L1 (except for a1; or QM-knowing, except for know_a1), QM-content L2 (except for b1), QM-applying, QM-reasoning, and QM-item type. With respect to the new cognitive process attributes, the Q-matrix were separated into four blocks—identifying, computing, judging, and reasoning. Then, each block of Q-matrix was randomly reordered with the Matlab program. Because the majority of items required the attributes "number" (a1 or know_a1) and "whole number or integer" (b1), these three columns were randomly ordered separately. The attribute "knowing" (a5) was still required by all items. Appendices F and G show the random Q-matrices according to the attributes based on the TIMSS assessment framework and item type for Booklets 2 and 3, respectively. The random Q-matrices according to the new cognitive process attributes for the two booklets are displayed in Appendices H and I. Thus, for each booklet, there were two specified Q-matrices and two random Q-matrices.

Then, the eight Q-matrices of both booklets were validated through eight Q-matrix models: QM1 ~ QM8. Each Q-matrix model was cross-validated using two methods: multiple regression and the LSDM. Based on the analysis results, the attributes and the specified Q-matrices were further refined. Some attributes were redefined, combined, or deleted. The Q-matrices were re-examined and revised. The Q-matrix models were simplified or remodeled. Finally, the revised Q-matrices were further assessed with eleven Q-matrix models: QM1 ~ QM11.

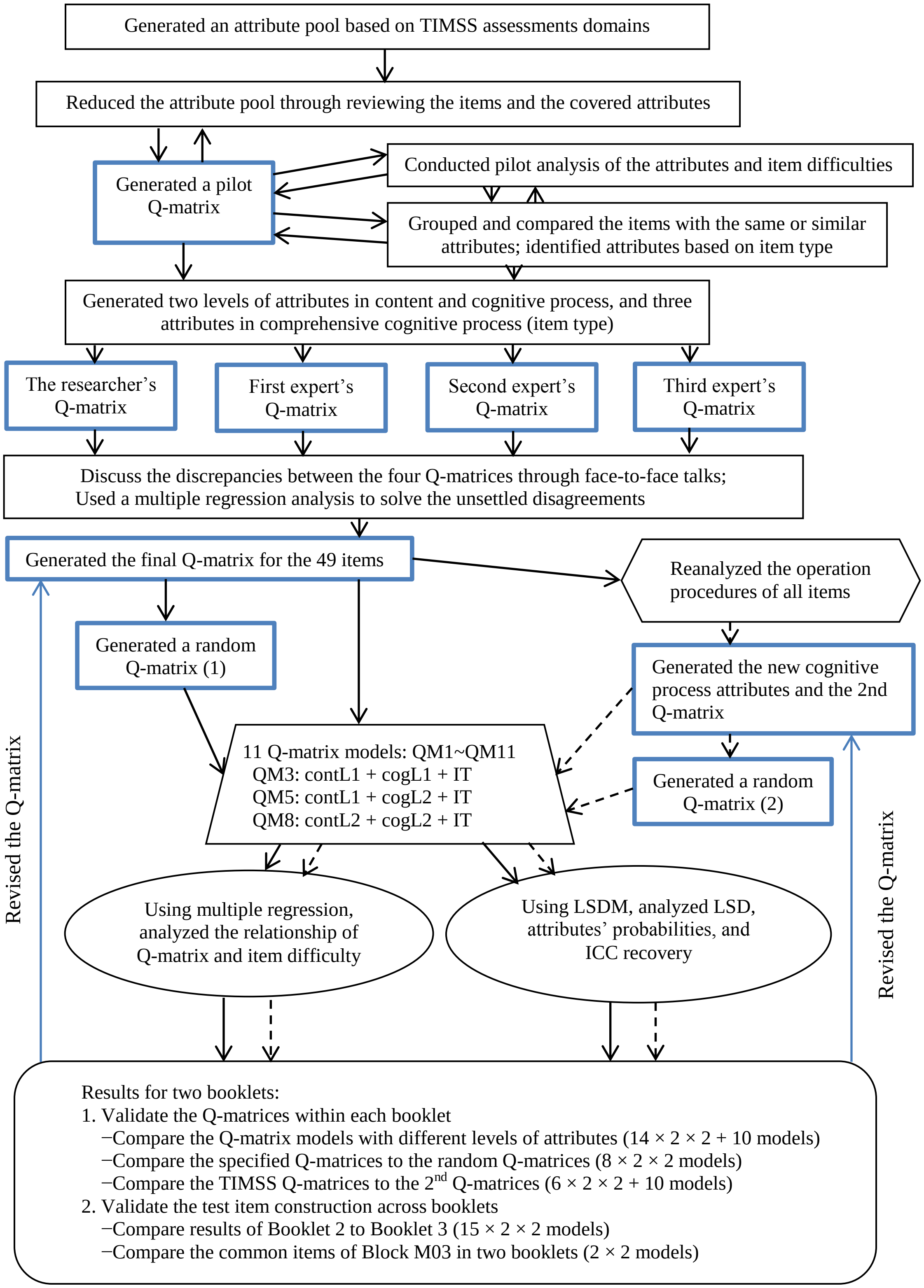
Generated an attribute pool based on TIMSS assessments domains
Reduced the attribute pool through reviewing the items and the covered attributes
Generated a pilot Q-matrix
Conducted pilot analysis of the attributes and item difficulties
Grouped and compared the items with the same or similar attributes; identified attributes based on item type
Generated two levels of attributes in content and cognitive process, and three attributes in comprehensive cognitive process (item type)
The researcher's Q-matrix
First expert's Q-matrix
Second expert's Q-matrix
Third expert's Q-matrix
Discuss the discrepancies between the four Q-matrices through face-to-face talks; Used a multiple regression analysis to solve the unsettled disagreements
Generated the final Q-matrix for the 49 items
Reanalyzed the operation procedures of all items
Generated a random Q-matrix (1)
Generated the new cognitive process attributes and the 2nd Q-matrix
11 Q-matrix models: QM1~QM11
QM3: contL1 + cogL1 + IT
QM5: contL1 + cogL2 + IT
QM8: contL2 + cogL2 + IT
Generated a random Q-matrix (2)
Revised the Q-matrix
Revised the Q-matrix
Using multiple regression, analyzed the relationship of Q-matrix and item difficulty
Using LSDM, analyzed LSD, attributes' probabilities, and ICC recovery
Results for two booklets:
1. Validate the Q-matrices within each booklet
−Compare the Q-matrix models with different levels of attributes (14 × 2 × 2 + 10 models)
−Compare the specified Q-matrices to the random Q-matrices (8 × 2 × 2 models)
−Compare the TIMSS Q-matrices to the 2nd Q-matrices (6 × 2 × 2 + 10 models)
2. Validate the test item construction across booklets
−Compare results of Booklet 2 to Booklet 3 (15 × 2 × 2 models)
−Compare the common items of Block M03 in two booklets (2 × 2 models)

*Figure 1*. Procedures for developing the Q-matrices and validating the Q-matrices.

## Analysis

The validation of the item-attribute matrix and validity of the TIMSS mathematics test were explored through a series of comparison studies using a multiple regression model and the least squares distance method (Dimitrov, 2007). The framework of validating the Q-matrices was displayed in Figure 1. Before the validation analyses, the mathematics test parameters (item difficulty and person ability range) were computed using an appropriate item-calibration model—the Rasch model with the Winsteps program (Linacre, 2013).

### Models of the Q-matrix.

As described above, there existed three types of attributes: *content/knowledge* (two levels, based on TIMSS assessment framework), *cognitive process* (two levels, based on the TIMSS assessment framework or hypothesized cognitive procedures), and *comprehensive cognitive process* (one level, based on item type). The different combinations of the content attributes and the cognitive process attributes were explored to search an adequate group of attributes that can effectively explain the item difficulty and recover the correct item probability. The three types of attributes were added step by step. As a result, there were eight analysis models of the Q-matrix: QM1 ~ QM8 (Table 6).

First, QM1 ~ QM3 investigated the combination of the content and cognitive process attributes at level one. QM1 included only the level 1 content attributes (contL1: a1 ~ a4). Following this, the level 1 cognitive process attributes (cogL1: a6 and a7) were added in QM2; and then, the IT attributes (IT: IT1 ~ IT3) were added in QM3. Second, QM4 ~ QM5 investigated the combination of the level 1 content attributes and the level 2 cognitive process attributes. Based on QM1, some of the level 2 cognitive process attributes (cogL2: appl_a1/a2/a3/a4 and reas_a1/a2/a3/a4) were included in QM4; and the IT attributes were contained in QM5. Third, QM6 ~ QM8 examined the combination of the content and cognitive process attributes at level two. QM6 contain only the level 2 content attributes (contL2: b1 ~ b8); and then, the eight level 2 cognitive process attributes were added in QM7, and the IT attributes were added in QM8. The attributes “a5-knowing” had four sub-attributes (know_a1 ~ know_a4), which were equal to the four content attributes at level one (a1 ~ a4), respectively. Thus, to avoid overlap among the attributes, a5 was not included in QM2 and QM3. Also, because a5 was required by all 49 items, it could not account for any variance in item difficulty, and a5 was excluded from the analysis using multiple regression. Because in the Q-matrix, the elements of “know_a1” to “know_a4” were the same as those of a1 to a4, respectively, the model QM4 (contL1 +cogL2: a1 ~ a4, appl_a1/a2/a3/a4, and reas_a1/a2/a3/a4) was actually equal to “all cogL2” (including know_a1/a2/a3/a4, appl_a1/a2/a3/a4, and reas_a1/a2/a3/a4).

Table 6

*Models of the Q-matrix*

| Model | Attributes |
|---|---|
| QM1 | contL1 |
| QM2 | contL1 + cogL1 |
| QM3 | contL1 + cogL1 + IT |
| QM4 | contL1 + cogL2 |
| QM5 | contL1 + cogL2 + IT |
| QM6 | contL2 |
| QM7 | contL2 + cogL2 |
| QM8 | contL2 + cogL2 + IT |

*Notes*: contL1—Level 1 content attributes (a1 ~ a4); contL2—Level 2 content attributes (b1 ~ b8); cogL1— Level 1 cognitive process attributes (a6 and a7); cogL2— Level 2 cognitive process attributes (appl_a1/a2/a3/a4 and reas_a1/a2/a3/a4); IT—the complex cognitive process attributes (IT1 ~ IT3).

Also, these eight analysis models were compared to the corresponding random Q-matrices. The results from the random Q-matrices provided a baseline for interpretation of analysis results for the specified Q-matrices. Better parameter estimates were expected from the non-randomly generated Q-matrices than those from the random Q-matrices. Specifically, the attributes of the randomly generated Q-matrices were hypothesized to be less useful in explaining differences in item difficulties and a larger average least squares distance (LSD) was expected to be found. The attribute probability curves (APCs) would more likely exhibit non-logical patterns.

There were two types of cognitive process attributes: one based on the TIMSS assessment framework (knowing, applying, and reasoning—called *the TIMSS cognitive process attributes, or the TIMSS process attributes*) and one based on hypothesized cognitive procedures (identifying, computing, judging, and reasoning—called *the new cognitive process attributes, or the new process attributes*). The Q-matrix totally based on the attributes from the TIMSS assessment framework (called *the TIMSS Q-matrix*: TIMSS content attributes + TIMSS process attributes + IT attributes) was compared to the refined Q-matrix with the new process attributes (called *the 2nd Q-matrix*: TIMSS content attributes + new process attributes + IT attributes) using the analysis models shown in Table 6. This analysis examined whether the 2nd Q-matrix provided a better explanation of the mathematics items. In addition, each model of the 2nd Q-matrix was compared to the relative random Q-matrices.

Booklets 2 and 3 contained a common assessment Block M03. The construct validity of the TIMSS mathematics assessment was cross-validated by comparing all results of Booklet 2 to those of Booklet 3. Specifically, the eight analysis models of the

TIMSS Q-matrix of the two booklets were contrasted. Then, the eight analysis models of the 2nd Q-matrix for the two booklets were compared. Moreover, the recovered item probabilities of the 15 common items in Block M03 were exanimated individually. Similar results were expected from the multiple regression analyses and the LSDM analyses. If substantially different results were found between Booklets 2 and 3, it indicated that the items of the two booklets were not equally well explained by the proposed attributes and Q-matrices, or the students might apply different cognitive strategies.

**Analysis steps.**

For each model of the Q-matrix, the relationship between the attributes and item difficulties was tested using a multiple regression model; the item probabilities recovered by the attribute probabilities were examined with the LSDM. The analysis procedures were as follows.

First, the relationship between the attributes and item difficulties was investigated based on the conception of Fischer's LLTM (1973)—item difficulty is a linear combination of the component subtasks with the Q-matrix as the weight of attributes for items. Using a multiple regression model, the variance in item difficulty explained by each Q-matrix was analyzed using SPSS. If a higher variance ($R^2$) in item difficulty can be explained by the identified attributes of a Q-matrix, it indicates that the regression line fits the data better and the specified Q-matrix is more reliable. The researcher expected that most of the variance in item difficulty (e.g. over 70%) can be accounted for by the attributes in the specified Q-matrices, while less variance in item difficulty would be explained by the attributes in the randomly generated Q-matrices. The adjusted $R^2$ was also reported, which penalizes the variance value when extra variables were added. If the adjusted $R^2$ increases, it suggests that the added attributes improve the regression model (Wikipedia, 2013).

Following this, each model of the Q-matrix was contrasted using the LSDM with the MATLAB program (MathWorks, 2005). First, a set of fixed ability levels were selected based on the parameters generated by Winsteps (Linacre, 2013). The probabilities of correct response to each item (*item probability*, $P_{ij}$) were calculated at selected ability levels ($\theta_i$) with the Rasch model.

Second, through minimizing the Euclidean norm of the vector $\|\mathbf{QX} - \mathbf{L}\|$ (see Equations 3 and 4), the probability of correct performance on every attribute (*attribute probability*, $P(\alpha_k = 1| \theta_i)$) was estimated. The attribute probability curves across all ability levels were analyzed. Also, the average least squares distance was computed across items and ability levels. The validity of the Q-matrix was supported if (1) the values of the LSDs decreased monotonically in a relatively small range, and (2) the APCs displayed logical and substantively meaningful patterns in terms of monotonicity, relative difficulty, and discrimination, that is,

> (a) the APCs would increase with the increase of the underlying verbal ability; (b) the relative difficulty of the attributes would make substantive sense; and (c) more difficult attributes would discriminate better among high-ability examinees and, conversely, relatively easy attributes would discriminate better at low ability levels. (Dimitrov, 2007, p. 373)

Third, based on a series of analyses using multiple regression and the LSDM, both the TIMSS Q-matrix and the 2nd Q-matrix were revised. With respect to the problem attributes and results, some attributes were re-defined, classified, or deleted. All elements of the Q-matrices were re-examined. In addition, the final analysis models of the Q-matrix were decided, with some models re-specified.

Fourth, the revised Q-matrices were further tested using multiple regression and the LSDM. The Q-matrix models with reasonable results were used to investigate the ICC recovery with the LSDM. The probability of correct response to each item was recovered by the product of the attribute probabilities estimated by the LSDM. The recovered ICC was compared to the corresponding ICC estimated by the Rasch model. The mean absolute difference (MAD) between the two ICCs was examined across ability levels for each item. A small MAD suggests that the attributes of a Q-matrix explain the test items well. Dimitrov (2007) provided the standards for evaluating the recovery degree: "(a) very good ($0.00 \leq \text{MAD} < 0.02$), (b) good ($0.02 \leq \text{MAD} < 0.05$), (c) somewhat good ($0.05 \leq \text{MAD} < 0.10$), (d) somewhat poor ($0.10 \leq \text{MAD} < 0.15$), (e) poor ($0.15 \leq \text{MAD} < 0.20$), and (f) very poor ($\text{MAD} \geq 0.20$)" (p. 373). These standards were implemented in the present study.

# Chapter Three

# Results

## Data Calibration with the Rasch Model

The data of students' response to each mathematics item were calibrated using the Rasch model for each Booklet. The results generated by Winsteps show that overall, the data of two assessment booklets fit the Rasch model well. For Booklet 2, both the mean square infit and the mean square outfit were 1.01, and the relative standardized fit estimates were 0.0. The person separation reliability was .88. The person logit positions ranged from -5.07 to 5.16. For the 31 test items, the item difficulties ranged from -1.95 to 2.17; the mean square infit estimates were in the .81 to 1.20 range, which suggests that the item fit was acceptable (Linacre, 2013). For Booklet 3, the mean square infit was 1.00 and the mean square outfit was 1.02; both infit and outfit standardized fit estimates were 0.0. The person separation reliability was .89. The logit person ability ranged from -5.23 to 5.19. For the 33 mathematics items, the item difficulties ranged from -2.45 to 2.16; the mean square infit estimates were in the range of .79 to 1.24.

Both Booklets 2 and 3 included 15 items of the assessment block M03. The same items had somewhat different item difficulties in the two booklets (Table 7). The average difficulty difference of the 15 items was .40. Although overall, the difference of item difficulty was not significant based on a t-test, the largest difference was .86 logits.

Table 7

*Item Difficulty of the Mathematics Items*

| No. | Block | Block Seq | Item ID | Item Difficulty |  | Difference in Item Difficulty (id1-id2) |
|---|---|---|---|---|---|---|
|  |  |  |  | Booklet 2 (id1) | Booklet 3 (id2) |  |
| 1 | M02 | 1 | M042003 | -0.69 |  |  |
| 2 | M02 | 2 | M042079 | -1.18 |  |  |
| 3 | M02 | 3 | M042018 | 0.76 |  |  |
| 4 | M02 | 4 | M042055 | 0.13 |  |  |
| 5 | M02 | 5 | M042039 | 0.19 |  |  |
| 6 | M02 | 6 | M042199 | -0.76 |  |  |
| 7 | M02 | 07A | M042301A | -0.11 |  |  |
| 8 | M02 | 07B | M042301B | 1.08 |  |  |
| 9 | M02 | 07C | M042301C | 2.17 |  |  |
| 10 | M02 | 8 | M042263 | 1.85 |  |  |
| 11 | M02 | 9 | M042265 | 0.14 |  |  |
| 12 | M02 | 10 | M042137 | 0.23 |  |  |
| 13 | M02 | 11 | M042148 | -0.98 |  |  |
| 14 | M02 | 12 | M042254 | -1.95 |  |  |
| 15 | M02 | 13 | M042250 | -1.07 |  |  |
| 16 | M02 | 14 | M042220 | 0.99 |  |  |
| 17 | M03 | 1 | M022097 | -0.42 | -0.48 | 0.06 |
| 18 | M03 | 2 | M022101 | -0.69 | -1.07 | 0.38 |
| 19 | M03 | 3 | M022104 | -0.62 | -0.89 | 0.27 |
| 20 | M03 | 4 | M022105 | 0.63 | 0.27 | 0.36 |
| 21 | M03 | 5 | M022106 | 1.11 | 1.11 | 0.00 |
| 22 | M03 | 6 | M022108 | -0.20 | -0.50 | 0.30 |
| 23 | M03 | 7 | M022110 | -0.62 | -0.89 | 0.27 |
| 24 | M03 | 8 | M022181 | -1.18 | -1.98 | 0.80 |
| 25 | M03 | 9 | M032307 | 1.93 | 1.55 | 0.38 |
| 26 | M03 | 10 | M032523 | 1.24 | 0.98 | 0.26 |
| 27 | M03 | 11 | M032701 | -1.59 | -2.45 | 0.86 |
| 28 | M03 | 12 | M032704 | -0.49 | -1.13 | 0.64 |
| 29 | M03 | 13 | M032525 | -0.16 | -0.57 | 0.41 |
| 30 | M03 | 14 | M032579 | -0.51 | -0.88 | 0.37 |
| 31 | M03 | 15 | M032691 | 0.80 | 0.20 | 0.60 |
| 32 | M04 | 1 | M042001 |  | -1.34 |  |
| 33 | M04 | 2 | M042022 |  | -0.10 |  |
| 34 | M04 | 3 | M042082 |  | 0.51 |  |
| 35 | M04 | 4 | M042088 |  | -0.43 |  |
| 36 | M04 | 05A | M042304A |  | -0.92 |  |
| 37 | M04 | 05B | M042304B |  | 2.16 |  |
| 38 | M04 | 05C | M042304C |  | 1.29 |  |
| 39 | M04 | 05D-1 | M042304D-1 |  | 0.28 |  |
| 40 | M04 | 05D-2 | M042304D-2 |  | 0.19 |  |
| 41 | M04 | 6 | M042267 |  | 0.54 |  |
| 42 | M04 | 7 | M042239 |  | 1.22 |  |
| 43 | M04 | 8 | M042238 |  | 0.80 |  |
| 44 | M04 | 9 | M042279 |  | -0.35 |  |
| 45 | M04 | 10 | M042036 |  | 0.69 |  |
| 46 | M04 | 11 | M042130 |  | -0.16 |  |
| 47 | M04 | 12A | M042303A |  | 0.50 |  |
| 48 | M04 | 12B | M042303B |  | 1.72 |  |
| 49 | M04 | 13 | M042222 |  | 0.12 |  |

### The TIMSS Q-Matrix with the Attributes Based on the TIMSS Assessment Framework and Item Type

The final Q-matrix of the 49 items specified by the three experts and the researcher is shown in Appendix D, named *the TIMSS Q-matrix* (TIMSS content attributes + TIMSS process attributes + IT attributes). The total number of items relative to each attribute is reported.

For the level one attributes, the four content attributes were required by from 6 to 28 items of Booklet 2 and by from 6 to 29 items of Booklet 3. A majority of the items (28 items in Booklet 2 and 29 items in Booklet 3) measured the attribute “number.” In Booklet 2, each item required 1 to 3 content attributes, on average 1.58 attributes per item. In Booklet 3, each item required 1 to 3 content attributes, on average 1.61 attributes per item. Most items (20 items in Booklet 2 and 22 items in Booklet 3) measured the cognitive process attribute “a6-applying.” Only four items in each booklet required the attribute “a7-reasoning.” On average, each item required .77 of the attributes a6 and a7.

For the level two attributes and Booklet 2, the eight content attributes (b1 ~ b8) were required by from 2 to 28 items. Except for one item requiring 4 attributes, each item required 1 to 3 content attributes, on average 2.16 attributes per item. The four “applying” attributes (appl_a1 ~ appl_a4) were specified for 3 to 14 items. Three items required the attribute “reasoning_a2algebra”. One item required the attribute “reasoning_a3geometry.” No item needed the attributes “reasoning_a1number” and “reasoning_a4data and chance.” On average, each item required .90 of the six applying and reasoning attributes.

For the level two attributes and Booklet 3, no item involved the level two attributes “b3-pattern” and “reasoning_a2algebra.” The other seven content attributes (b1, b2, b4 ~ b8) were required by from 3 to 29 items. Each item required 1 to 4 content attributes, on average 2.18 attributes per item. Two to 17 items required the four "applying" attributes. The number of items relative to the attributes "reas_a1", "reas_a3", and "reas_a4" were 2, 2, and 1, respectively. On average, each item was related to one of the six applying and reasoning attributes.

There was only one level of *complex cognitive process* attributes based on item type (IT1-multiple steps and/or responses, IT2-complexity, and IT3-constructed-response). For Booklet 2, the three attributes were required by 4, 5, and 11 items, respectively. On average, each item of Booklet 2 was identified with .65 attributes. For Booklet 3, the three attributes were specified to 9, 8, and 12 items, respectively. On average, each item contained .88 attributes.

The Q-matrix was validated with eight models displayed in Table 8. The three types of attributes (content, cognitive process, IT) were added step by step to examine the effect of attributes at different levels.

Table 8

*Models of the Q-matrix Based on the TIMSS Assessment Framework and Item Type*

| Model | Attributes | # of Attributes | |
|---|---|---|---|
| | | Booklet 2 | Booklet 3 |
| QM1 | contL1 | 4 | 4 |
| QM2 | contL1 + cogL1 | 6 | 6 |
| QM3 | contL1 + cogL1 + IT | 9 | 9 |
| QM4 | contL1 + cogL2 | 10 (9) | 11 |
| QM5 | contL1 + cogL2 + IT | 13 (12) | 14 |
| QM6 | contL2 | 8 | 7 |
| QM7 | contL2 + cogL2 | 14 (13) | 14 (13) |
| QM8 | contL2 + cogL2 + IT | 17 (16) | 17 (16) |

*Note*: The numbers in parentheses are the numbers of attributes included in the analyses of the specified Q-matrix models.

**The Specified Q-matrix and item difficulty.**

The relationship between the specified Q-matrix and item difficulty was examined through a series of multiple regression models. Tables 9 to 11 report the results of Booklet 2 regression analyses. Tables 12 to 14 show the results of Booklet 3 regression analyses. To compare the items of the two booklets, the summarized results for both booklets are displayed in Table 15.

***Booklet 2.*** QM1 ~ QM3 were computed for the Q-matrices with the attributes at level one (Table 9). The variance in item difficulty explained by QM1 with the four content attributes (a1 ~ a4) was .20 (adjusted $R^2$ = .08). When two cognitive process attributes (a6 and a7) were added, the variance in item difficulty explained by QM2 was as twice as that in QM1 ($R^2$ = .45, adjusted $R^2$ = .31, $R^2$ change = .25). In QM3, when the IT attributes (IT1 ~ IT3) were included, most variance in item difficulty was explained by the Q-matrix ($R^2$ = .78, adjusted $R^2$ = .68, $R^2$ change = .33).

Table 9

*Results of Multiple Regression of QM1 ~ QM3 for Booklet 2*

| Model | Attributes | # of Attributes | R | $R^2$ | Adjusted $R^2$ | Change Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $R^2$ Change | F Change | df1 | df2 | *p* |
| QM1 | contL1 | 4 | .45 | .20 | .08 | .20 | 1.67 | 4 | 26 | .186 |
| QM2 | contL1+cogL1 | 6 | .67 | .45 | .31 | .25 | 5.38 | 2 | 24 | .012 |
| QM3 | contL1+cogL1+IT | 9 | .88 | .78 | .68 | .33 | 10.36 | 3 | 21 | .000 |

QM4 ~ QM5 investigated the combination of the content attributes at level one and the cognitive process attributes at level two (Table 10). When the six cognitive process attributes (appl_a1/a2/a3/a4 and reas_a2/a3) were added to QM1, the Q-matrix—QM4 accounted for 51% of the variance in item difficulty (adjusted $R^2$ = .30, $R^2$ change

= .31). When the three IT attributes were contained, QM5 explained 79% of the variance in item difficulty (adjusted $R^2$ = .65, $R^2$ change = .28). Because of collinearity, the cognitive process attribute "applying_a2algebra" was excluded from the regression analyses in QM4 and QM5.

Table 10

*Results of Multiple Regression of QM4 and QM5 for Booklet 2*

| Model | Attributes | # of Attributes | R | $R^2$ | Adjusted $R^2$ | Change Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $R^2$ Change | F Change | df1 | df2 | *p* |
| QM4 | contL1+cogL2 | 9 | .71 | .51 | .30 | .31 | 2.62 | 5 | 21 | .054 |
| QM5 | contL1+cogL2+IT | 12 | .89 | .79 | .65 | .28 | 8.14 | 3 | 18 | .001 |

QM6 ~ QM8 were computed for the Q-matrices with the attributes at level two (Table 11). The Q-matrix of the eight content attributes (b1 ~ b8) accounted for 31% variance in item difficulty (adjusted $R^2$ = .06). In QM7, when the six cognitive process attributes were comprised, the Q-matrix took 52% of the variance in item difficulty (adjusted $R^2$ = .16, $R^2$ change = .21). In QM8, when the three IT attributes were encompassed, the variance in item difficulty explained by the Q-matrix (contL2+cogL2+IT) increased to .86 (adjusted $R^2$ = .70, $R^2$ change = .34). Because of collinearity, the attribute "applying_a2algebra" was excluded from QM7 and QM8.

Table 11

*Results of Multiple Regression of QM6 ~ QM8 for Booklet 2*

| Model | Attributes | # of Attributes | R | $R^2$ | Adjusted $R^2$ | Change Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $R^2$ Change | F Change | df1 | df2 | *p* |
| QM6 | contL2 | 8 | .56 | .31 | .06 | .31 | 1.24 | 8 | 22 | .325 |
| QM7 | contL2+cogL2 | 13 | .72 | .52 | .16 | .21 | 1.53 | 5 | 17 | .234 |
| QM8 | contL2+cogL2+IT | 16 | .93 | .86 | .70 | .34 | 11.18 | 3 | 14 | .001 |

***Booklet 3.*** The variances in item difficulty explained by QM1 and QM2 were very small ($R^2$ = .11, adjusted $R^2$ = -.02 by QM1, and $R^2$ = .19, adjusted $R^2$ = .002, $R^2$ change = .08 by QM2: Table 12). The increase of adjusted $R^2$ in QM2 indicated that inclusion of the cognitive process attributes improves the regression model. However, an adjusted $R^2$ of close to 0.0 suggested that the proposed Q-matrices did not predict item difficulty well. In QM3, adding the IT attributes significantly increased the explained variance, $R^2$ = .57, adjusted $R^2$ = .40, $R^2$ change = .38.

Table 12

*Results of Multiple Regression of QM1 ~ QM3 for Booklet 3*

| Model | Attributes | # of Attributes | R | $R^2$ | Adjusted $R^2$ | Change Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $R^2$ Change | F Change | df1 | df2 | *p* |
| QM1 | contL1 | 4 | .33 | .11 | -.02 | .11 | .87 | 4 | 28 | .496 |
| QM2 | contL1+cogL1 | 6 | .44 | .19 | .002 | .08 | 1.27 | 2 | 26 | .298 |
| QM3 | contL1+cogL1+IT | 9 | .76 | .57 | .40 | .38 | 6.84 | 3 | 23 | .002 |

Table 13

*Results of Multiple Regression of QM4 and QM5 for Booklet 3*

| Model | Attributes | # of Attributes | R | $R^2$ | Adjusted $R^2$ | Change Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $R^2$ Change | F Change | df1 | df2 | *p* |
| QM4 | contL1+cogL2 | 11 | .52 | .27 | -.11 | .16 | .65 | 7 | 21 | .711 |
| QM5 | contL1+cogL2+IT | 14 | .83 | .68 | .44 | .41 | 7.85 | 3 | 18 | .001 |

Results of QM4 and QM5 are reported in Table 13. In QM4 when the cognitive process attributes at level two (cogL2) were added to the content attributes at level one, the explained $R^2$ value increased to .27, but the adjusted $R^2$ reduced to -.11, indicating that adding “cogL2” to the Q-matrix did not improve the regression model after penalizing for the number of predictors—the attributes. In QM5, when the IT attributes were included, the *$R^2$* increased to .68, adjusted *$R^2$* = .44, *$R^2$* change = .41.

Similar results as QM4 and QM5 were found for QM6 ~ QM8 (Table 14). The explained variances in item difficulty were .14 (adjusted *$R^2$* = -.10) by QM6, .30 (adjusted *$R^2$* = -.19, *$R^2$* change = .15, *$R^2$* change = .15) by QM7, .71 (adjusted *$R^2$* = .41, *$R^2$* change = .41) by QM8. The decrease of adjusted *$R^2$* in QM7, compared to that of QM8, referred that the added attributes did not enhance the regression model. The attribute "applying_a4data and chance" was excluded from the regression models in QM7 and QM8 because of collinearity—the attributes "applying_a4" and "b7-data organization and representation" were specified to the same items.

Table 14

*Results of Multiple Regression of QM6 ~ QM8 for Booklet 3*

| Model | Attributes | # of Attributes | R | $R^2$ | Adjusted $R^2$ | Change Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $R^2$ Change | F Change | df1 | df2 | *p* |
| QM6 | contL2 | 7 | .38 | .14 | -.10 | .14 | .59 | 7 | 25 | .758 |
| QM7 | contL2+cogL2 | 13 | .54 | .30 | -.19 | .15 | .70 | 6 | 19 | .657 |
| QM8 | contL2+cogL2+IT | 16 | .84 | .71 | .41 | .41 | 7.49 | 3 | 16 | .002 |

***Comparing Booklet 2 to Booklet 3.*** Table 15 provides a summary of the analysis results of the eight Q-matrix models for both Booklets 2 and 3. For both booklets, the content attributes accounted for a small proportion of the variance in item difficulty (20% or less at level one, or 31% or less at level two), while the cognitive process attributes and the IT attributes explained much more variance in item difficulty than the content attributes. Overall, the effect of the IT attributes on the item difficulty was larger than other two types of attributes. Inclusion of the IT attributes resulted in an average increase of 30% or 40% in the $R^2$ values for Booklets 2 and 3, respectively. QM8, which included the level 2 attributes, accounted for most of the variance in item difficulty, .86 for Booklet 2 and .71 for Booklet 3.

Table 15

*Comparison of the Variances in Item Difficulty Explained by the Q-Matrices of Booklets 2 and 3*

| Model | Attributes | Booklet 2 | | | Booklet 3 | | |
|---|---|---|---|---|---|---|---|
| | | # of Attributes | $R^2$ | Adjusted $R^2$ | # of Attributes | $R^2$ | Adjusted $R^2$ |
| QM1 | contL1 | 4 | .20 | .08 | 4 | .11 | -.02 |
| QM2 | contL1+cogL1 | 6 | .45 | .31 | 6 | .19 | .002 |
| QM3 | contL1+cogL1+IT | 9 | .78 | .68 | 9 | .57 | .40 |
| QM4 | contL1+cogL2 | 9 | .51 | .30 | 11 | .27 | **-.11** |
| QM5 | contL1+cogL2+IT | 12 | .79 | .65 | 14 | .68 | .44 |
| QM6 | contL2 | 8 | .31 | .06 | 7 | .14 | -.10 |
| QM7 | contL2+cogL2 | 13 | .52 | .16 | 13 | .30 | **-.19** |
| QM8 | contL2+cogL2+IT | 16 | .86 | .70 | 16 | .71 | .41 |

*Note*: The bolded numbers indicate less variance explained by a model compared to that in its based model.

Compared to Booklet 2, Booklet 3 was found to have less of the variance in item difficulty explained by the proposed Q-matrices. All predicted variance values of Booklet 3 were less than the corresponding explained variances of Booklet 2, especially for the five Q-matrices without the IT attributes. On average, the variance in item difficulty of Booklet 3 explained by the Q-matrices without the IT attributes was only half of the explained variance for Booklet 2; the adjusted $R^2$ values were zero or negative. For Booklet 3, inclusion of the IT attributes greatly improved the regression model; and most of the variance in item difficulty was explained by the proposed Q-matrices, $R^2$ = 57% ~ 71%), which, however, were less than those explained variance of Booklet 2, $R^2$ = 78% ~

86%. The adjusted $R^2$ estimates of Booklet 3 were 40% ~ 44%, while the adjusted $R^2$ estimates of Booklet 2 were 65% ~ 70%. In sum, the results indicated that the proposed Q-matrices did not account for the item difficulties of Booklet 3 as well as the Q-matrices for the item difficulties of Booklet 2.

### The random Q-matrix and item difficulty.

The random Q-matrices for Booklets 2 and 3 (Appendices F and H) were generated using the Matlab package. The elements of the random Q-matrices were balanced by each category of attributes at the two levels through randomly reordering the specified Q-matrices of the two booklets. Thus, the number of attributes in each random booklet was the same as that in each specified booklet. The total numbers of 0's and 1's in the random Q-matrices were equal to the numbers of the relative entries in the specified Q-matrices. The random Q-matrices served as a baseline to validate the Q-matrices.

***Booklet 2*.** Table 16 displays the variances in item difficulty explained by the random Q-matrix and the specified Q-matrix for Booklet 2. The results of the eight Q-matrix models showed that, as expected, all of the variances explained by the random Q-matrices were smaller than those explained by the specified Q-matrices, especially the adjusted $R^2$ estimates. The findings indicated that the specified Q-matrix provided a stronger explanation of item difficulties than the random Q-matrix.

The variances in item difficulty explained by the three random Q-matrices with the level 1 attributes were .14 (adjusted $R^2$ = .01) by QM1, .17 (adjusted $R^2$ = -.04) by QM2, and .36 (adjusted $R^2$ = .09) by QM3. QM4 and QM5 account for .34 (adjusted $R^2$ = .01) and .51 (adjusted $R^2$ = .13) of the variance in item difficulty, respectively. For the

Q-matrices with the level 2 attributes, the explained variances in item difficulty were .25 (adjusted $R^2$ = -.02) by QM6, .50 (adjusted $R^2$ = .07) by QM7, and .63 (adjusted $R^2$ = .16) by QM8.

Table 16

*Variances in Item Difficulty Explained by the Specified and Random Q-Matrices (Booklet 2)*

| Model | Attributes | Booklet 2: Specified QM | | | Booklet 2: Random QM | | |
|---|---|---|---|---|---|---|---|
| | | # of Attributes | $R^2$ | Adjusted $R^2$ | # of Attributes | $R^2$ | Adjusted $R^2$ |
| QM1 | contL1 | 4 | .20 | .08 | 4 | .14 | .01 |
| QM2 | contL1+cogL1 | 6 | .45 | .31 | 6 | .17 | -.04 |
| QM3 | contL1+cogL1+IT | 9 | .78 | .68 | 9 | .36 | .09 |
| QM4 | contL1+cogL2 | 9 | .51 | .30 | 10 | .34 | .01 |
| QM5 | contL1+cogL2+IT | 12 | .79 | .65 | 13 | .51 | .13 |
| QM6 | contL2 | 8 | .31 | .06 | 8 | .25 | -.02 |
| QM7 | contL2+cogL2 | 13 | .52 | .16 | 14 | .50 | .07 |
| QM8 | contL2+cogL2+IT | 16 | .86 | .70 | 17 | .63 | .16 |

***Booklet 3.*** Except for the two random Q-matrix models (QM6 and QM7), the other six random models accounted for less variance in item difficulty than the relatively specified Q-matrices (Table 17). No explained variance by the random Q-matrices was more than 50%, while three identified Q-matrices accounted for 57% ~ 71% of the variance in item difficulty. All adjusted $R^2$ values were negative. The results suggested that overall, the identified relationship of the items by the attributes was more reliable than the randomly generated relationship between the items and the attributes.

With respect to the random Q-matrices, the $R^2$ estimates explained by QM1 ~ QM3 were .04 ~ .16 (adjusted $R^2$ = -.09 ~ -.17). The random QM4 and QM5 accounted for .23 and .27 (adjusted $R^2$ = -.17 and -.30) of the variance in item difficulty. The $R^2$ explained by the random QM6 was .19 (adjusted $R^2$ = -.04), which was higher than the variance explained by the relative specified Q-matrix ($R^2$ = .14, adjusted $R^2$ = -.10). The random QM7 accounted for more variance in item difficulty ($R^2$ = .33) than the corresponding specified Q-matrix ($R^2$ = .30). But the adjusted $R^2$ (= -.194) by the random QM7 was slight less than the adjusted $R^2$ (= -.185) by the specified Q-matrix. When the IT attributes were included in the random QM8, the explained variance ($R^2$ = .48, adjusted $R^2$ = -.10) was less than that by the specified Q-matrix.

Table 17

*Variances in Item Difficulty Explained by the Specified and Random Q-Matrices (Booklet 3)*

| Model | Attributes | Booklet 3: Specified QM | | | Booklet 3: Random QM | | |
|---|---|---|---|---|---|---|---|
| | | # of Attributes | $R^2$ | Adjusted $R^2$ | # of Attributes | $R^2$ | Adjusted $R^2$ |
| QM1 | contL1 | 4 | .11 | -.02 | 4 | .04 | -.09 |
| QM2 | contL1+cogL1 | 6 | .19 | .00 | 6 | .09 | -.12 |
| QM3 | contL1+cogL1+IT | 9 | .57 | .40 | 9 | .16 | -.17 |
| QM4 | contL1+cogL2 | 11 | .27 | -.11 | 11 | .23 | -.17 |
| QM5 | contL1+cogL2+IT | 14 | .68 | .44 | 14 | .27 | -.30 |
| QM6 | contL2 | 7 | .14 | -.10 | 7 | **.19** | **-.04** |
| QM7 | contL2+cogL2 | 13 | .30 | -.19 | 14 | **.33** | -.19 |
| QM8 | contL2+cogL2+IT | 16 | .71 | .41 | 17 | .48 | -.10 |

*Note*: The bolded numbers are the explained variances for the random Q-matrix models higher than those for the specified Q-matrix models.

### LSDM: the specified Q-matrix.

The Q-matrices were analyzed with the LSDM. The least squares distance (LSD) and the attribute probability ($P(\alpha_k=1|\theta_i)$) were examined across ability levels. The Q-matrices were considered to be reliable if (1) the values of the LSD decreased monotonically in a relatively small range, and (2) the attribute probability curves (APCs) displayed logical and substantively meaningful patterns in terms of monotonicity, relative difficulty, and discrimination. If the LSD and most APCs exhibited logical shapes, the mean absolute difference (MAD) between the recovered ICC and the ICC estimated by the Rasch model is reported for each item. The MAD results are reported at the end of all LSDM analyses.

The analyses found that in a Q-matrix, if the elements of two attributes are the same, that is, if two attributes are required by the same items, the attribute probability of one attribute is 100% across all ability levels. When the redundant column is excluded from a Q-matrix, the LSD, MAD, and the APCs of the rest of the attributes do not change, indicating that the attributes with redundant elements could be excluded from the LSDM analysis, as are collinear variables in regression analysis. For Booklet 2, the attribute "applying_a2algebra" had the same elements as "reasoning_a2algebra;" for Booklet 3, the attribute "applying_a4data and chance" had the same elements as "b7-data organization and representation." Thus, the attribute "applying_a2algebra" was excluded from the LSDM analysis of QM4, QM5, QM7, and QM8 for Booklet 2. The attribute "applying_a4data and chance" was excluded from the analysis of QM7 and QM8 for Booklet 3.

***Booklet 2*****.** Figure 2 show that the LSDs of the eight Q-matrix models decreased monotonically in a relatively small range (from .312 to .001), with an increase in ability level. The two models with only the content attributes had relatively higher LSDs (mean LSD = .126 for QM1 (conL1) and .124 for QM6 (conL2): Table 18). When adding the cognitive process attributes in the Q-matrices, the LSD became smaller (mean LSD = .104 for QM2, .101 for QM4, and .105 for QM7). The three models with the IT attributes had the smallest LSDs (mean LSD = .077 for QM5, .078 for QM8, and .81 for QM3).

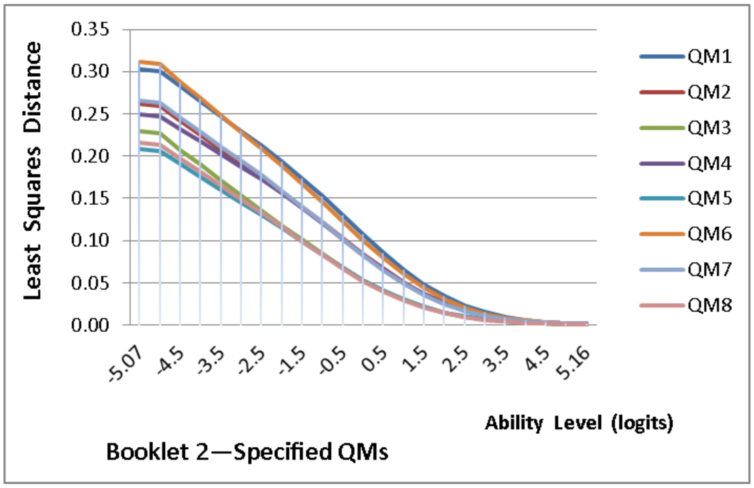

*Figure 2*. Least squares distance for the specified QMs of Booklet 2.

Table 18

*Mean Least Squares Distance for the Specified QMs of Booklet 2*

| | QM1 | QM2 | QM3 | QM4 | QM5 | QM6 | QM7 | QM8 |
|---|---|---|---|---|---|---|---|---|
| Mean LSD | .126 | .104 | .081 | .101 | .077 | .124 | .105 | .078 |
| Order of mean LSD | 8 | 5 | 3 | 4 | 1 | 7 | 6 | 2 |

Figure 3 displays the probability curve of each attribute for QM1 ~ QM8. Most APCs exhibited an acceptable monotonic pattern, relative difficulty, and discrimination. But, no model's APCs were all acceptable. Seven models (QM2 ~ QM8) had at least 2/3 of their attributes whose APCs were acceptable. Some attributes' APCs did not display a reasonable pattern. First, some attributes' probability was 100% across ability levels, indicating these attributes did not discriminate the students with different mathematical abilities. These attributes included those relative to the *data and chance* content, such as "a4-data and chance" (QM1 ~ QM5), "b7-data organization and representation" (QM6 ~ QM8), "b8-data interpretation and chance" (QM6 and QM7), and "applying_a4" (QM4, QM5, and QM7). The probabilities of "a2-algebra" (QM2 ~ QM4) and its sub-attribute "b4-algebraic expressions and equations/formulas functions" (QM7) were, or close to, 100%. Also, the probability of the attribute “IT1-multiple steps/responses” was not shown logically in the three models with the IT attributes, that is, being 100% across ability levels in QM3, QM5, and QM8, suggesting that IT1 might not be a necessary attribute for discriminating mathematical sub-skills. Second, some attributes’ probabilities at the lowest ability levels were higher than probability at a higher ability level, such as “reasoning_a2” in QM5, and “b5-geometric shapes” and “IT2-complexity” in QM8.

The APC results indicated that if the probability of a *content* attribute did not exhibit a logical pattern, the probabilities of the sub-content attributes and the relative cognitive process attributes would be affected, such as the attributes “a2-algebra” and “a4-data and chance.” Second, the models with more attributes (such as over 10 attributes) had more questionable APCs, suggesting that the number of attributes affects the parameter estimates. Moreover, the results suggested that the attribute “IT1-multiple steps/responses” can be excluded. After completing a series of comparison analyses with the new cognitive process attributes and the random Q-matrices, the Q-matrix entries related to *algebra* and *data/chance* knowledge were further examined.

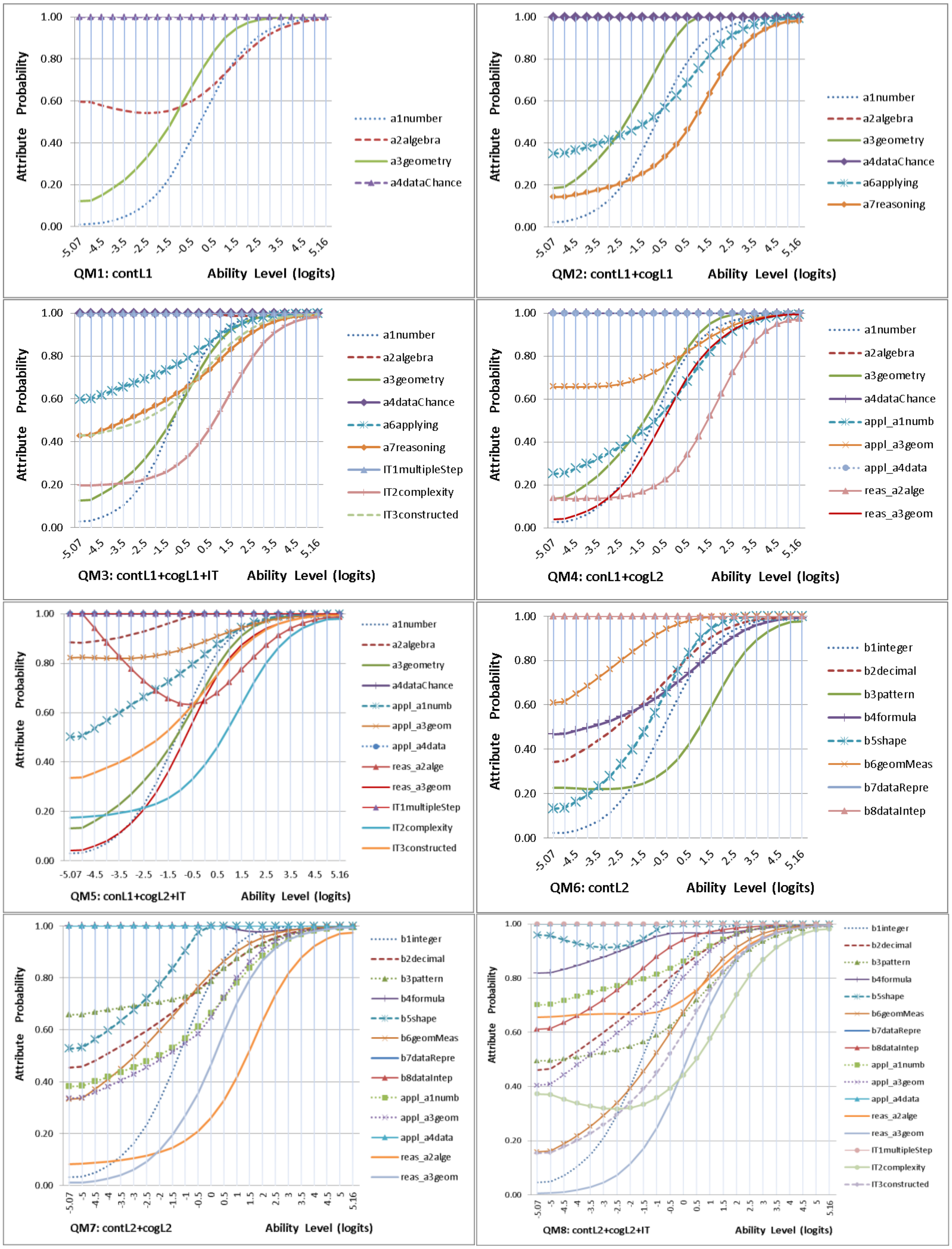

*Figure 3*. Attribute probability for the specified QM1 to QM8 of Booklet 2.

***Booklet 3.*** The LSDs of the eight Q-matrix models decreased monotonically in a small range (from .309 to .001), with an increase in ability level (Figure 4). QM1 and QM6, with the only content attributes, had high LSDs (Table 19). The LSDs were reduced for QM2, QM4, and QM7, which contained both the content and cognitive process attributes. QM3, QM5, and QM8 with the IT attributes had the smallest LSDs.

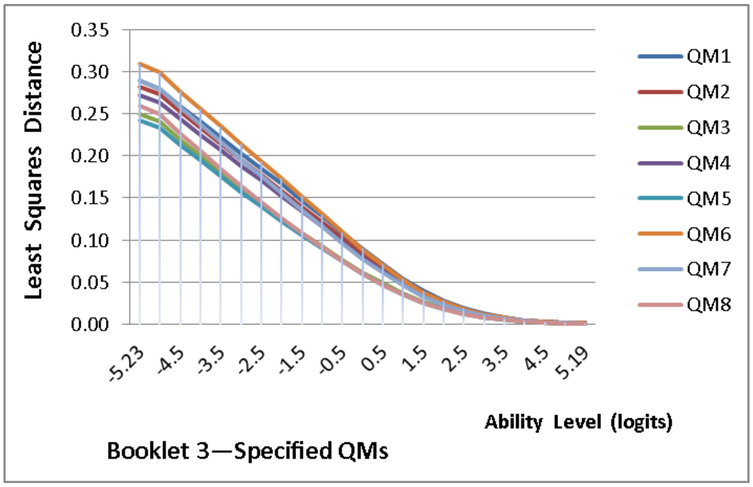

*Figure 4*. Least squares distance for the specified QMs of Booklet 3.

Table 19

*Mean Least Squares Distance for the Specified QMs of Booklet 3*

| | QM1 | QM2 | QM3 | QM4 | QM5 | QM6 | QM7 | QM8 |
|---|---|---|---|---|---|---|---|---|
| Mean LSD | .112 | .107 | .088 | .103 | .086 | .116 | .106 | .090 |
| Order of mean LSD | 7 | 6 | 2 | 4 | 1 | 8 | 5 | 3 |

Five models (QM1 ~ QM3, QM6, and QM7) had 2/3 or more of their attributes whose APCs were acceptable (Figure 5). Among them, all APCs of QM1 QM2, and QM6 displayed a clear pattern with respect to monotonicity, relative difficulty, and discrimination, although the probabilities of “a6-applying” (QM2) and “b6-geometric measurement and location movement” (QM6) at the lowest ability levels (from -5.23 to

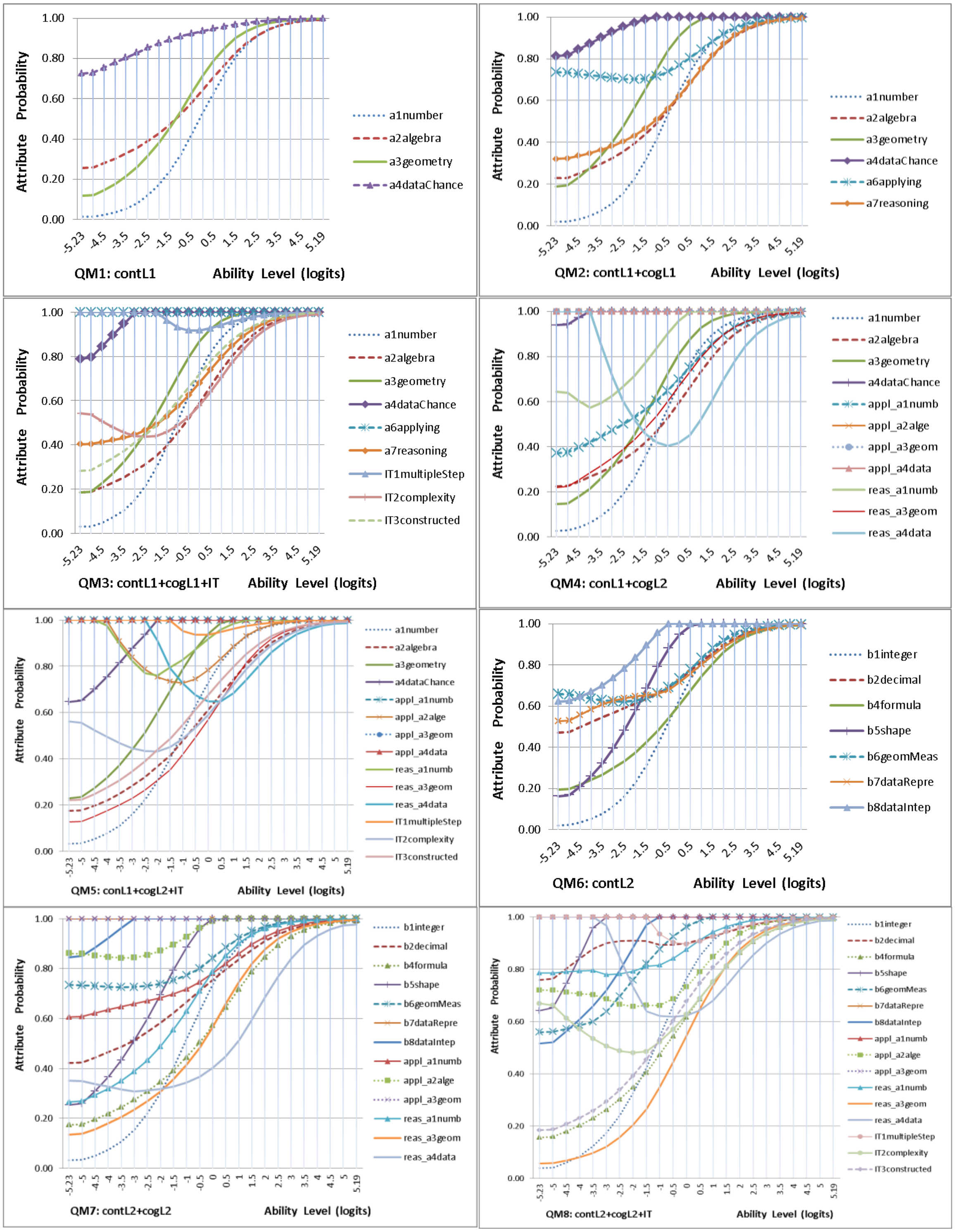

*Figure 5*. Attribute probability for the specified QM1 to QM8 of Booklet 3.

-2.0/-2.5) were slightly decreased with an increase in ability level. The results show that probability curves of a6 and its sub-attributes did not show a logical pattern in many models. The attributes with a probability of 100% across ability levels were a6 (QM3), “b7-data organization and representation” (QM7 and QM8), applying_a1 (QM5 and QM7), applying_a2 (QM4), applying_a3 (QM4, QM5, QM7, and QM8), and applying_a4 (QM4 and QM5). In the three models with the IT attributes (QM3, QM5, and QM8), the probability of “IT1-multiple steps/responses” was, or close to, 100% across ability levels. The probabilities of the following attributes did not always increase with an increase in ability level, such as b6 (QM7), applying_a2 (QM5, QM7, and QM8), reasoning_a1 (QM4 and QM5), reasoning_a4 (QM4, QM5, QM7, and QM8), and “IT2-complexity” (QM3, QM5, and QM8). The Q-matrix elements related to above questionable attributes were further examined.

***Comparing Booklet 2 to Booklet 3.*** With respect to LSD, the eight LSD curves of both booklets’ displayed a similar order. For two models with only the content attributes (QM1 and QM6), the LSDs of Booklet 2 were higher than those of Booklet 3, while for the other six models, the LSDs of Booklet 2 were relatively lower. In terms of APC, more APCs of Booklet 2 exhibited a logical style than APCs of Booklet 3. The results indicated that overall, the Q-matrix of Booklet 2 explained the test items better than the Q-matrix of Booklet 3. However, for Booklet 2, the attributes about the *algebra* and *data/chance* content did not discriminate different mathematical ability levels. In addition, the APCs of both booklets suggested that (1) the more complex Q-matrix, the more APCs that were not acceptable; (2) the attribute “IT1-multiple steps/responses” could be omitted. The entries of the questionable attributes were re-examined.

### LSDM: the random Q-matrix.

Results of both booklets showed that with an increase in ability level, the LSDs of the eight random Q-matrix models decreased monotonically in a small range, from .325 to .002 for Booklet 2 and from .348 to .002 for Booklet 3 (Figures 6 and 7), which were slightly higher than the LSD ranges of the specified Q-matrices. Second, the order of the eight LSD curves for the random Q-matrices was different from that for the specified Q-matrices. Comparatively, the order of the LSD curves for the specified Q-matrices was more reasonable than that for the random Q-matrices. For Booklet 2, the three random models with only the content and cognitive process attributes at level one had relatively higher LSDs (= .131, .127, and .124 for QM1, QM2, and QM3, respectively), while the three random models with the content and cognitive process attributes at level two had relatively lower LSDs (= .109, .103, and .101 for QM6, QM7, and QM8, respectively). For Booklet 3, the random models without the IT attributes (QM1, QM2, QM4, and QM5) had relatively higher LSDs, while the random models with the IT attributes (QM3, QM5, and QM8) had relatively lower LSDs; also, QM7 had small LSDs. Third, except for the random QM6 and QM7 of Booklet 2 and the random QM8 of Booklet 3, the mean LSDs of other random Q-matrices were higher than the corresponding LSDs of the specified Q-matrices (Table 20). On average, the LSD values of the random models were higher than those of the specified models, suggesting that the specified Q-matrices explained the test items better than the random Q-matrices.

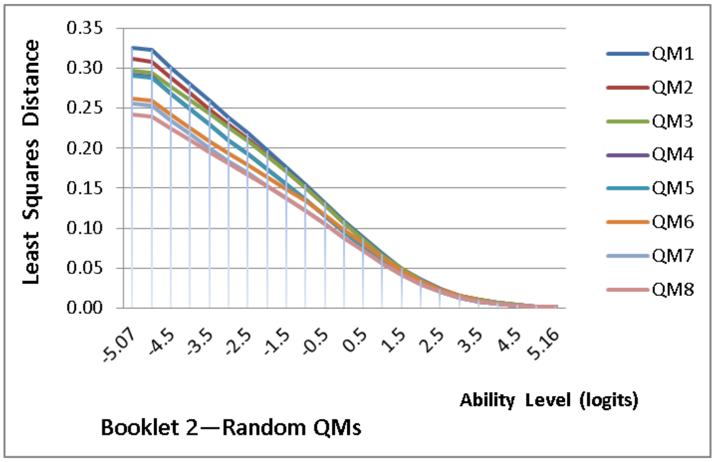

*Figure 6*. Least squares distance for the random QMs of Booklet 2.

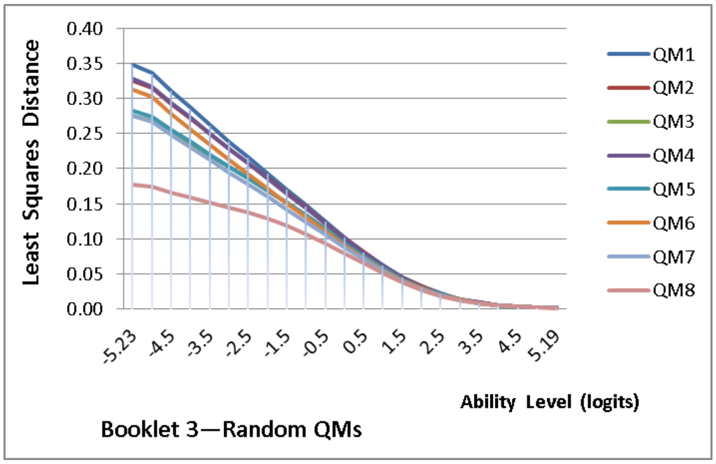

*Figure 7*. Least squares distance for the random QMs of Booklet 3.

Table 20

*Mean Least Squares Distance for the Specified and Random QMs*

| | | QM1 | QM2 | QM3 | QM4 | QM5 | QM6 | QM7 | QM8 |
|---|---|---|---|---|---|---|---|---|---|
| Booklet 2 specified | Mean LSD | .126 | .104 | .081 | .101 | .077 | .124 | .105 | .078 |
| | Order of mean LSD | 8 | 5 | 3 | 4 | 1 | 7 | 6 | 2 |
| Booklet 3 specified | Mean LSD | .112 | .107 | .088 | .103 | .086 | .116 | .106 | .090 |
| | Order of mean LSD | 7 | 6 | 2 | 4 | 1 | 8 | 5 | 3 |
| Booklet 2 random | Mean LSD | .131 | .127 | .124 | .116 | .116 | **.109** | **.103** | .101 |
| | Order of mean LSD | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
| Booklet 3 random | Mean LSD | .131 | .125 | .111 | .125 | .112 | .117 | .107 | **.081** |
| | Order of mean LSD | 8 | 6 | 3 | 7 | 4 | 5 | 2 | 1 |

*Note*: The bolded numbers are the mean LSDs for the random QMs that are lower than those for the specified QMs.

With respect to the APCs, overall, the specified Q-matrix models had more reasonable APCs than the random models (Figures 8 and 9). For both booklets, three random models (QM1 ~ QM3) had 2/3 or over of their total attributes that were acceptable. No model's APCs were all acceptable.  Thus, the APCs revealed that the specified Q-matrices were better than the random Q-matrices.

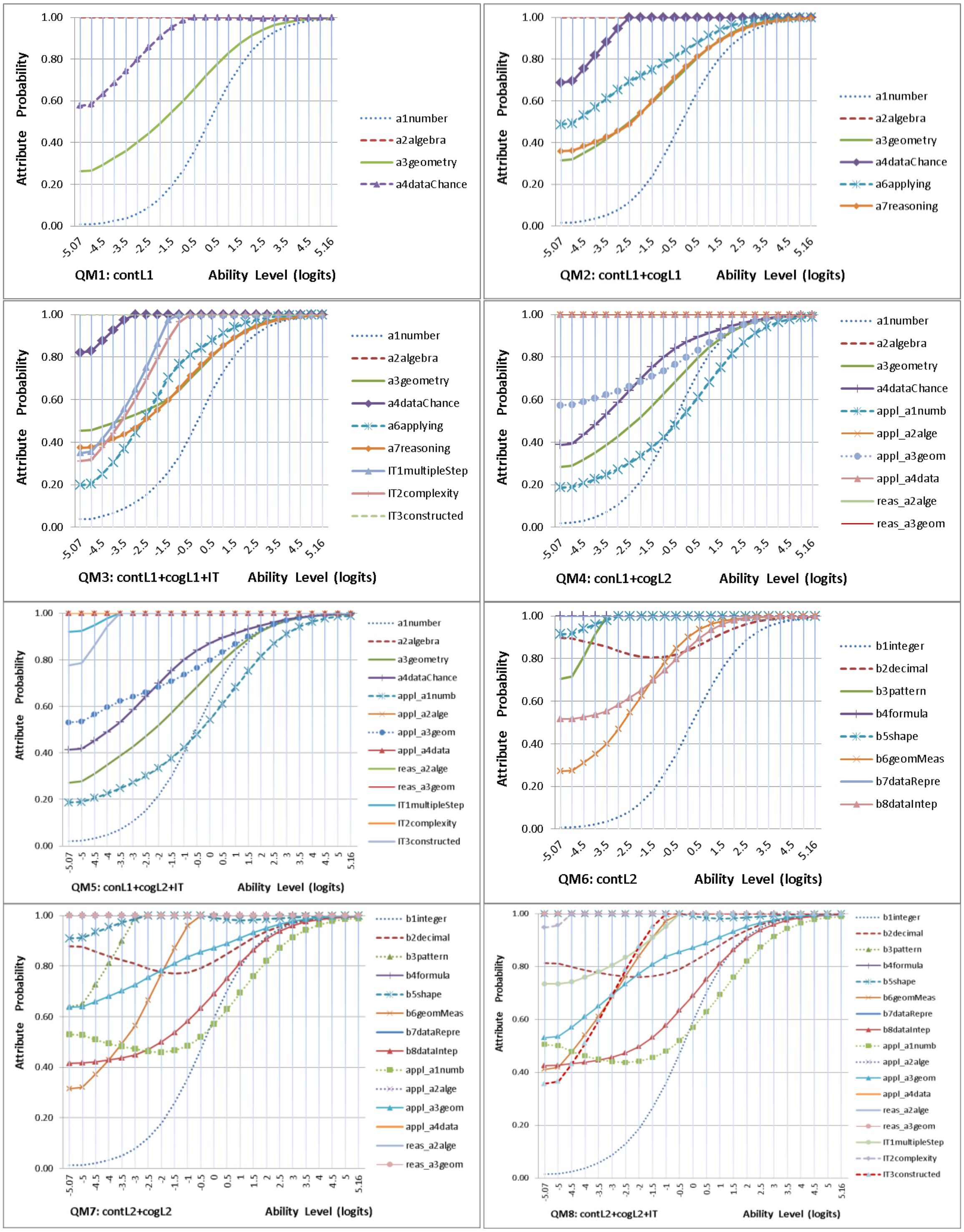

*Figure 8*. Attribute probability for the random QM1 to QM8 of Booklet 2.

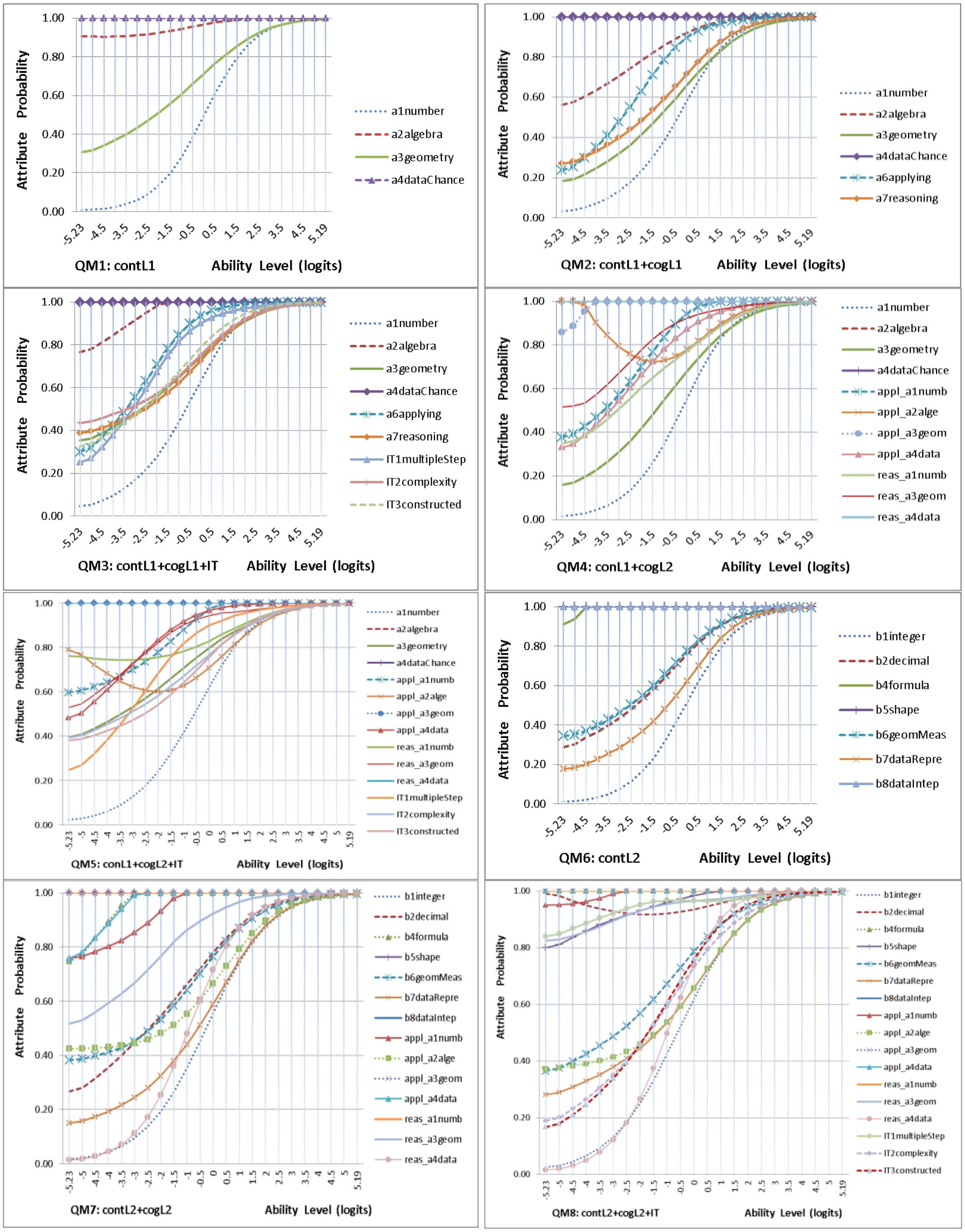

*Figure 9.* Attribute probability for the random QM1 to QM8 of Booklet 3.

## The 2nd Q-Matrix with the Attributes Based on the TIMSS Content, the New Cognitive Processes, and Item Type

The Q-matrix was further evaluated with the new cognitive process attributes. The three cognitive process attributes based on the TIMSS assessment framework (*knowing*, *applying*, and *reasoning*) were replaced by the four cognitive process attributes (*identifying*, *computing*, *judging*, and *reasoning*). There was no change in the two levels of content attributes and the IT attributes. The Q-matrix of the 49 items and the new cognitive process attributes are displayed in Appendix E. The total number of items related to each new attribute is also reported.

For the four new cognitive process attributes at level one (c1 ~ c4), most test item required the attributes "c2-computing" and "c3-judging". In Booklet 2, the numbers of items requiring the four attributes were 11, 24, 18, and 4, respectively. Each item required 1 to 3 attributes, on average 1.84 attributes per item. In Booklet 3, the numbers of items measuring the four attributes were 11, 25, 24, and 4, respectively. Each item required 1 to 3 attributes, on average 1.94 attributes per item.

For the 11 new cognitive process attributes at level two (d1 ~ d11), many mathematics items measured the attributes "d3-formulating," "d4-computing_number," and "d6-judging_number." No item in Booklet 2 required the attribute "d9-reasoning_number;" and no item in Booklet 3 required the attribute "d10-reasoning_algebra." The remaining 10 attributes were required by from 1 to 22 items of Booklet 2 and by from 2 to 25 items of Booklet 3, respectively. On average, each item required 2.35 attributes for Booklet 2 and 2.70 attributes for Booklet 3.

With the new cognitive process attributes, the TIMSS content attributes, and the IT attributes, validation of the 2nd Q-matrix was analyzed using the eight Q-matrix models exhibited in Table 21. Each type of attributes were added step by step. Because there was no change in the Q-matrix models with only the content attributes (QM1 and QM6),  only the models with the new cognitive process attributes were analyzed.

Table 21

*Models of the Q-matrix with the New Cognitive Process Attributes*

| Model | Attributes | # of Attributes | |
|---|---|---|---|
| | | Booklet 2 | Booklet 3 |
| QM2 | contL1 + cogL1 | 8 | 8 |
| QM3 | contL1 + cogL1 + IT | 11 | 11 |
| QM4 | contL1 + cogL2 | 14 | 14 (13) |
| QM5 | contL1 + cogL2 + IT | 17 | 17 (16) |
| QM7 | contL2 + cogL2 | 18 (17) | 17 (16) |
| QM8 | contL2 + cogL2 + IT | 21 (20) | 20 (19) |

*Note*: The numbers in parentheses are the numbers of attributes included in the analyses of the specified Q-matrix models.

### The specified Q-matrix and item difficulty.

Using multiple regression models, the relationship between the item difficulty and the Q-matrices with different types of attributes was examined.

***Booklet 2***. For the attributes at level one, after the four new cognitive process attributes (c1 ~ c4) were added to QM1 (contL1), QM2 accounted for 39% of the variance in item difficulty (adjusted $R^2$ = .16, $R^2$ change = .18: Table 22). When the IT attributes were included, the variance explained by QM3 increased to .81 (adjusted $R^2$ = .70, $R^2$ change = .42).

Table 22

*Results of Multiple Regression of QM2 and QM3 for Booklet 2 (with New Process Attributes)*

| Model | Attributes | # of Attributes | R | $R^2$ | Adjusted $R^2$ | Change Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $R^2$ Change | F Change | df1 | df2 | *p* |
| QM2 | contL1+cogL1 | 8 | .62 | .39 | .16 | .18 | 1.64 | 4 | 22 | .199 |
| QM3 | contL1+cogL1+IT | 11 | .90 | .81 | .70 | .42 | 13.77 | 3 | 19 | .000 |

Based on QM1 (contL1), the 10 new cognitive process attributes at level two (d1 ~ d8, d10, and d11) and the three IT attributes were added step by step (Table 23). When the cognitive process attributes were encompassed, the variance in item difficulty explained by QM4 became .61 (adjusted $R^2$ = .27, $R^2$ change = .41). In QM5, the explained variance was .82 (adjusted $R^2$ = .58, $R^2$ change = .21).

Table 23

*Results of Multiple Regression of QM4 and QM5 for Booklet 2 (with New Process Attributes)*

| Model | Attributes | # of Attributes | R | $R^2$ | Adjusted $R^2$ | Change Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $R^2$ Change | F Change | df1 | df2 | *p* |
| QM4 | contL1+cogL2 | 14 | .78 | .61 | .27 | .41 | 1.67 | 10 | 16 | .175 |
| QM5 | contL1+cogL2+IT | 17 | .91 | .82 | .58 | .21 | 5.03 | 3 | 13 | .016 |

Based on QM6 (contL2), the 10 new cognitive process attributes at level two were added in QM7 and the IT attributes were added in QM8 (Table 24). Addition of the cognitive process attributes greatly increased the explained variance in item difficulty by QM7 ($R^2$ = .71, adjusted $R^2$ = .34, $R^2$ change = .40). The variance in item difficulty explained by QM8 rose to .90 (adjusted $R^2$ = .71, $R^2$ change = .19). Because of collinearity (that is, the attributes “d5-computing_algebra” and “b4-algebraic expressions and equations/formulas functions" were required by the same items), d5 were excluded from the regression models of QM7 and QM8.

Table 24

*Results of Multiple Regression of QM7 and QM8 for Booklet 2 (with New Process Attributes)*

| Model | Attributes | # of Attributes | R | $R^2$ | Adjusted $R^2$ | Change Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $R^2$ Change | F Change | df1 | df2 | *p* |
| QM7 | contL2+cogL2 | 17 | .84 | .71 | .34 | .40 | 2.02 | 9 | 13 | .121 |
| QM8 | contL2+cogL2+IT | 20 | .95 | .90 | .71 | .19 | 6.60 | 3 | 10 | .010 |

***Booklet 3***. Based on QM1 (contL1), inclusion of the new cognitive process attributes (d1 ~ d4) in QM2 and then the IT attributes in QM3 raised the $R^2$ values (Table 25). However, the adjusted $R^2$ value of QM2 decreased, suggesting that the added cognitive process attributes did not improve the regression model when the variance was penalized with respect to the added attributes. The variance in item difficulty explained by QM2 was .18 (adjusted $R^2$ = -.09, $R^2$ change = .07). The variance in item difficulty explained by QM3 was .58 (adjusted $R^2$ = .36, $R^2$ change = .39).

Table 25

*Results of Multiple Regression of QM2 and QM3 for Booklet 3 (with New Process Attributes)*

| Model | Attributes | # of Attributes | R | $R^2$ | Adjusted $R^2$ | Change Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $R^2$ Change | F Change | df1 | df2 | *p* |
| QM2 | contL1+cogL1 | 8 | .43 | .18 | -.09 | .07 | .54 | 4 | 24 | .705 |
| QM3 | contL1+cogL1+IT | 11 | .76 | .58 | .36 | .39 | 6.54 | 3 | 21 | .003 |

When the 10 cognitive process attributes at level two (d1—d9 and d11) were added to QM1 (contL1: Table 26), the $R^2$ value explained by QM4 increased to .38 ($R^2$ change = .26); however, the adjusted $R^2$ value reduced to -.05, indicating that the added attributes did not improve the regression model. In QM5, the explained variance was .69 (adjusted $R^2$ = .38, $R^2$ change = .31). Because of collinearity (the attributes “d5-computing_algebra” and “a2-algebra”), d5 were excluded from the analysis of QM4 and QM5.

Table 26

*Results of Multiple Regression of QM4 and QM5 for Booklet 3 (with New Process Attributes)*

| Model | Attributes | # of Attributes | R | $R^2$ | Adjusted $R^2$ | Change Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $R^2$ Change | F Change | df1 | df2 | *p* |
| QM4 | contL1+cogL2 | 13 | .61 | .38 | -.05 | .26 | .89 | 9 | 19 | .548 |
| QM5 | contL1+cogL2+IT | 16 | .83 | .69 | .38 | .31 | 5.40 | 3 | 16 | .009 |

Based on QM6 that contained the 7 content attributes (b1, b2, b4—b8), the 10 cognitive process attributes at level two were added in QM7 (Table 27). QM7 accounted for more variance in item difficulty than QM6 (contL2) ($R^2$ = .45, $R^2$ change = .27). However, the value of adjusted $R^2$ (= -.10) was almost the same as that of QM6. Inclusion of the IT attributes in QM8 greatly increased the explained variance ($R^2$ = .75, adjusted $R^2$ = .39, $R^2$ change = .30). Because of collinearity (the attributes “d5-computing_algebra” and “b4-algebraic expressions and equations/formulas functions”), d5 were excluded from the regression models.

Table 27

*Results of Multiple Regression of QM7 and QM8 for Booklet 3 (with New Process Attributes)*

| Model | Attributes | # of Attributes | R | $R^2$ | Adjusted $R^2$ | Change Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $R^2$ Change | F Change | df1 | df2 | *p* |
| QM7 | contL2+cogL2 | 16 | .67 | .45 | -.10 | .31 | 1.00 | 9 | 16 | .477 |
| QM8 | contL2+cogL2+IT | 19 | .87 | .75 | .39 | .30 | 5.34 | 3 | 13 | .013 |

***Comparing Booklet 2 to Booklet 3.*** Results of the two booklets showed that when the new cognitive process attributes and then the IT attributes were added in the models, the Q-matrices accounted for more variance in item difficulty (Table 28). The models with the IT attributes (QM3, QM5, and QM8) explained most of the variance in item difficulty. For Booklets 2 and 3, the three Q-matrices with all types of attributes (QM3, QM5, and QM8) accounted for 81% ~ 90% and 58% ~ 75% of the variance in item difficulty, respectively.

Table 28

*Comparison of the Variances in Item Difficulty Explained by the Q-Matrices of Booklets 2 and 3 with the New Cognitive Process Attributes*

| Model | Attributes | Booklet 2 | | | Booklet 3 | | |
|---|---|---|---|---|---|---|---|
| | | # of Attributes | $R^2$ | Adjusted $R^2$ | # of Attributes | $R^2$ | Adjusted $R^2$ |
| QM2 | contL1+cogL1 | 8 | .39 | .16 | 8 | .18 | **-.09** |
| QM3 | contL1+cogL1+IT | 11 | .81 | .70 | 11 | .58 | .36 |
| QM4 | contL1+cogL2 | 14 | .61 | .27 | 13 | .38 | **-.05** |
| QM5 | contL1+cogL2+IT | 17 | .82 | .58 | 16 | .69 | .38 |
| QM7 | contL2+cogL2 | 17 | .71 | .34 | 16 | .45 | -.10 |
| QM8 | contL2+cogL2+IT | 20 | .90 | .71 | 19 | .75 | .39 |

*Note*: The bolded numbers indicate less variance explained by a model compared to that in its based model QM1.

All variance values of Booklet 3 were much less than the corresponding variances of Booklet 2. For Booklet 2, the adjusted $R^2$ estimates increased as additional predictors were included; and all adjusted $R^2$ estimates were positive. For Booklet 3, the adjusted $R^2$ estimates did not always increase as more predictors were included; and four adjusted $R^2$ values were negative. In QM2, QM4 and QM6, when the cognitive process attributes (at level one or level two) were added, the adjusted $R^2$ estimates of Booklet 3 deceased or stayed the same. When all attributes were included, although the identified Q-matrices of both Booklets accounted for most of the variance in item difficulty, the adjusted $R^2$ estimates of Booklet 3 were less than 50% (adjusted $R^2$ = .36 ~ .39), while the adjusted $R^2$ values of Booklet 2 were .58 ~ .71. The results indicated that the identified Q-matrices of Booklet 3 did not account for the test items as effectively as Booklet 2, especially the Q-matrices without the IT attributes.

### The random Q-matrix and item difficulty.

The random Q-matrices of the new cognitive process attributes are shown in Appendices H and I for the two booklets. The elements of the random Q-matrices were produced by randomly reordering the specified Q-matrices. Within each of the four cognitive process attributes (*identifying*, *computing*, *judging*, and *reasoning*), the total numbers of 0's and 1's in the random Q-matrices were equal to the relative numbers in the specified Q-matrices. These random elements were combined with the random elements of the Q-matrix with the content attributes and the IT attributes as displayed in Appendices F and G. Here only the results with the cognitive process attributes were reported, because no change existed in QM1 and QM6 with only the content attributes.

***Booklet 2*.** As shown in Table 29, all variance values ($R^2$ and adjusted $R^2$) explained by the random Q-matrices were less than the corresponding variance estimated explained by the specified Q-matrices, as expected. When the attributes (cogL1, cogL2, and IT) were added in the models, the $R^2$ and adjusted $R^2$ values explained by the specified Q-matrices increased, while the variances explained by the random Q-matrices did not always increase and some adjusted $R^2$ values were negative. For the random Q-matrices, the $R^2$ estimates of QM2, QM4, and QM7 without the IT attributes were .24

~ .60 (adjusted $R^2$ = -.04 ~ -.01). The $R^2$ estimates of QM3, QM5, and QM8 with the IT attributes were .45 ~ .78 (adjusted $R^2$ = .13 ~ .26). The results indicated that the random Q-matrices were not as predictive as the specified Q-matrices in interpreting the items' difficulty.

Table 29

*Variances in Item Difficulty Explained by the Specified and Random Q-Matrices with the New Cognitive Process Attributes (Booklet 2)*

| Model | Attributes | Booklet 2: Specified QM | | | Booklet 2: Random QM | | |
|---|---|---|---|---|---|---|---|
| | | # of Attributes | $R^2$ | Adjusted $R^2$ | # of Attributes | $R^2$ | Adjusted $R^2$ |
| QM2 | contL1+cogL1 | 8 | .39 | .16 | 8 | .24 | -.04 |
| QM3 | contL1+cogL1+IT | 11 | .81 | .70 | 11 | .45 | .13 |
| QM4 | contL1+cogL2 | 14 | .61 | .27 | 14 | .45 | -.03 |
| QM5 | contL1+cogL2+IT | 17 | .82 | .58 | 17 | .59 | .06 |
| QM7 | contL2+cogL2 | 17 | .71 | .34 | 17 | .60 | -.01 |
| QM8 | contL2+cogL2+IT | 20 | .90 | .71 | 20 | .78 | .26 |

***Booklet 3.*** Most of the variances explained by the random Q-matrices were less than those explained by the specified Q-matrices (Table 30). All adjusted $R^2$ values of the random Q-matrices were negative. The explained $R^2$ value (= .48) by the random QM7 was a little higher than that by the specified QM7, while both adjusted $R^2$ values were equal to -.10. The explained variance values by the random QM2 were the same as those by the specified QM2 ($R^2$ = .18, adjusted $R^2$ = -.09). The remaining variance values of the random models were less than those explained by the specified Q-matrices. Thus, the results of Booklet 3 also indicated that the specified Q-matrices were better than the random Q-matrices in accounting for the test item difficulties.

Table 30

*Variances in Item Difficulty Explained by the Specified and Random Q-Matrices with the New Cognitive Process Attributes (Booklet 3)*

| Model | Attributes | Booklet 3: Specified QM | | | Booklet 3: Random QM | | |
|---|---|---|---|---|---|---|---|
| | | # of Attributes | $R^2$ | Adjusted $R^2$ | # of Attributes | $R^2$ | Adjusted $R^2$ |
| QM2 | contL1+cogL1 | 8 | .18 | -.09 | 8 | .18 | -.09 |
| QM3 | contL1+cogL1+IT | 11 | .58 | .36 | 11 | .23 | -.18 |
| QM4 | contL1+cogL2 | 13 | .38 | -.05 | 14 | .35 | -.16 |
| QM5 | contL1+cogL2+IT | 16 | .69 | .38 | 17 | .37 | -.35 |
| QM7 | contL2+cogL2 | 16 | .45 | -.10 | 17 | **.48** | -.10 |
| QM8 | contL2+cogL2+IT | 19 | .75 | .39 | 20 | .52 | -.27 |

*Note*: The bolded numbers are the explained variances for the random Q-matrix models higher than those for the specified Q-matrix models.

### LSDM: the specified Q-matrix.

The six Q-matrix models with the new cognitive process attributes were investigated using the LSDM. The attribute “d5-computing_algebra” was excluded from the analyses of QM7 and QM8 for both booklets and QM4 and QM5 for Booklet 3, because it had the same elements as the attributes “b4-algebraic expressions and equations/formulas functions” or “a2-algebra”.

***Booklet 2.*** The LSDs of the six models decreased monotonically in a small range (from .240 to .001), with an increase in ability level (Figure 10). The three models with the IT attributes (QM3, QM5, and QM8) had relatively smaller LSDs, mean LSDs =.071, .063, and .056. The three models without the IT attributes (QM2, QM4, and QM7) had relatively higher LSDs, mean LSDs =.100, .082, and .075 (Table 31).

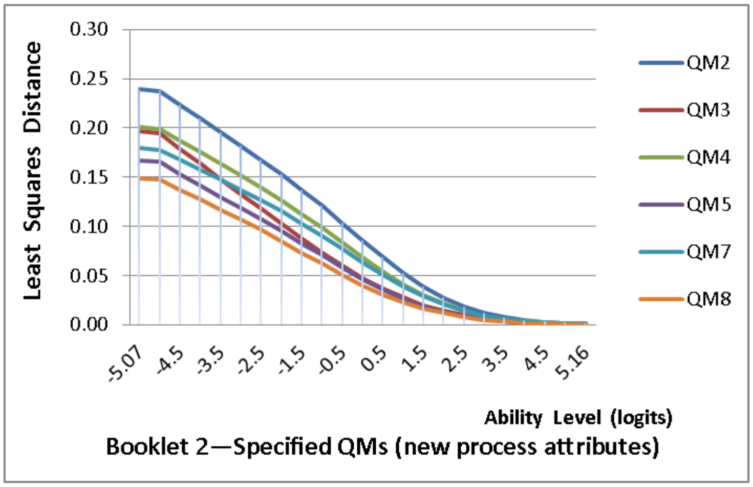

*Figure 10*. Least squares distance for the specified QMs of Booklet 2 (with new process attributes).

Table 31

*Mean Least Squares Distance for the Specified QMs of Booklet 2 (with New Process Attributes)*

| | QM2 | QM3 | QM4 | QM5 | QM7 | QM8 |
|---|---|---|---|---|---|---|
| Mean LSD | .100 | .071 | .082 | .063 | .075 | .056 |
| Order of mean LSD | 6 | 3 | 5 | 2 | 4 | 1 |

With respect to the APCs, most attributes displayed a reasonable pattern (Figure 11). However, no model had an acceptable probability curve for every including attribute. The two models with the level 1 attributes (QM2 and QM3) had at least 2/3 of the total attributes whose APCs were acceptable. In these two models, the attribute probabilities of "a2-algebra" and "a4-data and chance" were 100% across ability levels. In QM3, the probability of "IT1-multiple steps/responses" was 100% across ability levels. The probabilities of a2, a4, and IT1 were also a constant (= 100%) in all involved models (QM4, QM5, and QM8). The problem attributes in QM4, QM5, QM7, and QM8 also included b4 ~ b8, "d4-computing_number," "d5-computing_algebra," "d6-judging_number," and "d10-reasoning_algebra".

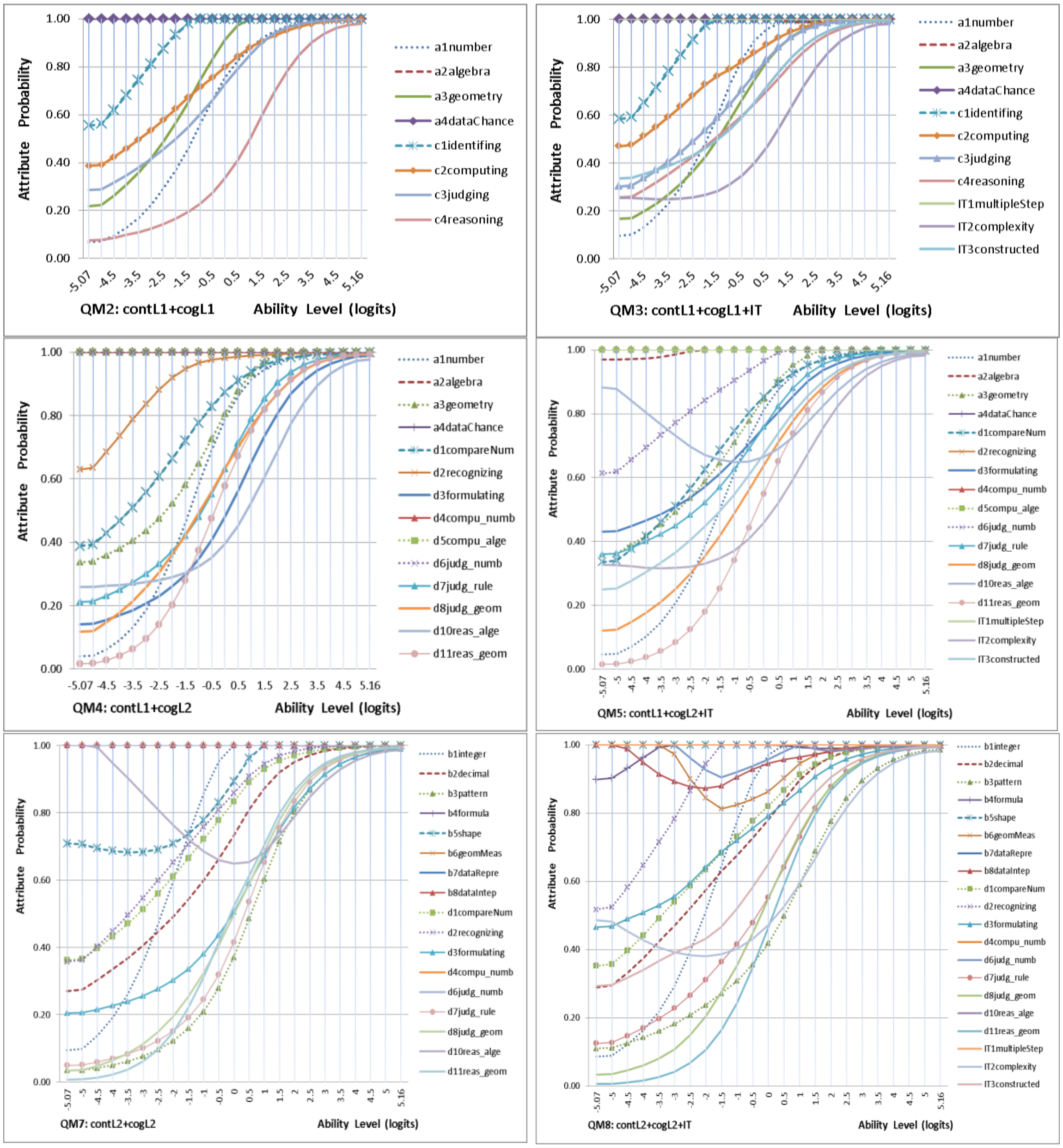

*Figure 11*. Attribute probability for the specified QM2-QM5, QM7 and QM8 of Booklet 2 (with new process attributes).

***Booklet 3.*** The LSDs of the six models decreased monotonically in a small range (from .251 to .001), with an increase in ability level (Figure 12). The three models with the IT attributes (QM3, QM5, and QM8) had relatively smaller LSDs, mean LSDs =.081, .075, and .075. The three models without the IT attributes (QM2, QM4, and QM7) had a little higher LSDs, mean LSDs =.101, .087, and .086 (Table 32).

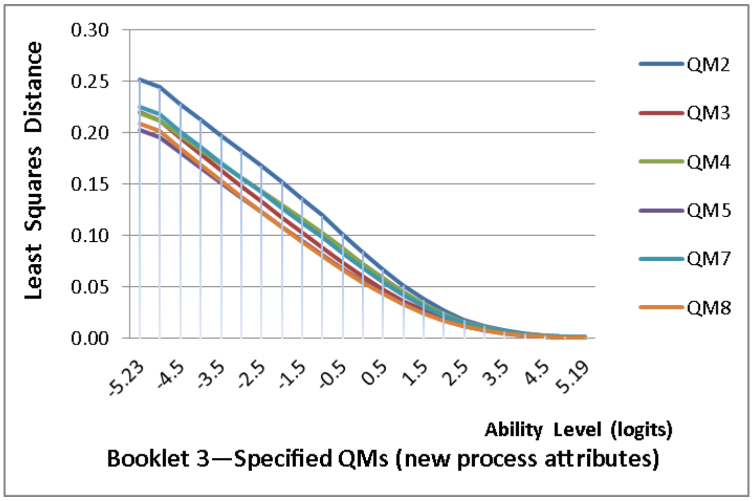

*Figure 12*. Least squares distance for the specified QMs of Booklet 3 (with new process attributes).

Table 32

*Mean Least Squares Distance for the Specified QMs of Booklet 3 (with New Process Attributes)*

| | QM2 | QM3 | QM4 | QM5 | QM7 | QM8 |
|---|---|---|---|---|---|---|
| Mean LSD | .101 | .081 | .087 | .075 | .086 | .075 |
| Order of mean LSD | 6 | 3 | 5 | 1 | 4 | 2 |

With respect to the APCs, two models with the level 1 content and cognitive process attributes (QM2 and QM3) had over 2/3 of their attributes with an acceptable APC (Figure 13). In these two models, the probability of “a4-data and chance” was 100% across ability levels. The probability of “IT1-multiple steps/responses” was, or close to, 100% across ability levels in the models with the IT attributes (QM3, QM5, and QM8).

In these three models, the probabilities of "IT2-complexity" at the lowest ability levels (from -5.23 to -2) did not rise with an increase in ability level. Moreover, the probability of "d6-judging_number" was a constant of 100% in the four relative models (QM4, QM5, QM7, and QM8). In these four models, the problem attributes also included b5 ~ b8, "d4-computing_number," "d8-judging_geometry," and "d9-reasoning_number".

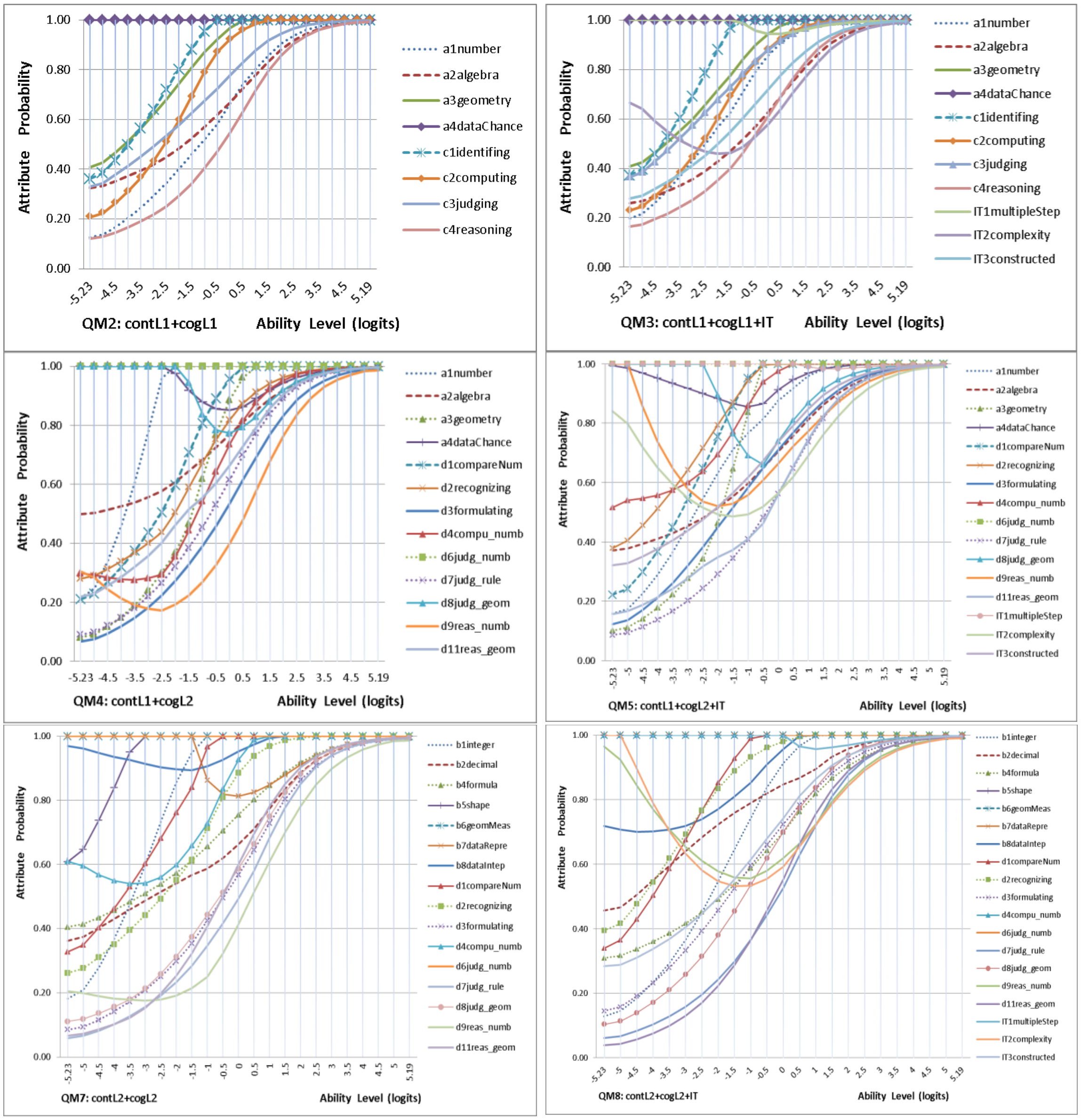

*Figure 13*. Attribute probability for the specified QM2-QM5, QM7 and QM8 of Booklet 3 (with new process attributes).

***Comparing Booklet 2 to Booklet 3.*** For both booklets, the LSDs displayed acceptable curves; the order of the LSD curves was similar to each other, that is, those models with the IT attributes had higher LSDs than those without the IT attributes, indicating consistency in specifying the Q-matrices for the two booklets. For each model, the mean LSD of Booklet 3 was a slightly higher than that of Booklet 2. So, the LSD results suggested that Booklet 2's Q-matrices explained the items better than Booklet 3's Q-matrices.

According to the APCs, both booklets had similar numbers of attributes whose probability exhibited a logical pattern. Meanwhile, only QM2 and QM3 of both booklets had over 2/3 of the attributes with an acceptable APC. Thus, the APC results revealed that the attributes of the two booklets held the test items as well as each other.

In addition, both booklets showed that several attributes' probability curves were not acceptable in many models, such as "a2-algebra," "a4-data and chance," "d4-computing_number," "d6-judging_number," "IT1-multiple steps/responses," and "IT2-complexity," suggesting that the elements of these attributes needed to be further examined, or that some attributes could be revised or even excluded.

### LSDM: the random Q-matrix.

With an increase in ability level, the LSDs of the six random Q-matrix models of both booklets decreased monotonically in a small range, from .289 to .002 for Booklet 2 and from .270 to .002 for Booklet 3 (Figures 14 and 15). The order of the LSD curves of the random models was different from that of the specified models. The order of the LSD curves of the specified models was consistent with the multiple regression analysis. Except for the random QM2 of Booklet 3, the mean LSDs of other eleven random models were higher than those of the specified models (Table 33), suggesting that overall, the specified Q-matrices explained the test items better than the random Q-matrices.

With respect to the APCs, the random Q-matrices were a little better than the specified Q-matrices (Figures 16 and 17). The following random models had at least 2/3 of their attributes with an acceptable APC: four models for Booklet 2 (QM2, QM4, QM5, and QM8) and five models for Booklet 3 (QM2, QM3, QM4, QM7, and QM8).

Thus, the results combining both the LSDs and APCs did not reveal if the specified Q-matrices were absolutely better than the random Q-matrices.

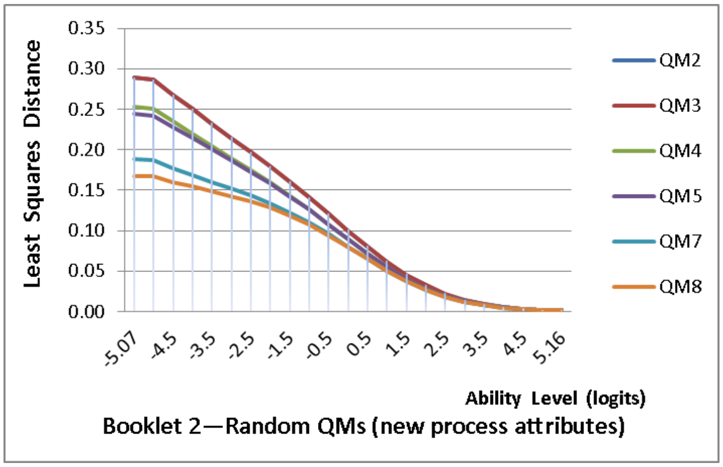

*Figure 14*. Least squares distance for the random QMs of Booklet 2 (with new process attributes).

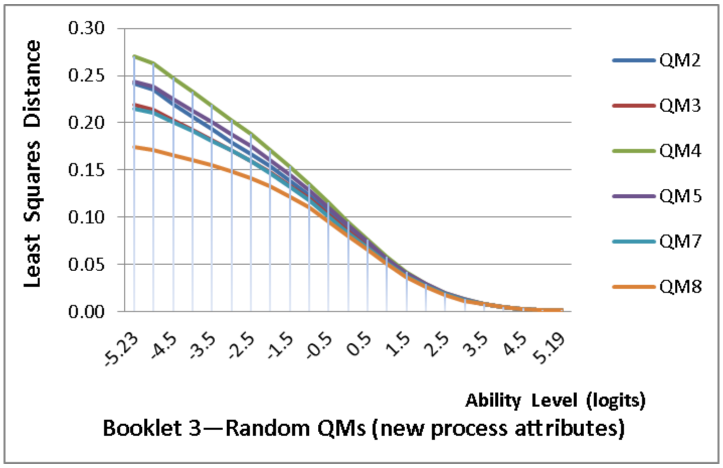

*Figure 15*. Least squares distance for the random QMs of Booklet 3 (with new process attributes).

Table 33

*Mean Least Squares Distance for the Specified and Random QMs (with New Process Attributes)*

| | | QM2 | QM3 | QM4 | QM5 | QM7 | QM8 |
|---|---|---|---|---|---|---|---|
| Booklet 2 specified | Mean LSD | .100 | .071 | .082 | .063 | .075 | .056 |
| | Order of mean LSD | 6 | 3 | 5 | 2 | 4 | 1 |
| Booklet 3 specified | Mean LSD | .101 | .081 | .087 | .075 | .086 | .075 |
| | Order of mean LSD | 6 | 3 | 5 | 1 | 4 | 2 |
| Booklet 2 random | Mean LSD | .119 | .118 | .105 | .103 | .085 | .080 |
| | Order of mean LSD | 6 | 5 | 4 | 3 | 2 | 1 |
| Booklet 3 random | Mean LSD | **.100** | .094 | .111 | .103 | .093 | .082 |
| | Order of mean LSD | 4 | 3 | 6 | 5 | 2 | 1 |

*Note*: The bolded number is the mean LSD for the random QM2 which is lower than that for the specified QM2.

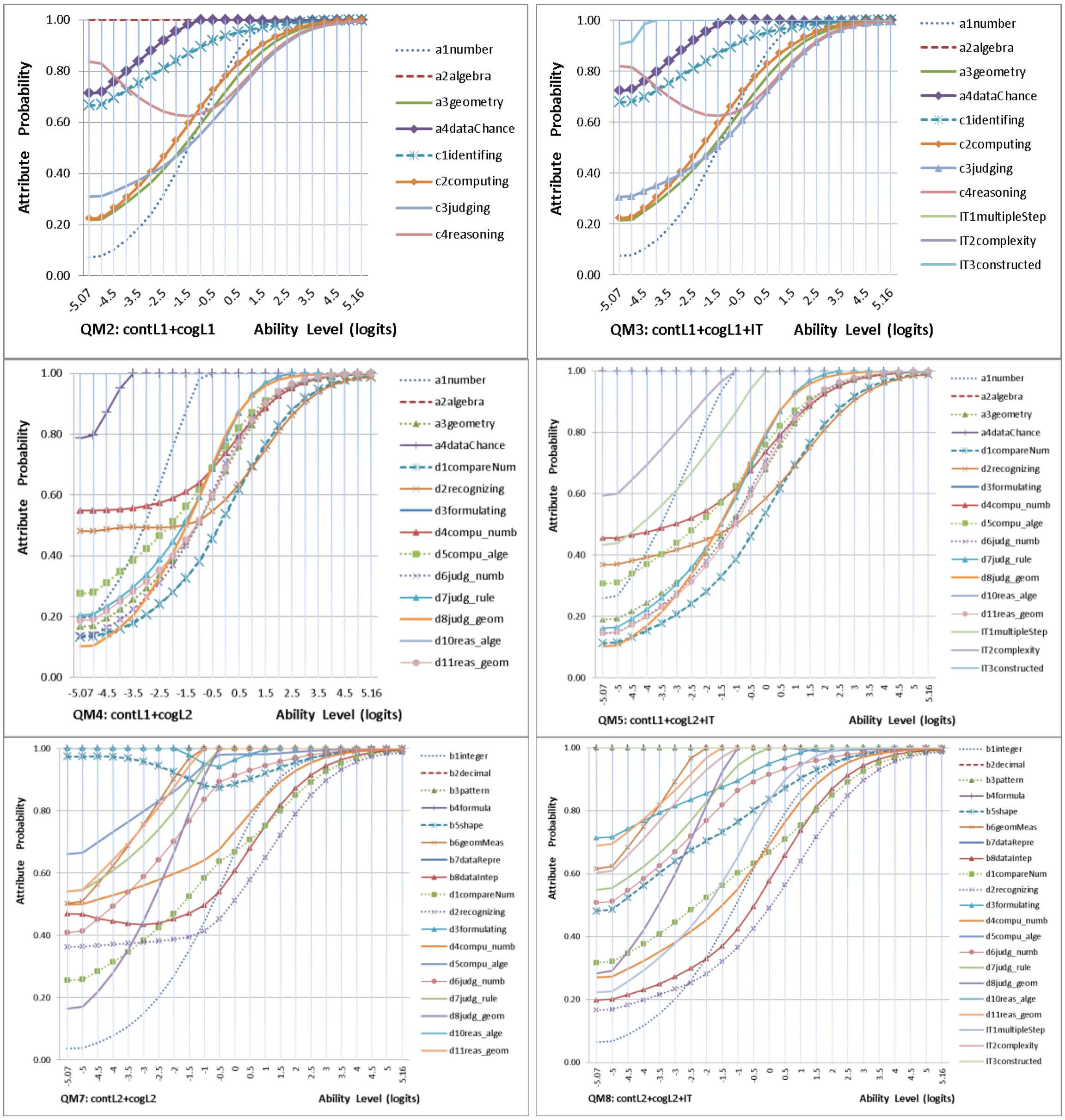

*Figure 16*. Attribute probability for the random QM2-QM5, QM7 and QM8 of Booklet 2 (with new process attributes)

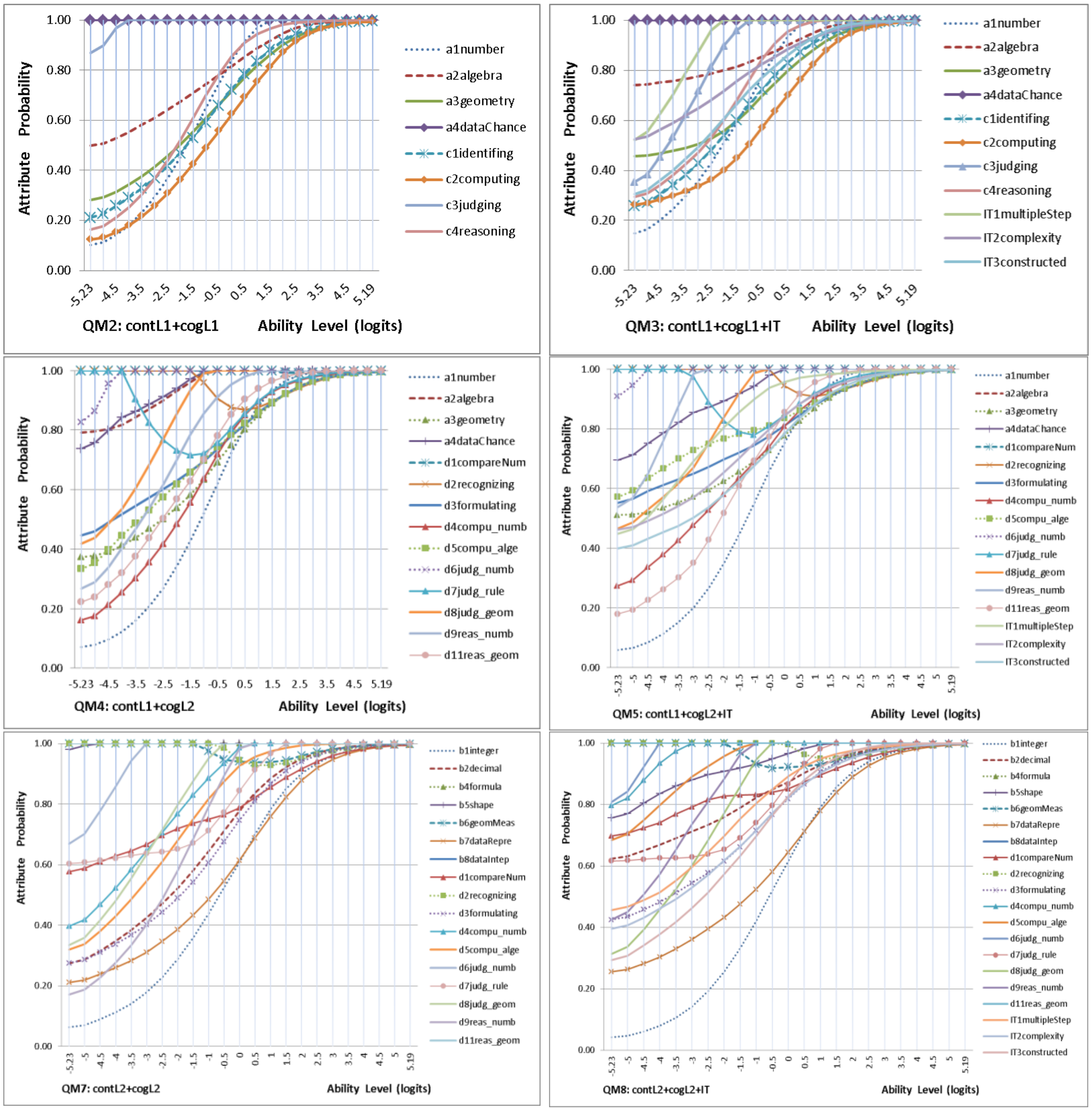

*Figure 17*. Attribute probability for the random QM2-QM5, QM7 and QM8 of Booklet 3 (with new process attributes).

## Results Summary of the Regression Analysis and the LSDM Analysis for the Original Q-Matrices

Each Q-matrix was validated using both multiple regression and the LSDM. A reliable Q-matrix should be useful with both methods. With respect to multiple regression, it was expected that a higher variance in item difficulty was explained by the better-specified Q-matrices. With respect to the LSDM, small LSDs and reasonable APCs were anticipated to be produced for the better Q-matrices. The results of both analyses are summarized in Table 34.

There were some similar results found for the specified Q-matrices with the TIMSS process attributes and the new process attributes. First, the Q-matrices with the IT attributes explained most of the variance in item difficulty (.78 ~ .90 for Booklet 2, .57 ~ .75 for Booklet 3). The content attributes explained a very small proportion of the variance in item difficulty (e.g. $R^2$ = .20 and .11). For Booklet 2, the higher adjusted $R^2$ values for QM2, QM4, and QM7 than those for QM1 and QM6 indicated that the cognitive process attributes explained more variance in item difficulty than the content attributes. However, for Booklet 3, only the adjusted $R^2$ value for QM2 with the TIMSS process attributes was slightly higher than that for QM1. The IT attributes explained much more variance than both the content and process attributes. Second, the LSDs of all models (including the random models) were acceptable. Third, most attributes' APCs were acceptable. However, among 28 specified models, only three models' APCs were all acceptable (QM1, QM6, and QM2 with the TIMSS process attributes for Booklet 3). The probabilities of the attributes a2, a4, d6, and IT1 did not exhibited a meaningful pattern in many models. Fourth, for the three most complex models—QM5, QM7 and

QM8, although the explained variances were relatively higher than those for the other models, many APCs did not show logically. Fifth, both regression and LSDM analyses showed that the Q-matrices explained the item difficulties for Booklet 2 better than for Booklet 3. Sixth, according to the explained variances and the LSDs, the specified Q-matrices were absolutely better than the random Q-matrices, but the APCs did not absolutely distinguish the specified from the random Q-matrices. Overall, the specified Q-matrices were still better than the random Q-matrices.

In sum, no Q-matrix model was decidedly a good Q-matrix that could adequately explained the test items for all analyses. Thus, based on the above analyses, the attributes and Q-matrices were further refined. All elements of the Q-matrices were re-examined.

Table 34

*Results Summary of the Analyses Using Multiple Regression and the LSDM*

| | Q-matrix | Multiple Regression: Variance | LSDM: LSD | | LSDM: Models with at Least 2/3 Acceptable APCs | LSDM: Attributes with Problem APC | |
|---|---|---|---|---|---|---|---|
| | | | Range | Mean LSD | | probability = 100% | Non-monotonically increasing curve |
| Booklet 2 (8 models: with TIMSS process attributes) | Specified (A1) | 8 QMs': > A2's variance | .312 ~ .001 | 6 QMs': < A2's LSD<br>2 QMs': > A2's LSD | 7 QMs (QM2 ~ QM8) | a2 (QM2-QM4), a4 (QM1-QM5), b4 (QM7), b7 (QM6-QM8), b8 (QM6, QM7), appl_a4 (QM4, QM5, QM7), IT1 (QM3, QM5, QM8) | b5 (QM8*), reas_a2 (QM5), IT2 (QM8) |
| | A1's Random | 8 QMs': < A1's variance | .325 ~ .002 | 2 QMs': < A1's LSD<br>6 QMs': > A1's LSD | 3 QMs (QM1, QM2, QM3) | | |
| Booklet 3 (8 models: with TIMSS process attributes) | Specified (A2) | | .309 ~ .001 | | 5 QMs (QM1, QM2, QM3, QM6, QM7) | a6 (QM3), b7 (QM7, QM8), appl_a1 (QM5, QM7), appl_a2 (QM4), appl_a3 (QM4, QM5, QM7, QM8), appl_a4 (QM4, QM5) | a6 (QM2*), b6 (QM7), appl_a2 (QM5, QM7, QM8), reas_a1 (QM4, QM5), reas_a4 (QM4, QM5, QM7, QM8), IT1 and IT2 (QM3, QM5, QM8) |
| | A2's Random | 6 QMs': < A2's variance<br>2 QMs': > A2's variance | .348 ~ .002 | 1 QMs': < A2's LSD<br>7 QMs': > A2's LSD | 3 QMs (QM1, QM2, QM3) | | |
| Booklet 2 (6 models: with new process attributes) | Specified (A3) | 6 QMs': > A4's variance | .240 ~ .001 | 5 QMs': < A4's LSD<br>1 QMs': > A4's LSD | 2 QMs (QM2, QM3) | a2 (QM2 ~ QM4), a4 (QM2 ~ QM5), b4 (QM7), b5 (QM8), b6 (QM7), b7 (QM7, QM8), b8 (QM7), d4 (QM4, QM5, QM7, QM8), d5 (QM4, QM5), d6 (QM4, QM7), IT1 (QM3, QM5, QM8) | b5 (QM7), b6 (QM8), b8 (QM8), d6 (QM8), d10 (QM5, QM7), IT2 (QM8) |
| | A3's Random | 6 QMs': < A3's variance | .289 ~ .002 | 6 QMs': > A3's LSD | 4 QMs (QM2, QM4, QM5, QM8) | | |
| Booklet 3 (6 models: with new process attributes) | Specified (A4) | | .251 ~ .001 | | 2 QM (QM2, QM3) | a4 (QM2, QM3), b5 (QM8), b6 (QM7, QM8), b7 (QM8), d4 (QM8), d6 (QM4, QM5, QM7, QM8), IT1 (QM5) | a4 (QM4, QM5), b7 (QM7), b8 (QM7), d4 (QM7), d8 (QM4, QM5), d9 (QM4, QM5, QM8), IT1 (QM3, QM8), IT2 (QM3, QM5, QM8) |
| | A4's Random | 4 QMs': < A4's variance<br>2 QMs': > A4's variance | .270 ~ .002 | 1 QMs': < A4's LSD<br>5 QMs': > A4's LSD | 5 QMs (QM2, QM3, QM4, QM7, QM8) | | |

* The difference between the probability at the lowest ability level and the lowest probability was less than .05.

## Revision of the Q-Matrices

Considering the performance of every attribute and the relative high ratios of the number of attributes to the number of items, some attributes were redefined or excluded, and the Q-matrix models were simplified or remodeled. The entries of the Q-matrices were re-examined, especially those attributes with problem APCs.

First, four attributes were deleted, which were “applying_a4data and chance,” “reasoning_ a4data and chance,” “d6-judging_number,” and “IT1-multiple steps/responses.” The APCs of “applying_a4” and “reasoning_a4” were not well displayed in many models. The attribute “a4-data and chance” contained data organization and representation, data interpretation, and chance. Actually, a4 primarily involved displaying, reading and calculating the *number* content (a1). Analyzing the items showed that “applying_a4” and “reasoning_a4” could be combined into “applying_a1” and “reasoning_a1”, respectively. Also, only one item of the 49 items measured “reasoning_a4.” After being revised, three TIMSS reasoning attributes (reas_a1/a2/a4) were equal to the three new process attributes (d9-reasoning_number, d10-reasoning_algebra, and d10-reasoning _geometry), respectively. Moreover, the probabilities of d6 and IT1 were, or very close to, a constant of 100% across ability levels in most models, indicating they might be covered by the other cognitive process attributes and IT attributes, respectively. In addition, the cancelation of “d6-judging_number” resulted in re-specifying the elements of the *judging* attribute at level one, “c3-judging.”

Second, the refined validation excluded the models with both content and cognitive process attributes at level two (QM7 and QM8). Many attributes’ APCs were not clearly displayed because of the model complexity, with too many attributes in one model. Second, LSDM analysis assumed independence of attributes. However, the cognitive process attributes at level two were classified in terms of the category of content attributes, which led to overlap between the level 2 content attributes and level 2 process attributes in QM7 and QM8. The model with only the level 2 content attributes (QM6) was examined.

Third, the models of QM4 and QM5 with the new cognitive process attributes were redefined. For QM4 and QM5 with the TIMSS cognitive process attributes, the four sub-attributes of “a5-knowing” (know_a1/a2/a3/a4) actually were equal to the four content attributes at level one (a1-number, a2-algebra, a3-geometry, and a4-data and chance). Thus, the combination of the content attributes at level one (a1, a2, a3, and a4) and the process attributes at level two (appl_a1/a2/a3/a4 and reas_a1/a2/a3/a4) was equal to the combination of all process attributes at level two (contL1 + (partial) cogL2 = all cogL2). For the new process attributes, QM4 and QM5 should include only the process attributes at level two (d1 ~ d11), without the content attributes at level one. The number, algebra, and geometry content had been integrated in the attributes d1 ~ d11, such as “d1-comparing number,” “d2-recognizing,” “d3-formulating,” “d4-computing_number,” and “d5-computing_algebra.” Moreover, when “a4-data and chance” were added, the probability of a4 was, or close to, a constant of 100%, and the probabilities of IT2 (QM5 of Booklet 2) became unreasonable across ability levels; the probabilities of other attributes did not change. Thus, a4 also was excluded from QM4 and QM5. The revised models reduced overlap between the attributes a1 ~ a4 and d1 ~ d11 and so the risk of violating the assumption—independence of attributes.

Fourth, the elements of “b1-whole numbers and integers” were re-specified. An item measuring “b2-fractions, decimals, ratio, proportion, and percent” measured the attribute b1, which led to the construction that the elements of b1 were the same as those of its level 1 attribute “a1-number.” Thus, if an item required both b1 and b2, only b2 was specified; while b1 was specified to the items measuring only whole numbers and integers.

Fifth, two models with the IT attributes (QM3 and QM5) were investigated with two *number* attributes—revised b1 and/or b2. A majority of the items measured “a1-number” (28/31 items in Booklet 2 and 29/33 items in Booklet 3). As a result, the APC of a1 was always close to a perfect S-shape. To effectively display the probability of the *number* attribute and the effect of the attribute involving fractions, decimals, ratio, proportion, and percent, a1 in QM1, QM3 and QM5 was replaced by b1 and b2, or b2 was added to QM5 with the new process attribute at level two.

Sixth, the two sub-attributes of “a4-data and chance” (b7 and b8) were reclassified. B7 measured skills of data organization, representation, and interpretation in figures and graphs; b8 measured skills about computing chance and probability. Thus, in the model with only the level 2 content attributes (QM6), b1, b7, and b8 were revised; no change was made in b2 ~ b6.

Finally, a few elements of the Q-matrices had been revised after examining the relationship between the mathematics items and the attributes. The attributes with a changed elements included “a2-algebra” (or “knowing_a2”), “a6-applying,” “applying_a1,” “d3-formulating,” “d5-computing_algebra,” “d7-judging_operationRule,” and “IT2-complexity.” The refined Q-matrices are displayed in Appendices J and K.

Thus, after revisions, there were still four content attributes at level one (a1 ~ a4), eight content attributes at level two (b1 ~ b8: b1, b7, and b8 were redefined), three TIMSS process attributes at level one, and four new process attributes at level one (Table 35). But there were only 10 TIMSS process attributes at level two and 10 new process attributes at level two; “applying_a4,” “reasoning_a4,” d6, and IT1 were deleted.

Table 35

*Attributes after Revising the Q-Matrices*

| | Level 1 Attributes | Level 2 Attributes |
|---|---|---|
| Content attributes | a1, a2, a3, a4 | b1, b2, b3, b4, b5, b6, b7, b8 |
| TIMSS cognitive process attributes | a5, a6 a7 | know_a1/a2/a3/a4, appl_a1/a2/a3, reas_a1/a2/a3 |
| New cognitive process attributes | c1, c2, c3, c4 | d1, d2, d3, d4, d5, d7, d8, d9, d10, d11<br>(d9 = reas_a1, d10 = reas_a2, d11 = reas_a3) |
| Comprehensive cognitive process attributes | IT2, IT3 | |

## Analysis of the Revised Q-Matrices

### The revised Q-matrix and item difficulty.

Variances in item difficulty explained by the revised Q-matrices were analyzed using multiple regression. First, relationships among the attributes and the item difficulties were examined with the Spearman's rank correlation coefficient (or Spearman's rho). “A5-knowing” was not included because it was specified to all items. The results are reported in Appendix L.

For Booklet 2, several attributes (know_a4, apply_a2, reas_a2, a7-reasoning, IT2-complexity, and IT3-constructed-response) significantly correlated with the item difficulty, $p < .05$ or .01; among them, only the correlation coefficients between two IT attributes and the item difficulty were strong, $r \geq .54$ (Rosenthal, 2001). Strong relationship ($r \geq .50$) also was found between the attributes “a1-number” (or know_a1) and “a3-geometry” (or know_a3), a1 and reas_a3, “a2-applying” and appl_a2/reas_a2, and a3 and appl_a1/appl_a3. Very strong correlation was observed between “a6-applying” and appl_a1, “a7-reasoning” and appl_a2/reas_a2, and “b1-whole numbers and integers” and “b2-fractions, decimals, ratio, proportion, and percent”, $r \geq .76$. The attributes appl_a2 and reas_a2 had the same elements in the Q-matrix, $r = 1.00$; so, apply_a2 was excluded from the regression analysis. For the new process attributes, the following attributes strongly correlated to each other: “c1-identifying” and “a4-data and chance” (or know_a4), “c2-computing” and a1/a3, “c3-judging” and b2, “d2-recognizing” and a4, and “d4-computing_number” and b1/b2. The correlation between “d5-computing_algebra” and a2 was very strong, $r = .91$. Moreover, strong relationship among the IT attributes and the cognitive process attributes was noticed only between IT2 and the new process attributes “d3-formulating”.

For Booklet 3, a less strong relationship was found among the attributes and the item difficulties than for Booklet 2. Only IT2 and IT3 significantly correlated with the item difficulty, $r = .60$ and .39, respectively. The following attributes strongly ($r \geq .50$), or very strongly ($r \geq .70$), correlated to each other: a3 and appl_a3, a6 and appl_a1, a7 and reas_a2/reas_a3, b1 and b2, c1 and a4/c2, c3 and b1/b2/a3, d1 and a4, d2 and a4, and “d8-judging_geometry” and a3. The attributes a2 and “d5-computing_algebra” had the same elements in the Q-matrix, $r = 1.00$.

In sum, it is reasonable that the level 1 attributes strongly related to the relative level 2 attributes. However, for a Q-matrix model, it is not expected that many attributes are highly correlated with respect to the assumption of attribute independence. Results indicated that in general, the IT attributes moderately or weakly related to the TIMSS and new process attributes. Overall, the relationship of the attributes in the Q-matrices with the TIMSS process attributes was stronger than that with the new process attributes. For the models with the level 2 new process attribute, except for IT2 and “d3-formulating”, the correlations among the other attributes were moderate or low.

***Models with only the content attributes*.** For Booklets 2 and 3, the variances in item difficulty explained by the content attributes at level one (QM1) decreased ($R^2$ = .19 and adjusted $R^2$ = .06 for Booklet 2, $R^2$ = .05 and adjusted $R^2$ = -.09 for Booklet 3: Table 36). The variances explained by the revised QM6, with the content attributes at level two, were the same as those for the original QM6 ($R^2$ = .31 and adjusted $R^2$ = .06 for Booklet 2, $R^2$ = .14 and adjusted $R^2$ = -.10 for Booklet 3).

***Models with the TIMSS cognitive process attributes*.** For Booklet 2, the variances explained by the revised models (QM2 ~ QM5-2) were higher than those by the original models (Table 36). Variance explained by the level 1 attributes (QM3 = contL1+cogL1+ IT) was .81 (adjusted $R^2$ = .73); variance explained by the level 2 process attributes (QM5 = contL1+cogL2+ IT) was .81 (adjusted $R^2$ = .72). When the attribute “a1-number” was divided into two attributes “b1-whole numbers and integers” and “b2-fractions, decimals, ratio, proportion, and percent,” no change was observed in $R^2$ values, while the values of adjusted $R^2$ were slightly reduced (adjusted $R^2$ = .72 for QM3-2, adjusted $R^2$ = .70 for QM5-2).

Table 36

*Variances in Item Difficulty Explained by the Revised Q-Matrices with the TIMSS Process Attributes*

| Model | Attributes | Booklet 2 | | | Booklet 3 | | |
|---|---|---|---|---|---|---|---|
| | | # of Attributes | $R^2$ | Adjusted $R^2$ | # of Attributes | $R^2$ | Adjusted $R^2$ |
| QM1 | contL1 | 4 | .19 | .06 | 4 | .05 | -.09 |
| QM2 | contL1+cogL1 | 6 | **.48** | **.36** | 6 | .12 | -.08 |
| QM3 | contL1+cogL1+IT | 8 | **.81** | **.73** | 8 | .57 | **.43** |
| QM4 | contL1+cogL2 | 8 | **.54** | **.37** | 9 | .17 | -.15 |
| QM5 | contL1+cogL2+IT | 10 | **.81** | **.72** | 11 | .64 | **.45** |
| QM6 | contL2 | 8 | .31 | .06 | 7 | .14 | -.10 |
| QM3-2 | contL1+cogL1+IT (b1b2) | 9 | **.81** | **.72** | 9 | .57 | .40 |
| QM5-2 | contL1+cogL2+IT (b1b2) | 11 | **.81** | **.70** | 12 | .64 | .43 |

*Note*: The bolded numbers are the explained variances higher than those for the original Q-matrix models.

For Booklet 3, the adjusted $R^2$ values explained by the revised QM3 and QM5 were a little higher than those by the original QM3 and QM5, respectively. No change existed in the $R^2$ value for QM3. The other variance values explained by the revised models decreased. A lower variance explained for QM4 and QM5 might be partially caused by the reduced number of the level two process attributes. When a1 was replaced by b1 and b2, no change was found in $R^2$ values, but a little lower adjusted $R^2$ was found (= .40 for QM3-2 and .43 for QM5-2).

***Models with the new cognitive process attributes*.** For Booklet 2, variance explained by the revised QM2 was the same as that by the original QM2 ($R^2$ = .39 and adjusted $R^2$ = .17: Table 37). Variance explained by the revised QM3 decreased ($R^2$ = .78 and adjusted $R^2$ = .67), partially due to one less attribute included in model. Although there was a small number of attributes in revised QM4 and QM5 and lower $R^2$ values, the adjusted $R^2$ values were higher than those for the original QM4 and QM5 ($R^2$ = .60 and adjusted $R^2$ = .42 for QM4, $R^2$ = .78 and adjusted $R^2$ = .65 for QM5). The results of QM3-2 and QM5-2 displayed that the use of b1 and b2 almost did not change the $R^2$ values, but reduced the adjusted $R^2$ values slightly.

Table 37

*Variances in Item Difficulty Explained by the Revised Q-Matrices with the New Process Attributes*

| Model | Attributes | Booklet 2 | | | Booklet 3 | | |
|---|---|---|---|---|---|---|---|
| | | # of Attributes | $R^2$ | Adjusted $R^2$ | # of Attributes | $R^2$ | Adjusted $R^2$ |
| QM2 | contL1+cogL1 | 8 | .39 | .17 | 8 | .13 | -.16 |
| QM3 | contL1+cogL1+IT | 10 | .78 | .67 | 10 | **.62** | **.44** |
| QM4 | cogL2 | 9 | .60 | **.42** | 9 | .25 | **-.04** |
| QM5 | cogL2+IT | 11 | .78 | **.65** | 11 | .63 | **.43** |
| QM3-2 | contL1+cogL1+IT (b1b2) | 11 | .78 | .66 | 11 | **.62** | **.42** |
| QM5-2 | cogL2+IT+b2 | 12 | .79 | **.64** | 12 | .63 | **.40** |

*Note*: The bolded numbers are the explained variances higher than those for the original Q-matrix models.

For Booklet 3, the revised QM2 accounted for less variance in item difficulty than the original QM2. The variance explained by the revised QM3 increased ($R^2$ = .62 and adjusted $R^2$ = .44). With fewer attributes, the $R^2$ values explained by QM4 and QM5 decreased ($R^2$ = .25 for QM4 and .63 for QM5), while both adjusted $R^2$ values increased (adjusted $R^2$ = -.04 for QM4 and .43 for QM5), indicating that the new process attributes at level two explained more variance in item difficulty than the original models after penalizing for the number of attributes. When b2 included in QM3-2 and QM5-2, no change was founded in the $R^2$ values, while the adjusted $R^2$ values declined slightly.

***Results summary of multiple regression analysis***. For the revised Q-matrices (with the TIMSS process attributes and the new process attributes), first, both booklets showed the variances explained by only the content attributes were lowest ($R^2$ = .39 and .13 for QM1 and QM6, respectively). When the cognitive process attributes were added (QM2 and QM4), the four adjusted $R^2$ values for Booklet 2 increased; but for Booklet 3, only two adjusted $R^2$ values increased slightly. For the models with the IT attributes (QM3, QM3-2, QM5, and QM5-2), the explained variances were the highest ($R^2$ = .62 ~ .81) and all adjusted $R^2$ values adequately increased. The results suggested that the cognitive process attributes, especially those complex process attributes, played an important role in explaining the test item difficulties. Second, all variances in item difficulty explained by Booklet 2's Q-matrices were higher than those by Booklet 3's Q-matrices. Also, for Booklet 3's models without the IT attributes (QM1, QM2, QM4, and QM6), the adjusted $R^2$ values were negative. The results indicated that the Q-matrices of Booklet 2 interpreted the items better than the Q-matrices of Booklet 3. Third, when the attribute "a1-number" was divided into two sub-attributes b1 and b2 or when b2 was added, no substantial change in the explained variance was found.

#### LSDM: models with the TIMSS content and TIMSS cognitive process attributes.

***Least squares distance of Booklets 2 and 3.*** For both booklets, the LSDs of the eight Q-matrix models decreased monotonically in a small range, from .31 to .001 (Figures 18 and 19). The orders of eight LSD curves of two booklets were similar to each other. The mean LSDs of two models with only the content attributes (QM1 and QM6) were the largest (Table 38). When the cognitive process attributes were included in QM2 and QM4, the LSD values were reduced. The models with the IT attributes (QM3 and QM5) had the lowest LSDs. Moreover, when one number attribute a1 was replaced by b1 and b2, the LSDs of QM3-2 and QM5-2 were further reduced, indicating including b1 and b2 helped to recover the items. Compared to Booklet 3, two models' mean LSDs of Booklet 2 (QM1 and QM6) were slightly higher; the other six models' mean LSDs of Booklet 2 were a little lower, suggesting that overall, the attributes recovered the items of Booklet 2 better than the item of Booklet 3.

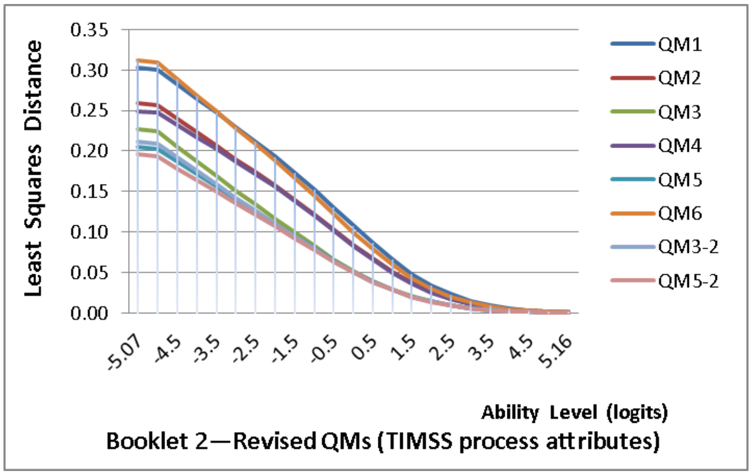

*Figure 18*. Least squares distance for the revised QMs of Booklet 2 (with TIMSS process attributes).

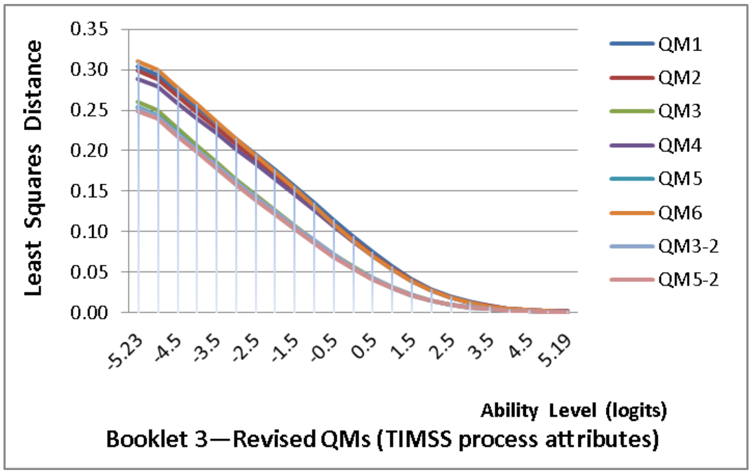

*Figure 19*. Least squares distance for the revised QMs of Booklet 3 (with TIMSS process attributes).

Table 38

*Mean Least Squares Distance for the Revised QMs of Booklet 2 (with TIMSS Process Attributes)*

| | | QM1 | QM2 | QM3 | QM4 | QM5 | QM6 | QM3-2 | QM5-2 |
|---|---|---|---|---|---|---|---|---|---|
| Booklet 2 | Mean LSD | **.126** | .104 | .080 | .101 | .075 | **.124** | .076 | .072 |
| | Order of mean LSD | 8 | 6 | 4 | 5 | 2 | 7 | 3 | 1 |
| Booklet 3 | Mean LSD | .118 | **.114** | **.089** | **.111** | **.086** | .117 | **.087** | **.085** |
| | Order of mean LSD | 8 | 6 | 4 | 5 | 2 | 7 | 3 | 1 |

*Note*: The bolded numbers are the mean LSDs for Booklets 2 or 3 that are higher than the corresponding mean LSDs for another booklet.

***Attribute probability curves of Booklets 2 and 3.*** For Booklet 2, compared to the APCs of the original QM1 ~ QM6, the APCs of three revised models (QM1, QM3, and QM6) were improved (Figure 20). No obvious change was noted in APCs of QM2, QM4, and QM5. The APCs of QM3-2 and QM5-2 were improved slightly when “a1-number” was divided into two sub-attributes b1 and b2. In all eight models, more than 2/3 of the included attributes had an acceptable APC. However, problems still existed in the APCs of attributes a2, a4, b7, b8, and “reasoning_a2”.

First, the probability of “a2-algebra” in QM2 was a constant of 100% across ability levels. In QM1 and QM5, the probabilities of a2 at the lowest ability levels (-5.07 ~ -3.5 in QM1 and -5.07 ~ -3.0 in QM5) were slightly higher than that at the ability level of -3.0 or -2.5, respectively; but the differences in probabilities were less than .05 and thus, acceptable. In QM3, the APC of a2 was a U-shaped curve, close to a probability line of 100%. Second, the APC of “reasoning_a2” displayed a U-shaped curve in QM4, QM5, and QM5-2. Third, the probability of a4 was a constant of 100% across ability levels in all eight models, as well as its sub-attributes in QM6. The results indicated that the attribute of data and chance did not discriminate the students’ mathematical ability.

For Booklet 3, with the revised Q-matrix, the APCs of QM3 were improved slightly, while the APCs of a4 in QM2 and QM4 and “b7-data” in QM6 became worse, with a value of 100% across ability levels (Figure 21). No substantial change existed in the other three models. Among the eight models, five models (QM1 ~ QM3, QM3-2, and QM6) had at least 2/3 of their attributes with an acceptable APC. All attributes’ APCs in QM1 exhibited a logical pattern. With b1 and b2 instead of a2, the APCs of QM3-2 and QM5-2 were not improved. The probability of “a6-applying” was a constant of 100% in QM3 and QM3-2. In QM4, QM5, and QM5-2, the probabilities of the level 2 attributes of a6 (appl_a1/a2/a3) were not displayed reasonably; also, the probabilities of “reasoning_a1” exhibited a U-shaped curve or were close to 100% across ability levels. In addition, the probabilities of “IT2-complexity” did not increase with an increase in ability level from -5.23 to -2.0/-2.5.

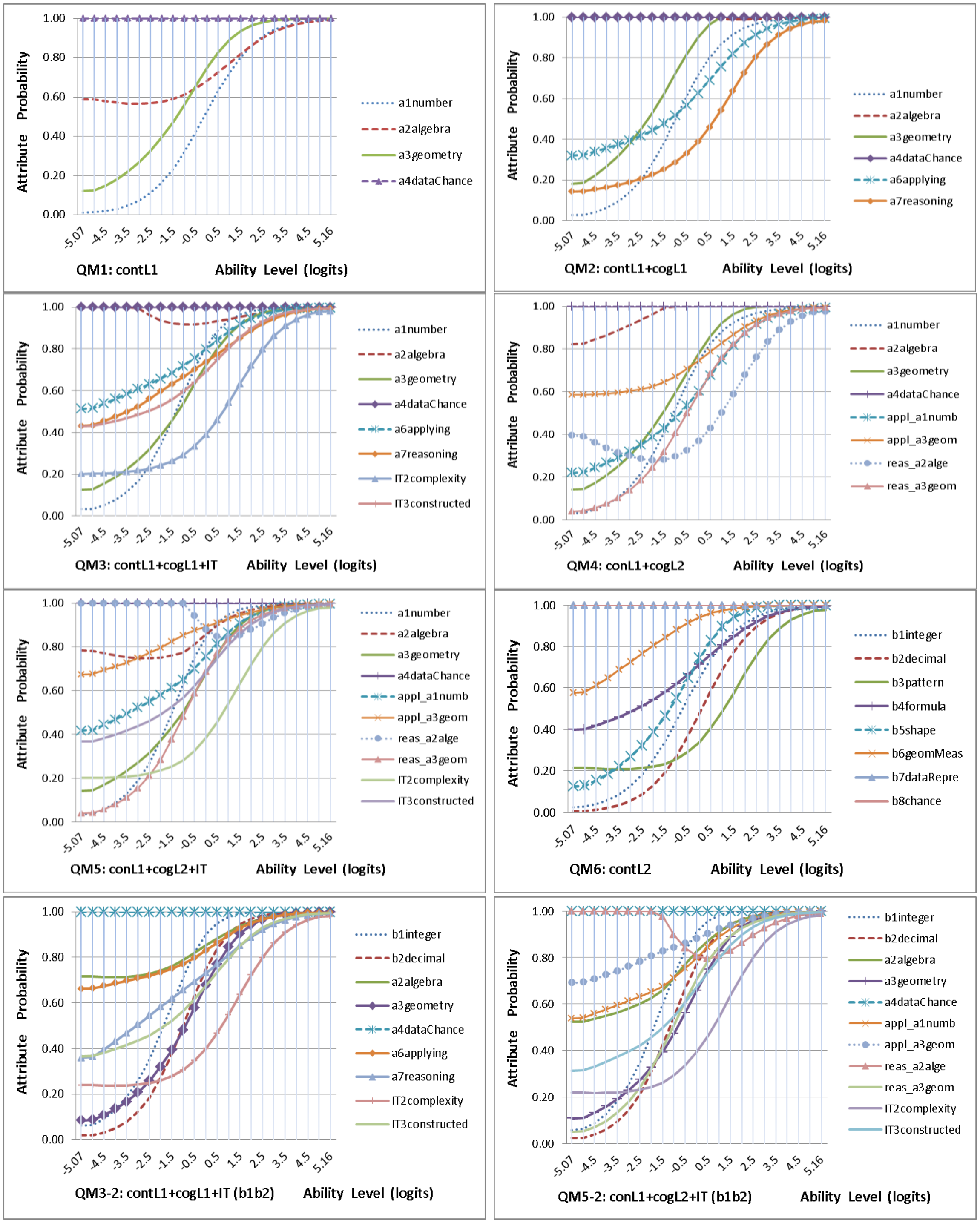

*Figure 20*. Attribute probability for the revised QM1 to QM5-2 of Booklet 2 (with TIMSS process attributes).

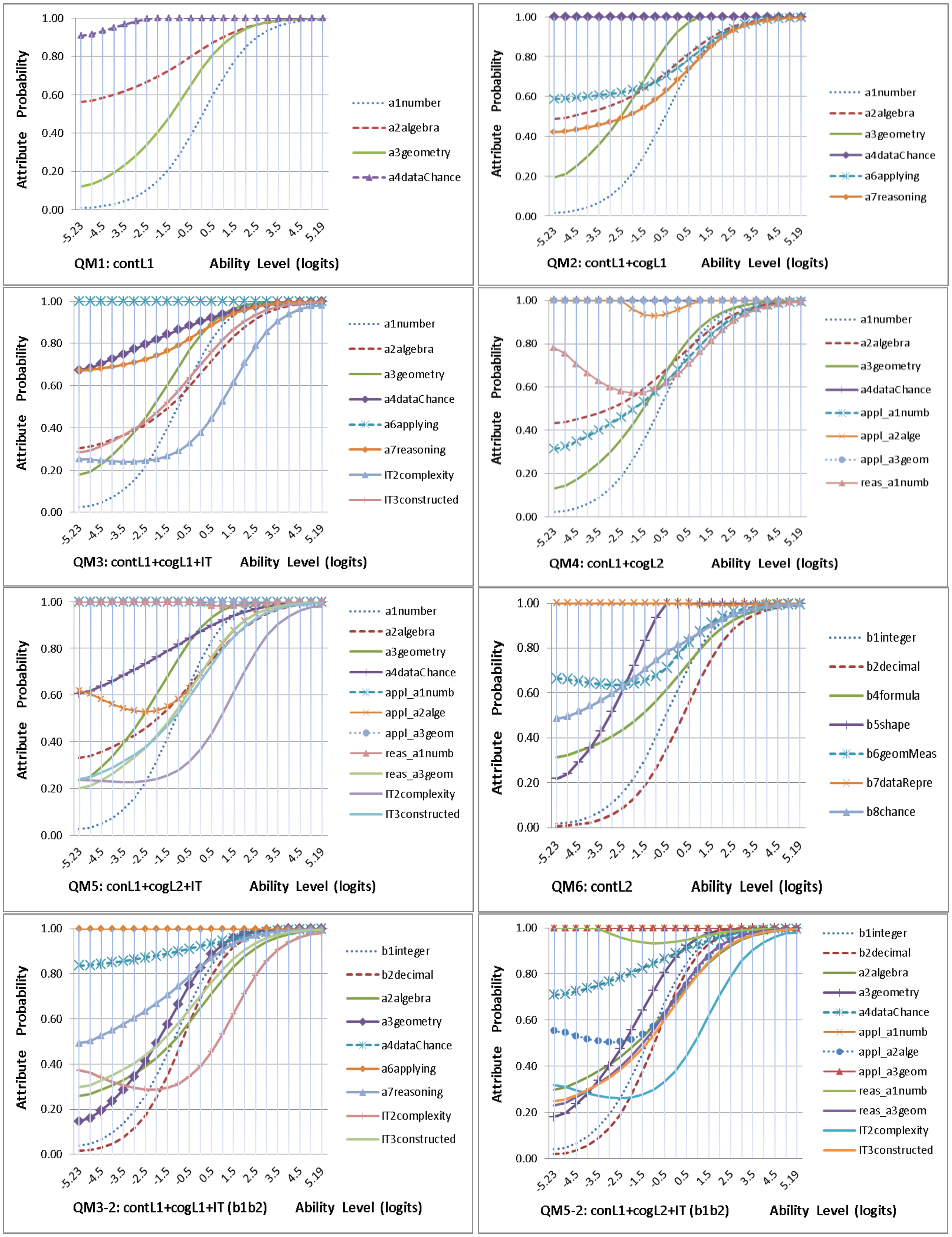

*Figure 21*. Attribute probability for the revised QM1 to QM5-2 of Booklet 3 (with TIMSS process attributes).

**LSDM: models with the new cognitive process attributes.**

***Least squares distance of Booklets 2 and 3.*** For both booklets, six models' LSDs decreased monotonically in a small range, from .28 to .001, with an increase in ability level (Figures 22 and 23). The models with the IT attributes (QM3 and QM5) had lower LSDs than QM2 and QM4 (Table 39). Including b1 and b2 in QM3-2 and b2 in QM5-2 reduced the LSDs. Except for QM4, Booklet 2 had lower LSDs than Booklet 3 in the other five models.

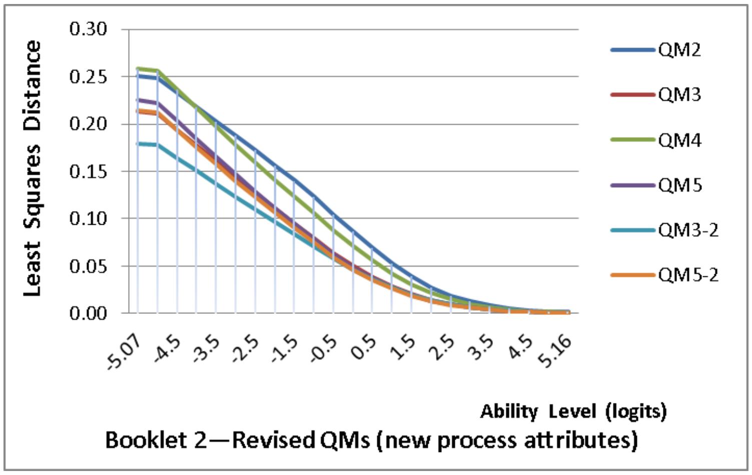

*Figure 22*. Least squares distance for the revised QMs of Booklet 2 (with new process attributes)

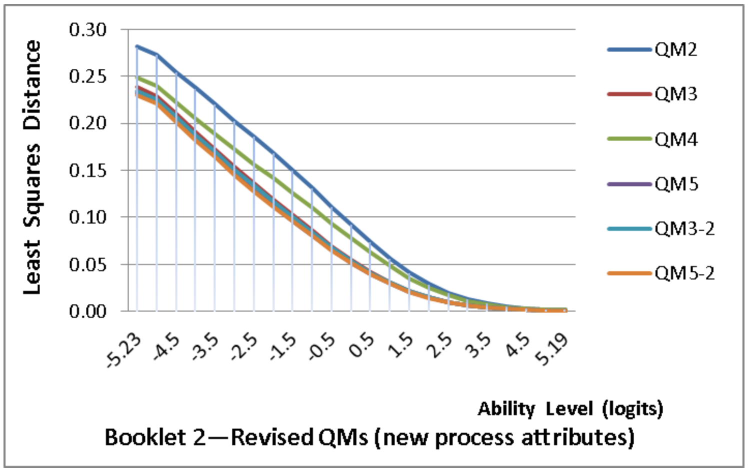

*Figure 23*. Least squares distance for the revised QMs of Booklet 3 (with new process attributes)

Table 39

*Mean Least Squares Distance for the Revised QMs of Booklet 2 (with New Process Attributes)*

| | | QM2 | QM3 | QM4 | QM5 | QM3-2 | QM5-2 |
|---|---|---|---|---|---|---|---|
| Booklet 2 | Mean LSD | .103 | .076 | **.096** | .078 | .066 | .074 |
| | Order of mean LSD | 6 | 3 | 5 | 4 | 1 | 2 |
| Booklet 3 | Mean LSD | **.111** | **.082** | .095 | **.079** | **.081** | **.078** |
| | Order of mean LSD | 6 | 4 | 5 | 2 | 3 | 1 |

*Note*: The bolded numbers are the mean LSDs for Booklets 2 or 3 that are higher than the corresponding mean LSDs for another booklet.

***Attribute probability curves of Booklets 2 and 3.*** For Booklet 2, the APCs of QM4 and QM5 were improved with the revised Q-matrix, but no obvious change was found in QM2 and QM3 (Figure 24). The addition of b2 refined the APCs of QM5-2, but not the APCs of QM3-2. All six models had more than 2/3 of the included attributes with acceptable APCs. Among them, all attributes' probabilities in QM4 and QM5-2 displayed a logical pattern in terms of monotonicity, relative difficulty, and discrimination. In QM2 and QM3, the probabilities of both "a2-number" and "a4-data and chance" were 100% across ability levels. In QM5, the probabilities of "d7-judging_operationRule" and "IT2-complexity" declined with an increase in ability level in the -5.23 ~ -2.0 range.

For Booklet 3, the APCs of the four models (QM2 ~ QM5) were improved with the revised Q-matrix (Figure 25). The use of b1 and b2 did not substantially change the APCs of QM3-2 and QM5-2. Every APC in QM4 was acceptable. The other five models had one unacceptable APC. No attribute's probability was 100% across all ability levels. The attributes with a problem APC were "a4-data and chance," "c3-judging," and "IT2-complexity," which did not increase monotonically across ability levels. In QM2, QM3, and QM3-2, the APC of a4 displayed a U-shape. In QM2, the probability of c3 decreased slightly in the lowest ability level range (from -5.23 to -3.0) with an increase in ability level, but the difference between the probability at the ability of -5.23 and the probability at the ability of -3.0 was only .03. In addition, in QM5 and QM5-2, the probability of IT2 decreased in the lowest range of ability level (from -5.23 to -1.5/-2.0) with an increase in ability level.

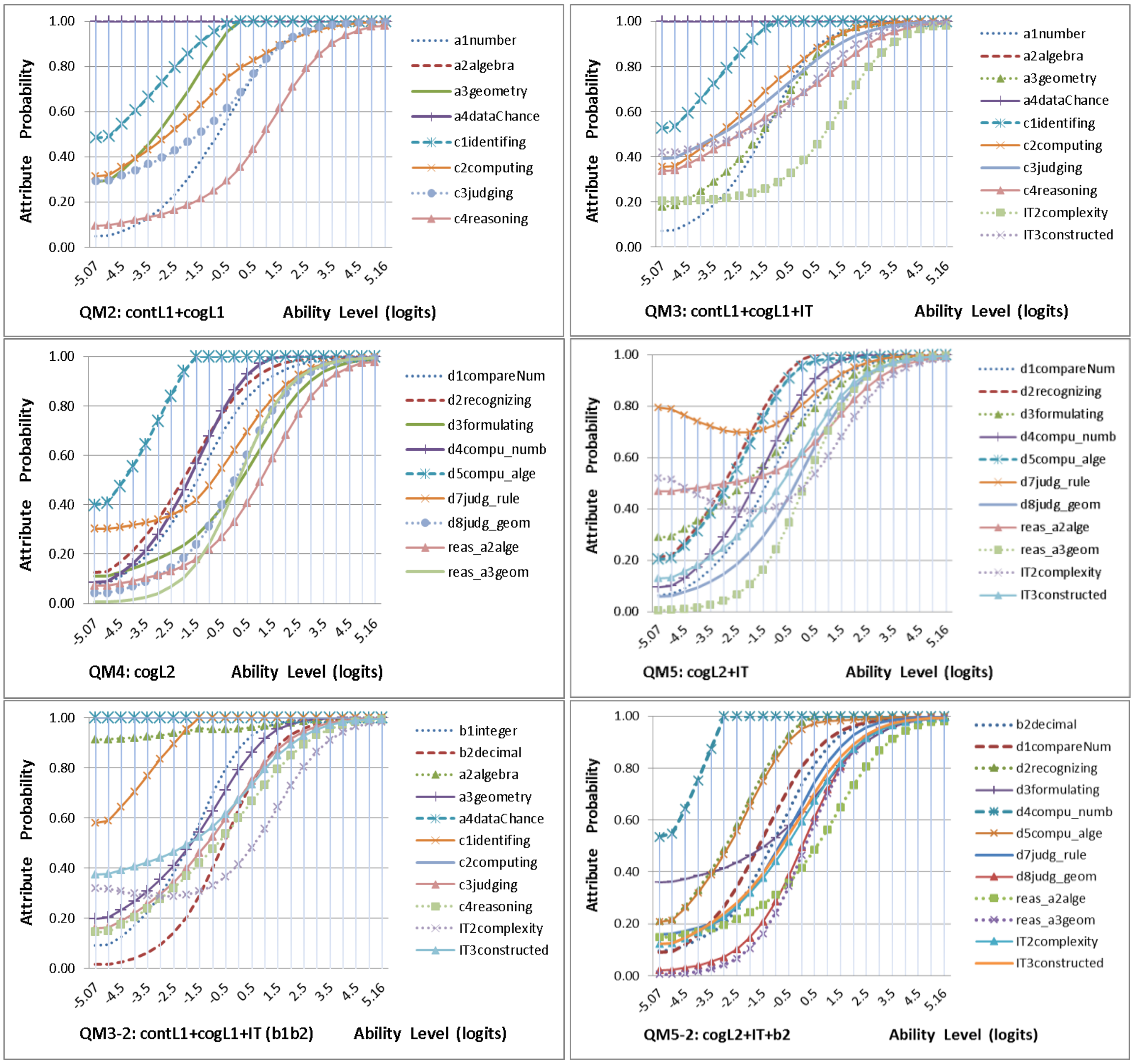

*Figure 24*. Attribute probability for the revised QM2 to QM5-2 of Booklet 2 (with new process attributes).

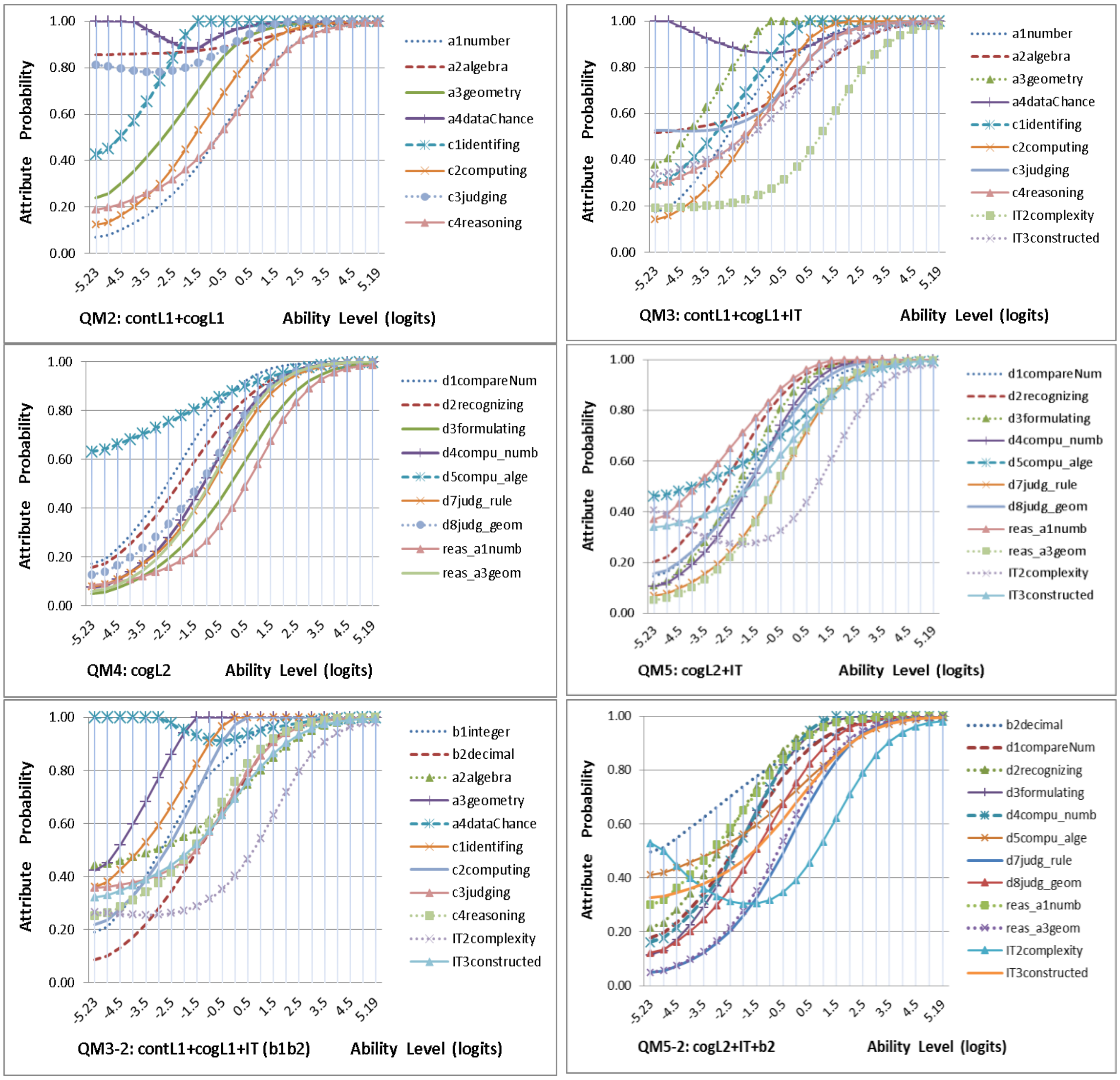

*Figure 25*. Attribute probability for the revised QM2 to QM5-2 of Booklet 3 (with new process attributes).

***Results summary of LSDM analysis***. Table 40 reports the summary results of the revised Q-matrices. With respect to LSD, both booklets generated similar results with the TIMSS cognitive process attributes and the new cognitive process attributes. The LSDs of all models declined monotonically in a small range, with an increase in ability level. Adding the process attributes and the complex process (IT) attributes reduced the LSD values. Also, including “b2-fractions, decimals, ratio, proportion, and percent” resulted in lower LSDs of four models, and improved the APC pattern of QM5-2 for Booklet 2. Although revision of the Q-matrices led to slightly worse LSDs in many models, more APCs of the revised models were acceptable. The APC of “a4-data and chance” still was not displayed reasonably in most models. A problem APC was also found for the sub-attributes of a4, such as “b7-data” and “b8-chance.” The probability of “a2-algebra” was not logically exhibited in the models with the level one attributes. Furthermore, the probability of “IT2-complexity” in several models declined with an increase in the lowest ability levels.

Comparing the results of Booklet 2 to Booklet 3, the LSDs and APCs indicated that the attributes explained the items of Booklet 2 better than the items of Booklet 3. Most models of Booklet 2 had lower mean LSDs than those of Booklet 3. With the TIMSS attributes, all eight models of Booklet 2 had at least 2/3 of APCs acceptable; but, every model had one or two attributes’ probabilities that did not increase with an increase in ability level. For Booklet 3, all APCs of QM1 were acceptable; but, only five models had at least 2/3 of APCs acceptable. With the new process attributes, all six models of both booklets had at least 2/3 of APCs acceptable. Overall, Booklet 3 had slightly more acceptable APCs. All QM4’s APCs of both booklets were displayed logically. When b2 was added, all APCs of QM5-2 for Booklet 2 became reasonable.

Table 40

*Results Summary of the Revised Q-Matrices Using Multiple Regression and the LSDM*

| | | | LSDM: LSD | | LSDM: APC | | | LSDM: Attributes with Problem APC | |
|---|---|---|---|---|---|---|---|---|---|
| | Q-matrix | Multiple Regression: Variance | Range | Mean LSD | Models with at Least 2/3 Acceptable Curves | Models: Very Good | Good Models if without considing a2, a4, b7 and b8 | probability = 100% | Non-monotonically increasing curve |
| Booklet 2 (8 models: withTIMSS process attributes) | Revised (A5) | 8 QMs': > A6's variance | .312 ~ .001 | 6 QMs': < A6's LSD<br>2 QMs': > A6's LSD | all 8 models | | QM1, QM2, QM3, QM3-2, QM6 | a2 (QM2), a4 (7 models excepting QM6), b7 and b8 (QM6) | a2 (QM1*, QM3, QM5*), reas_a2 (QM4, QM5, QM5-2) |
| Booklet 3 (8 models: withTIMSS process attributes) | Revised (A6) | | .311 ~ .001 | | 5 models (QM1, QM2, QM3, QM3-2, QM6) | QM1 | QM2, QM6 | a4 (QM2, QM4), b7(QM6), a6 (QM3, QM3-2), appl_a1 (QM5, QM5-2), appl_a3 (QM4, QM5, QM5-2), reas_a1 (QM4) | b6 (QM6*), appl_a2 (QM4, QM5, QM5-2), reas_a1 (QM5, QM5-2), IT2 (QM3-2, QM5-2) |
| Booklet 2—final (6 models: with new process attributes) | Revised (A7) | 6 QMs': > A8's variance | .259 ~ .001 | 5 QMs': < A8's LSD<br>1 QMs': > A8's LSD | all 6 models | QM4, QM5-2 | QM2, QM3 | a2 (QM2, QM3), a4 (QM2, QM3, QM3-2), c2 (QM3-2) | d7 (QM5), IT2 (QM3-2*, QM5) |
| Booklet 3—final (6 models: with new process attributes) | Revised (A8) | | .281 ~ .001 | | all 6 models | QM4 | QM2, QM3, QM3-2 | | a4 (QM2, QM3, QM3-2), c3 (QM2*), IT2 (QM5, QM5-2) |

* The difference between the probability at the lowest ability level and the lowest probability was less than .05.

**Results summary of the regression analysis and the LSDM analysis for the revised Q-matrices.**

For both booklets, the revised Q-matrices with the TIMSS process attributes and the new process attributes were validated by 28 models using two methods. The results of both analyses are summarized in Table 40. First, results generated from the level 1 attributes and the level 2 attributes were consistent. Second, most of the variance in item difficulty was explained by the Q-matrices with the IT attributes (.78 ~ .81 for Booklet 2, .57 ~ .64 for Booklet 3), but generally not by those without the IT attributes (.19 ~ .60 for Booklet 2, .05 ~ .25 for Booklet 3). The variances explained by the Q-matrix with only the content attributes were the smallest. Third, the LSDs of all models were small and displayed reasonably. Fourth, compared to the original Q-matrices, more revised models' APCs were improved, especially the models with the new process attributes. Most attributes' APCs were acceptable, but the APCs of a2 and a4 still were not logical in many models. For the following models, almost all APCs were acceptable: QM1 (contL1) of Booklet 3 and three models with the new process attributes for both booklets—QM4 (cogL2), QM5 (congL2+IT), and QM5-2 (cogL2+IT+b2). Finally, the Q-matrices explained the items for Booklet 2 better than for Booklet 3.

In sum, both the regression analysis and the LSDM analysis indicated that two Q-matrices (QM5 (congL2+IT) and QM5-2 (cogL2+IT+b2)) could effectively explain the difficulties of the mathematics items. These two Q-matrices included the information of all three categories of attributes. For QM5, the explained variances in item difficulty were $R^2$ = .78 and adjusted $R^2$ = .65 for Booklet 2, and $R^2$ = .63 and adjusted $R^2$ = .43 for

Booklet 3. For QM5-2, the variances explained were almost equal to those for QM5, respectively. For these two models, almost all attributes' APCs were acceptable.

**Comparing the TIMSS cognitive process attributes to the new cognitive process attributes.**

***The Q-matrix models of QM2 ~ QM5-2.*** This study had two sets of cognitive process attributes, one based on the TIMSS assessment framework (*knowing*, *applying*, and *reasoning*, called *the TIMSS process attributes*) and another based on hypothesized cognitive procedures (*identifying*, *computing*, *judging*, and *reasoning*, called *the new process attributes*). After revising the Q-matrices, each type of process attributes had 10 sub-attributes, with three same attributes for reasoning (d9 = reasoing_a1number, d10 = reasoning_a2algebra, d11 = reasoning_a3geometry). The Q-matrix models including these attributes were QM2 ~ QM5, QM3-2, and QM5-2.

Multiple regression analyses showed that overall, the Q-matrices with the TIMSS process attributes explained more variance in item difficulty than those with the new process attributes (Table 41). For Booklet 2, except for QM4, variances explained by the other five Q-matrices with the TIMSS process attributes were higher than those with the new process attributes. With the TIMSS process attributes, both explained $R^2$ values of QM3 and QM5 were .81, adjusted $R^2$ = .73 and .72. With the new process attributes, both $R^2$ values of QM3 and QM5 were .78, adjusted $R^2$ = 67 and .65. For Booklet 3, Two models (QM5 and QM5-2) with the TIMSS process attributes had slightly higher variances explained ($R^2$ = .64 and adjusted $R^2$ = .45 for QM5) than those with the new process attributes. Three models (QM3, QM4, and QM3-2) with the new process attributes had higher variances (*$R^2$* = .62 and adjusted *$R^2$* = .43 for QM3 and *$R^2$* = .25 and adjusted *$R^2$* = .43 for QM3) than those with the TIMSS process attributes.

Table 41

*Variances in Item Difficulty Explained by the Revised Q-Matrices: TIMSS Process Attributes Vs. New Process Attributes*

| Model | Attributes | Booklet 2 | | | Booklet 3 | | |
|---|---|---|---|---|---|---|---|
| | | # of Attributes | $R^2$ | Adjusted $R^2$ | # of Attributes | $R^2$ | Adjusted $R^2$ |
| TIMSS Process Attributes | | | | | | | |
| QM2 | contL1+cogL1 | 6 | **.48** | **.36** | 6 | .12 | **-.08** |
| QM3 | contL1+cogL1+IT | 8 | **.81** | **.73** | 8 | .57 | .43 |
| QM4 | contL1+cogL2 (= all cogL2) | 8 | .54 | .37 | 9 | .17 | -.15 |
| QM5 | contL1+cogL2+IT | 10 | **.81** | **.72** | 11 | **.64** | **.45** |
| QM3-2 | contL1+cogL1+IT (b1b2) | 9 | **.81** | **.72** | 9 | .57 | .40 |
| QM5-2 | contL1+cogL2+IT (b1b2) | 11 | **.81** | **.70** | 12 | **.64** | **.43** |
| New Process Attributes | | | | | | | |
| QM2 | contL1+cogL1 | 8 | .39 | .17 | 8 | **.13** | -.16 |
| QM3 | contL1+cogL1+IT | 10 | .78 | .67 | 10 | **.62** | **.44** |
| QM4 | cogL2 | 9 | **.60** | **.42** | 9 | **.25** | **-.04** |
| QM5 | cogL2+IT | 11 | .78 | .65 | 11 | .63 | .43 |
| QM3-2 | contL1+cogL1+IT (b1b2) | 11 | .78 | .66 | 11 | **.62** | **.42** |
| QM5-2 | cogL2+IT+b2 | 12 | .79 | .64 | 12 | .63 | .40 |

*Note*: The bolded numbers are the explained variances (with either the TIMSS process attributes or the new process attributes) higher than those for the corresponding models with another type of process attributes.

LSDM analyses of a total of 12 models for two booklets displayed that except for QM5-2 for Booklet 2, the other 11 models with the new process attributes had a little lower LSDs than those with the TIMSS process attributes, LSD differences between two booklets = .001 ~ .015. Also, more APCs of the models with the new process attributes displayed a logical pattern. QM4 for both booklets and QM5-2 for Booklet 3, which included all new process attributes at level two, had all APCs acceptable.

The APCs of the two QM4 showed that “d5-computing_algebra,” “d1-comparing numbers,” and “d2-recognizing” were relatively easy for students, while “d3-formulating” was relatively difficult. QM4 of Booklet 2 also indicated “d8-judging_geometry,” “reasoning_a3”, and “reasoning_a2” were relatively difficult. “D7-judging_operationRule” and “reasoning_a1” were relatively difficult for students who took Booklet 3. When two IT attributes were added, “IT3-constructed-response” was generally easier than “IT2-complexity.” For those students with ability above the average level, IT2 was generally the most difficult attribute. When the IT attributes were added, the probability curves of “reasoning_a2” for Booklet 2 and “reasoning_a1” for Booklet 3 moved up obviously, suggesting there were compound effect between the IT attributes and “reasoning_a2” or “reasoning_a2.”

***The Q-matrix models of QM9 ~ QM11.*** The study further analyzed the models with only the two types of cognitive process attributes at level one (QM9 = IT, QM10 = cogL1, QM11 = cogL1+IT). This analysis compared effects of individual type of attributes on the item difficulties and item performance, and investigated the Q-matrix models without the content attributes.

The multiple regression analysis showed that with only the two IT attributes, the Q-matrix (QM9) accounted for .59 of the variance (adjusted $R^2$ = .56) in the item difficulties for Booklet 2, and .43 (adjusted $R^2$ = .39) for Booklet 3 (Table 42). With only the level 1 TIMSS process attributes (QM10), the explained $R^2$ and adjusted $R^2$ estimates were .31 and .26 for Booklet 2, and .06 and .001 for Booklet 3, respectively. When both types of the cognitive process attributes—IT and TIMSS process attributes were included in QM11, the explained $R^2$ and adjusted $R^2$ values increased to .64 and .58 for Booklet 2, and .45 and .37 for Booklet 3, respectively. Moreover, for the model with only the new process attributes (M10), the explained $R^2$ and adjusted $R^2$ estimates were .34 and .24 for Booklet 2, and .12 and -.01 for Booklet 3, respectively. Following this, when the IT attributes were added (M11), the explained variances increased, $R^2$ = .74 and adjusted $R^2$ = .68 for Booklet 2, and $R^2$ = .55 and adjusted $R^2$ = .45 for Booklet 3.

Table 42

*Variances in Item Difficulty Explained by the Q-Matrices for QM1, QM6, and QM9 ~ QM11*

| Model | Attributes | Booklet 2 | | | Booklet 3 | | |
|---|---|---|---|---|---|---|---|
| | | # of Attributes | $R^2$ | Adjusted $R^2$ | # of Attributes | $R^2$ | Adjusted $R^2$ |
| *QM1* | *contL1* | *4* | *.19* | *.06* | *4* | *.05* | *-.09* |
| *QM6* | *contL2* | *8* | *.31* | *.06* | *7* | *.14* | *-.10* |
| QM9 | IT | 2 | .59 | .56 | 2 | .43 | .39 |
| TIMSS Process Attributes | | | | | | | |
| QM10 | cogL1 (a6a7) | 2 | .31 | **.26** | 2 | .06 | **.001** |
| QM11 | cogL1+IT | 4 | .64 | .58 | 4 | .45 | .37 |
| New Process Attributes | | | | | | | |
| QM10 | cogL1 (c1~c4) | 4 | **.34** | .24 | 4 | **.12** | -.01 |
| QM11 | cogL1+IT | 6 | **.74** | **.68** | 6 | **.55** | **.45** |

*Note*: The bolded numbers are the explained variances (with either the TIMSS process attributes or the new process attributes) higher than those for the corresponding models with another type of process attributes.

Comparing the Q-matrix models with each of the three types of attributes, the adjusted $R^2$ estimates displayed that the variances explained by the cognitive process attributes (either TIMSS or new: adjusted $R^2$ = .26/.24 for Booklet 2 and .001/-.01 for Booklet 3) were higher than those by the content attributes (adjusted $R^2$ = .06/.06 for Booklet 2 and -.09/-.10 for Booklet 3); and the variances explained by the IT attributes (adjusted $R^2$ = .56 for Booklet 2 and .39 for Booklet 3) were much higher than those by both the content and process attributes. Compared to the model with only the TIMSS process attributes (adjusted $R^2$ = .26 for Booklet 2 and .001 for Booklet 3), the new process attributes explained slightly less of the variance in item difficulty (adjusted $R^2$ = .24 for Booklet 2 and -.01 for Booklet 3). When the IT attributes were included, the adjusted $R^2$ by the Q-matrices with the new process attributes were higher than those with the TIMSS process attributes.

The LSDM analysis showed that for QM10 and QM11 with either the TIMSS or new process attributes, the LSDs decreased monotonically in a small range, from .31 to .001 (Figures 26 to 29). However, the LSDs for QM9 with only the IT attributes were within a large range, from .68 to .001, suggesting that it would be not appropriate to recover ICC with only the IT attributes although the two APCs for QM9 exhibited good patterns in Figures 30 and 31.

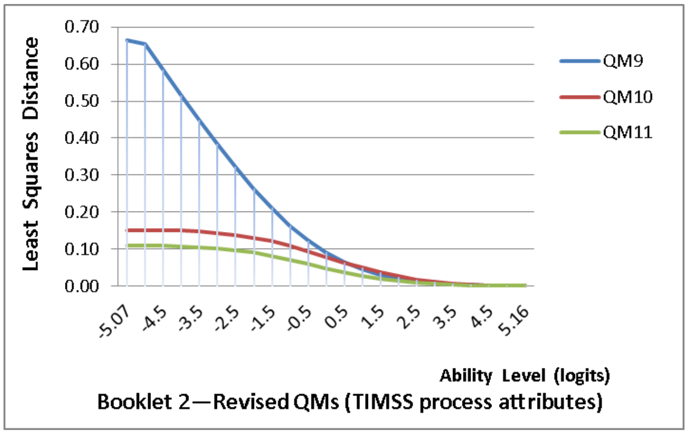

*Figure 26*. Least squares distance for QM9 to QM11 of Booklet 2 (with TIMSS process attributes)

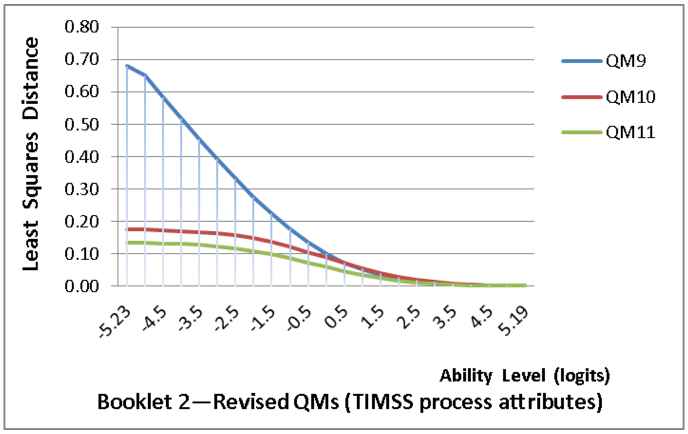

*Figure 27*. Least squares distance for QM9 to QM11 of Booklet 3 (with TIMSS process attributes)

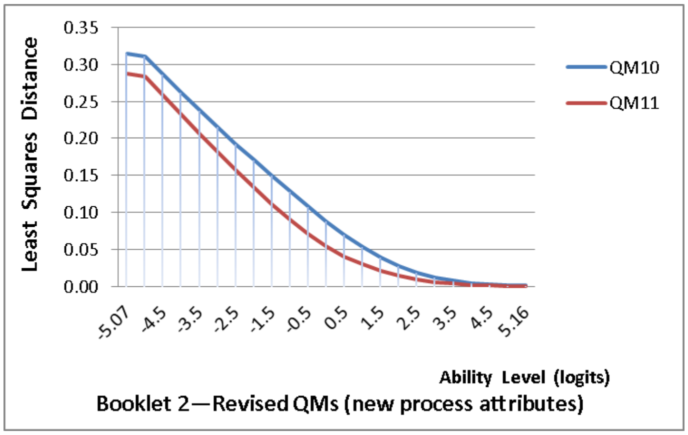

*Figure 28*. Least squares distance for QM10 and QM11 of Booklet 2 (with new process attributes)

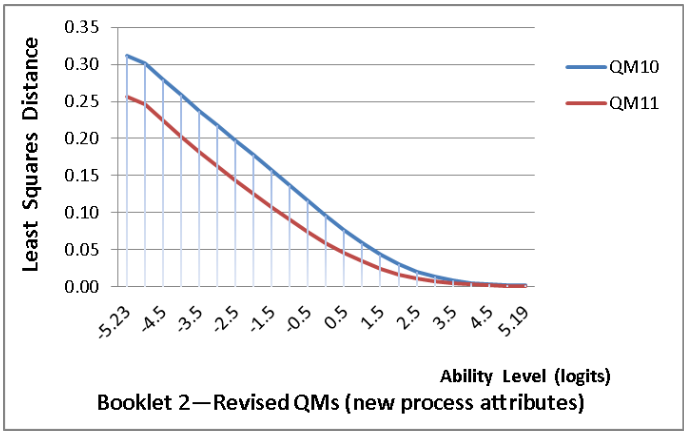

*Figure 29*. Least squares distance for QM10 and QM11 of Booklet 3 (with new process attributes)

The APCs of the new process attributes at level one displayed a better pattern than the APCs of the TIMSS process attributes at level one (Figures 30 and 31). Because “a5-knowing” was specified for all items, its APC was always a perfect S-shape; as a result, it could not show its relative difficulty to “a6-applying” and “a7-reasoning.” In general, a6 was relatively easier than a7. When the IT attributes were added, the probability of a6 became a constant of 100% across ability levels for Booklet 3, suggesting that a6 might not effectively discriminate student skills.

For the new process attributes at level one , QM10 of Booklet 2 clearly showed the order of difficulty of four attributes, from “c1-identifying” (easiest), “c3-judging,” “c2-computing,” to “c4-reasoning” (most difficult: Figure 27). QM10 of Booklet 3 also revealed that c1 and c3 were easier than c2 and c4. When the IT attributes were added, the APCs of the four process attributes changed slightly; but overall, their relative difficulty did not change. Both APCs of IT2 and IT3 crossed the other APCs. IT2 was still the most difficult attribute. But in QM11 of Booklet 2, the probability of IT2 decreased with an increase in ability level, from -5.07 to -2.0, suggesting that IT2 did not effectively measure the mathematical skills of students at the lowest ability levels.

In sum, the Q-matrices with the TIMSS process attributes explained a little more variance in item difficulty than those with the new process attributes. However, the LSDM analyses indicated that the probabilities of the new process attributes exhibited more logical patterns in terms of monotonicity, relative difficulty, and discrimination. Thus, the new process attributes were better than the process attributes based on the TIMSS assessment framework.

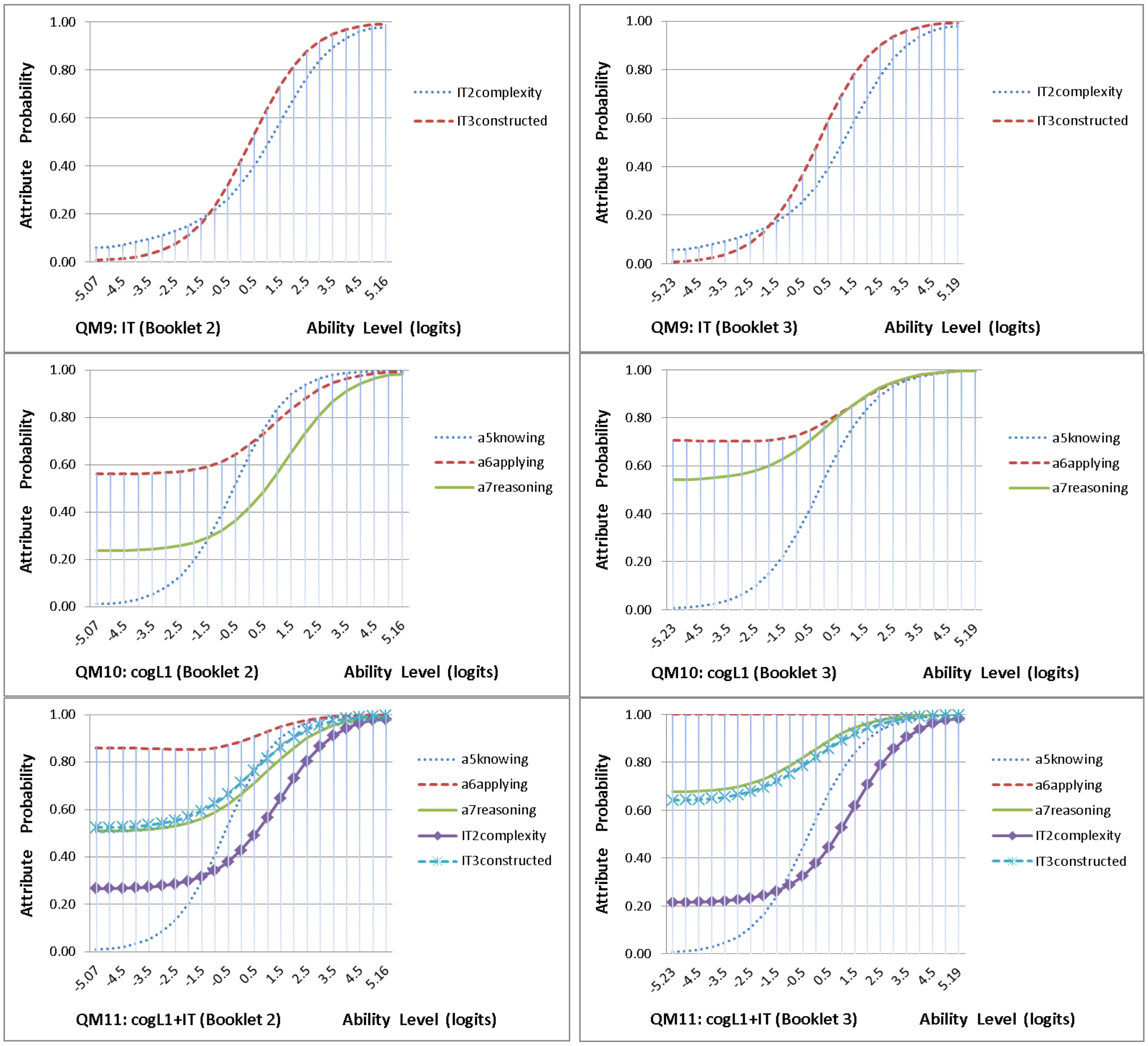

*Figure 30*. Attribute probability for QM9 to QM11 of Booklets 2 and 3 (with level 1 TIMSS process attributes).

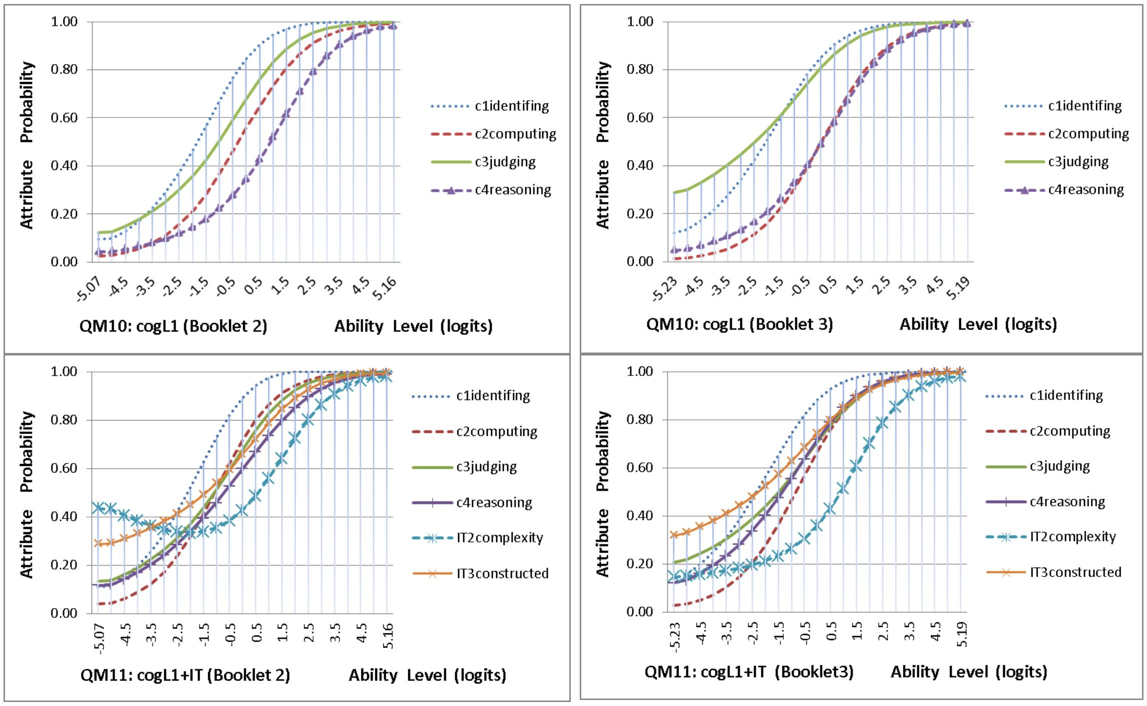

*Figure 31*. Attribute probability for QM10 and QM11 of Booklets 2 and 3 (with level 1 new process attributes).

### Recovery of item characteristic curve with the LSDM.

The Q-matrices of QM4 (cogL2) and QM5-2 (cogL2+IT+b2) with the new process attributes were used to recover the ICC of every item in the two booklets. The reasons for choosing these two models were that (1) almost all APCs in the two models were acceptable (excepting IT2 in QM5-2 for Booklet 3, its probability decreased when the ability level went up from -5.23 to -2.0); and (2) these two models included all new process attributes at level two, which also covered the content attributes at level one. The mean absolute difference (MAD) between the ICC recovery and the ICC estimated by the Rasch model was calculated for each item of Booklets 2 and 3 (Figures 32 and 33).

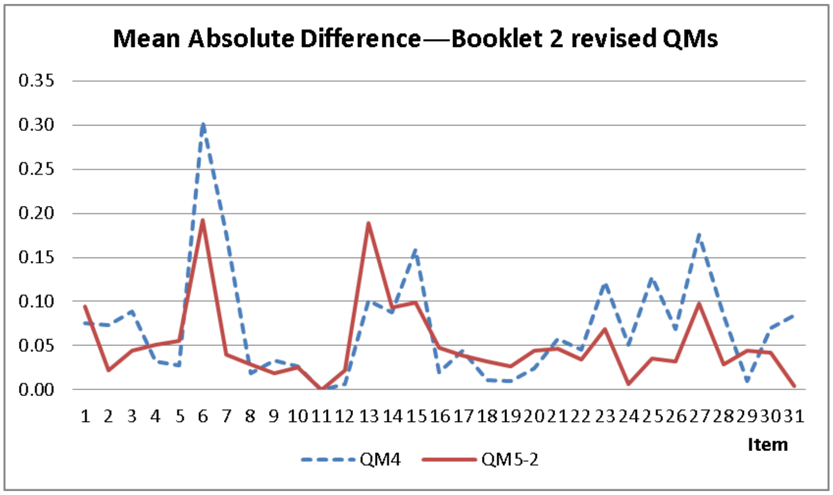

*Figure 32*. Mean absolute difference of the Booklet 2 items for QM4 and QM5-2.

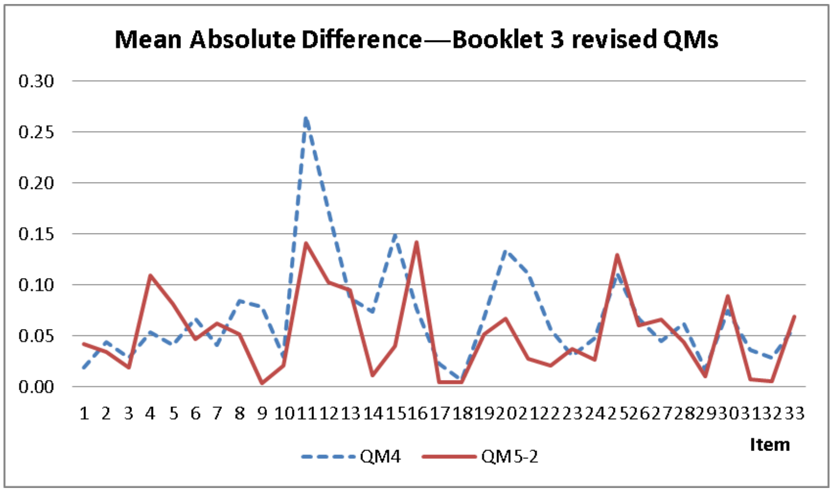

*Figure 33*. Mean absolute difference of the Booklet 3 items for QM4 and QM5-2.

Based on the MAD values, numbers of items with different recovery degrees are reported, as well as the proportion of items falling into the categories of “somewhat poor,” “poor,” and “very poor” (Table 42). For QM4, the ICC recovery of Booklet 3 was better than that of Booklet 2. For QM5-2, the ICC recovery of Booklet 2 was slightly better than Booklet 3. Overall, the items had lower MAD values when they were recovered by the attributes in QM5-2, which had the IT attributes and b2. The recovery of several items was improved substantially when adding the attributes IT2 and b2, such as Items M02-7A, M02-13, M03-7, M03-9, M03-11, M03-12, M03-15, M04-5A, and M04-5B. However, some items still were not well recovered by the proposed attributes. The two items with a poor recovery (both MADs = .19) in QM5-2 had only one attribute; these items were M02-6 (measuring “d5-computing_algebra”) and M02-11 (measuring “d2-recognizing”).

Table 43

*Number of Items with Different Recovery Degrees*

| Item Recovery Degree | Mean Absolute Difference | Booklet 2 | | Booklet 3 | |
|---|---|---|---|---|---|
| | | QM4 | QM5-2 | QM4 | QM5-2 |
| Very good | 0 ≤ MAD < 0.02 | 5 | 3 | 1 | 7 |
| Good | 0.02 ≤ MAD < 0.05 | 9 | 17 | 11 | 10 |
| Somewhat good | 0.05 ≤ MAD < 0.10 | 10 | 7 | 15 | 11 |
| Somewhat poor | 0.10 ≤ MAD < 0.15 | 3 | 2 | 3 | 5 |
| Poor | 0.15 ≤ MAD < 0.20 | 3 | 2 | 2 | 0 |
| Very poor | MAD ≥ 0.20 | 1 | 0 | 1 | 0 |
| Proportion of items with poor MAD* | | 0.23 | 0.13 | 0.18 | 0.15 |

* Proportion of items with poor MAD = number of items within somewhat poor, poor, and very poor ÷ total number of items

The 15 items of assessment Block M03 were included in both Booklets 2 and 3. These items’ MADs are reported by booklet in Figure 34. For the same model, the MADs of the same items in two booklets were different. Overall, the ICC recovery of Booklet 3 was slightly worse than that of Booklet 2. The results suggested that the required attributes held the Booklet 2 items better than the Booklet 3 items.

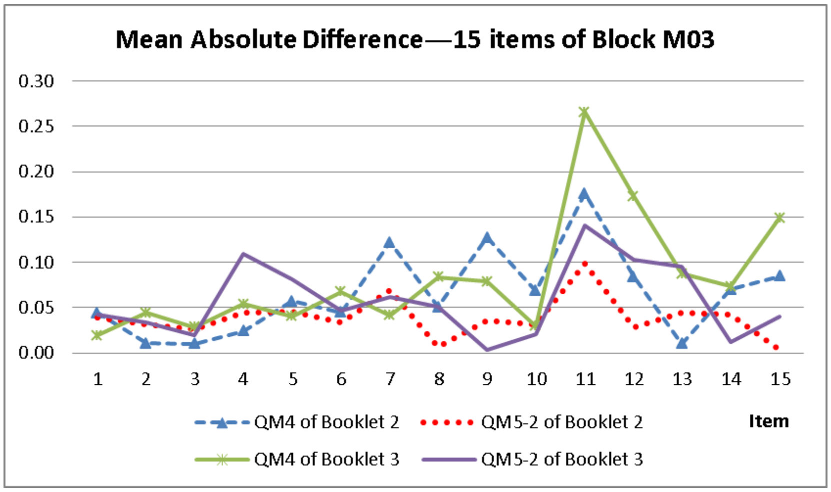

*Figure 34*. Mean absolute difference of the items in Block M03 for QM4 and QM5-2.

# Chapter Four

## Discussion

No matter the kind of assessment, the key point is to improve learning. Assessment should serve to improve learning, and provide support for learning and instruction, instead of testing for the sake of testing. Although nowadays approaches to assessment are much improved over decades past, the educational assessment community still faces challenges to integrate assessment into the learning environment. Cognitive diagnostic assessment was developed under multiple streams of influence, such as educational policy and goals, development of educational measurement, and advances in computational science and computational techniques. CDA generates more detailed information about students' performance profiles than the scores derived from traditional CTT or IRT approaches; as a result, CDA supports better learning and teaching (Nichols, 1994; Pellegrino, 2004). Moreover, a useful assessment should be rooted in reliable design of the test items, which is vital for widely administrated assessments, such as TIMSS. Research on CDA and CDMs aids in advancing the quality of educational assessments. As the largest and most comprehensive study of educational achievement, TIMSS provides a cross-national perspective on the trends in student achievement, school organizational and instructional practices, and education systems. However, studies of TIMSS assessment using CDMs are limited.

For many CDMs, the item-attribute matrix is an essential component. However, the construction of attributes and a reliable Q-matrix is a challenge. Study of the Q-matrix and CDMs is still at a very beginning stage (Leighton & Gierl, 2007). Researchers investigated impacts of the Q-matrix misspecification, approaches to validation of the Q-matrix, and exploratory methods to develop a Q-matrix using different CDMs. However, many issues still are not settled in building a Q-matrix with reliable quality. Thus, the present study explored the validation of the Q-matrix based on the TIMSS mathematics items using multiple regression and the LSDM.

This chapter presents a summary of the study, major findings with respect to the research questions, limitations of the study, and recommendations for future study.

**Summary of the Study**

The purposes of this study were two-fold: (1) to validate the item-attribute matrix using two levels of attributes, and (2) through retrofitting the diagnostic models to the TIMSS test, to evaluate the construct validity of TIMSS mathematics assessment by comparing the results of two assessment booklets in TIMSS 2007.

The data used were from the released mathematics items for the 8th grade in TIMSS 2007. The present study analyzed the students' performance on the 49 items of three assessment blocks (M02, M03, and M04) in Booklets 2 and 3. According to a review of the literature, the TIMSS assessment framework, and the 49 mathematics items, an attribute pool were developed with respect to the TIMSS assessment domains: the content domains and the cognitive domains. Using the two levels of proposed attributes, the researcher conducted a pilot study: specified the relationship between the items and the attributes and analyzed the draft Q-matrix based upon the LLTM. Finally, the researcher developed 7 level 1 attributes (4 content and 3 cognitive process) and 20 level

2 attributes (8 content and 12 cognitive process) based on the TIMSS assessment domains and 3 IT attributes based on item type.

The 7 attributes at level one included four *content attributes* (contL1: a1-*number*, a2-*algebra*, a3-*geometry*, and a4-*data and chance*) and three *cognitive process attributes* (cogL1: a5-*knowing*, a6-*applying*, and a7-*reasoning*). At level two, every level 1 content attribute was divided into two sub-attributes; as a result, there were 8 level 2 *content* attributes (contL2: b1 ~ b8). At level two, every level 1 cognitive process attribute was further classified into four sub-skills according to the content domains; thus, there were 12 level 2 cognitive process attributes (cogL2: *know_a1/a2/a3/a4*, *appl_a1/a2/a3/a4*, and *reas_a1/a2/a3/a4*). Moreover, based on item type, three *comprehensive cognitive process attributes* (or called *the IT attributes*) were identified, which represented integrated and more comprehensive cognitive processing that could not be defined with the listed cognitive process attributes. There were only one level of the IT attributes: IT1-*multiple steps and/or responses*, IT2-*complexity*, and IT3-*constructed-response*.

Compared to the content attributes, the cognitive process attributes were harder to define. Limitations were found for the cognitive process attributes based on the TIMSS assessment framework (called *the TIMSS cognitive process attributes*). Then, the 2nd classification of cognitive process attributes (called *the new cognitive process attributes*) were produced according to further analyzing the procedures required to solve the 49 test items. The new process attributes encompassed 4 level 1 attributes (c1-*identifying*, c2-*computing*, c3-*judging*, and c4-*reasoning*) and 11 level 2 attributes (d1 ~ d11).

In total, four Q-matrices were developed for the 49 items, two of them were specified and two of them were randomly produced according to the specified Q-matrices.

*The TIMSS Q-matrix* were specified by three experts and the researcher, using the attributes based on the TIMSS assessment framework (with TIMSS content attributes + TIMSS process attributes + IT attributes). Then, the TIMSS process attributes were replaced by the new process attributes and *the 2nd Q-matrix* was formed (with TIMSS content attributes + new process attributes + IT attributes). The relationship between the items and the new process attributes was specified by the researcher. The specified Q-matrices were cross-validated through comparing results between the specified Q-matrices and the random Q-matrices, between the Q-matrices with the TIMSS process attributes, and with the new process attributes, and also between Booklet 2 and Booklet 3.

The 49 items were grouped into two booklets. Therefore, four Q-matrices were generated for each booklet. Validation of each Q-matrix was investigated with eleven Q-matrix models (QM1 ~ QM11) using multiple regression and the LSDM. Based on the LLTM, multiple regression examined variance in item difficulty explained by the Q-matrices. A higher variance explained by a Q-matrix suggests that the Q-matrix is more reliable. The LSDM investigated performance on the attributes and extent of the item characteristic curve (ICC) recovery by the attributes' probabilities. According to the LSDM, the smaller the LSDs, the better the attributes or the Q-matrix explain the items; and also, the attribute probability curves (APCs) exhibit a reasonable pattern in terms of monotonicity, relative difficulty, and discrimination. In addition, a smaller mean absolute difference (MAD) between the recovered ICC and the estimated ICC suggests a good recovery. With respect to the MAD, Dimitrov (2007) suggested six recovery degrees, from "very good" to "very poor".

Eight Q-matrix models (QM1 ~ QM8) were implemented to explore effect of different categories of attributes (content, cognitive process, and IT) at the two levels. QM1 ~ QM3 (QM3 = contL1+cogL1+IT) tested the Q-matrices with the attributes at level one. QM4 and QM5 (QM5 = contL1+cogL2+IT) investigated the combination of the level 1 content attributes and the level 2 process attributes. QM6 ~ QM8 (QM8 = contL2+cogL2+IT) analyzed the Q-matrices with the attributes at level two. Each category of attributes was added step by step for the comparison analyses.

Based on a series of analyses, the Q-matrices were further refined with respect to the problem APCs of some attributes. Three attributes were re-defined ("b1-whole numbers and integers," "b7-data," and "b8-chance"). Four attributes were deleted ("applying_a4data and chance," "reasoning_ a4data and chance," "d6-judging_number," and "IT1-multiple steps/responses"). Due to the complexity, overlap among the attributes, and the problem APCs, the Q-matrix models QM7 and QM8 were excluded from the final analyses. Also, QM4 and QM5 with the new process attributes were re-modeled because of overlap between the level 1 content attributes and the level 2 cognitive process attributes. In the two models with the IT attributes (QM3-2 and QM5-2), the attribute "a1-number" were replaced by b1 and "b2-fractions, decimals, ratio, proportion, and percent" because a1 were required by a majority of the items and so, the APC of a1 did not display meaningful. In addition, the relationship between the attributes and the items were re-examined; and a few elements of the Q-matrices were revised.

Following this, the revised Q-matrices were further analyzed using multiple regression and the LSDM, with eleven Q-matrix models (QM1 ~ QM8 and QM9 ~ QM11). The models of QM9 ~ QM11 compared the Q-matrices with only one type of attributes at level one. The Q-matrices with the three types of attributes accounted for most of the variance in item difficulty; and most ACPs exhibited a logical pattern in terms of monotonicity, relative difficulty, and discrimination. Overall, the proposed Q-matrices were acceptable, but with a few unreasonable APCs. For Booklet 3, the Q-matrices without the IT attributes explained low variances in item difficulty, with negative adjusted $R^2$ values, and there were more problem APCs, indicating that the items of Booklet 3 were not well explained by the proposed attributes and the Q-matrices.

Except for the attribute "IT2-complexity" for Booklet 3, all APCs of QM4 and QM5-2 for both booklets were reasonable. Therefore, QM4 and QM5-2 were chosen to analyze degree of the ICC recovery by the LSDM.

In sum, the present study analyzed 12 Q-matrices for two booklets: 4 originally specified Q-matrices, 4 random Q-matrices, and 4 revised Q-matrices. Among them, 6 Q-matrices involved the TIMSS process attributes and 6 Q-matrices included the new process attributes. Each Q-matrix was analyzed with the eight Q-matrix models; also, the revised Q-matrix was tested with QM9 ~ QM11. Thus, in total, the validation study tested 98 models using multiple regression and the LSDM. The proposed Q-matrices were validated and cross-validated through comprehensive analyses.

**Major Findings**

With respect to each research question, major findings are summarized and discussed below.

Research question one: *Which type of attributes contributes more to item difficulty: content, cognitive process, or complex cognitive process (item type)?* To address this question, the eleven Q-matrix models (QM1 ~ QM11) were employed; and, each category of attributes (content, cognitive process, and IT) at the two levels were added step by step. Consistent results were found from the specified Q-matrices, original and revised, with the TIMSS process attributes and the new process attributes. Results demonstrated that: the content attributes accounted for a very small proportion of the variance in item difficulty; the cognitive process attributes accounted for more variance in item difficulty than the content attributes; the complex cognitive process attributes accounted for much more variance than both the content and process attributes.

For Booklet 2, when the cognitive process attributes, at both level one and level two, were added to the models with only the content attributes, the explained adjusted $R^2$ values became over twice as high as those of the models with only the content attributes. For Booklet 3, when the cognitive process attributes were added to the models with only the content attributes, the explained adjusted $R^2$ values slight increased or even decreased. However, the models with each type of attributes (QM10 *vs.* QM1/QM6) displayed that the adjusted $R^2$ values explained by only the cognitive process attributes were higher than those by the content attributes. Results suggested that the ability to apply knowledge and reasoning with knowledge may be more important than just remembering knowledge. However for many assessments, test scores generally are tied to content domains rather than cognitive mechanisms (Nichols, 1994); based upon such test scores, it would be hard to provide effective support for better teaching and learning.

Generally, the content attributes and the cognitive process attributes in total explained less than 50% of the variance in item difficulty. When the IT attributes were included, all explained $R^2$ values increased significantly ($p \leq .003$); most variance in item difficulty (81% for Booklet 2 and about 60% for Booklet 3) was explained by the specified Q-matrices. It could be that the IT attributes were easier to identify than the cognitive process attributes; and so, the IT attributes were more predictive. Also, results indicated that item type, which represents integrated cognitive processing, played a more important role in deciding the item difficulties. Thus, design of test items not only need to consider knowledge topics and cognitive skills to be measured, but also should consider item type comprehensively, such as wording, length of item, distracting information, and use of multiple choice or constructed-response formats.

Also, as expected, almost all of the variances explained by the random Q-matrices were smaller than those explained by the specified Q-matrices. This indicated that the specified Q-matrices, which were constructed based upon the underlying content and cognitive behaviors, were better the Q-matrices that were randomly generated. The development of CDA and CDMs endeavors to integrate cognitive psychology with educational assessment. An adequate Q-matrix, to some degree, can reveal factors affecting cognitive processing.

However, the cognitive processing for a problem could be a complex mechanism, within which many factors would affect each other reciprocally. As a result, it would be difficult to demonstrate the mechanisms of solving some mathematics items with several simple content and process attributes, which is supported by the large percentage of variance in item difficulty explained by the IT attributes. Also, some items required the same attributes; but, their item difficulties were obviously different. For Booklet 3, about 40% of the variance in item difficulty was not explained by the proposed attributes and the models without the IT attributes had negative adjusted $R^2$ values, indicating the unfound attributes still largely affect the item difficulties. Thus, there is still room for refining the attributes and improving the Q-matrices.

Moreover, the different attributes' contribution to item difficulty suggested that CDA can generate richer information about students' strength and weakness than the overall scores estimated by the CTT and IRT methods. For the 49 items, a correct response to an item received one score point or two score points (only for two items). However, the items measured different knowledge and cognitive process; and their difficulties were significantly different. Thus, if based on feedback from a CDA, students would be provided specific remedial instructions with respect to their skill profiles.

Research question two: *Do the new cognitive process attributes provide an explanation of the items? What different results were found between the TIMSS cognitive process attributes and the new cognitive process attributes?* With respect to the limitations of some TIMSS process attributes, the new process attributes were proposed. After revision, the TIMSS process attributes included 3 attributes at level one and 10 attributes at level two; the new process attributes contained 4 attributes at level one and 10 attributes at level two. These two types of attributes had a common attribute "*reasoning*" at level one and three common reasoning attributes at level two.

Results of QM2 ~ QM5-2 displayed that the new process attributes explained the item difficulties well. For QM3 (contL1+cogL1+IT) and QM5 (cog2+IT), the explained variances in item difficulty were about .78 and .62 for Booklets 2 and 3, respectively. The LSDs decreased monotonically in a relatively small range. The APCs of all new process attributes, at both levels, exhibited a reasonable pattern.

Overall, the explained variances in item difficulty by the Q-matrices with the new process attributes were close to those with the TIMSS process attributes. For Booklet 2, excepting QM4, the variances explained by the Q-matrices with the TIMSS process attributes were slightly higher than those with the new process attributes. However, for Booklet 3, excepting QM5 and QM5-2, the variances explained by the Q-matrices with the new process attributes were a little higher than those with the TIMSS process attributes.

The LSDM analyses of both booklets indicated that the Q-matrices with the new process attributes interpreted the items better than those with the TIMSS process attributes. Except for QM5 and QM5-2 of Booklet 2, the mean LSDs of the other 10 models with the new process attributes were less than those with the TIMSS process attributes. More APCs of the TIMSS process attributes did not display logically than those of the new process attributes. Except for "c2-computing" (QM3-2) and "d7-judging_operationRule" (QM5) for Booklet 2, the APCs of other new process attributes at both levels exhibited a meaningful pattern. Almost all APCs of QM4 and QM5-2 were acceptable.

In sum, according to all analysis results and the limitations of some TIMSS process attributes, the new process attributes explained the mathematics items better than the TIMSS process attributes.

Research question three: *What differences are generated from the level 1 attributes and the level 2 attributes?* Using attributes at a very small grain size could lead to unstable parameters and overly complex diagnostic feedbacks about students' performance. On the other hand, the coarse-grained attributes do not produce detailed information about students' knowledge states. Thus, choosing an appropriate grain size for attributes is very important for a CDM. However, the literature review shows that studies of CDM and the Q-matrix generally employed one level of attributes.

In this study, two levels of attributes were developed for the content attributes and the process attributes. There was only one level of the IT attributes. At level two, each level 1 content attributes were divided into two sub-attributes. The TIMSS process attributes at level two were classified based on the level 1 process attributes and the level 1 content attributes. At level two, each new process attributes at level one were divided into two or three sub-skills, mainly based on the cognitive behaviors in solving the items. With respect to the classification, the level 2 process attributes embraced the information of both the content and process attributes at level one. Thus, more detailed information underlying the item difficulty and the students' performance could be generated with the level 2 attributes. Meanwhile, because models with the level 2 attributes were more complex, more attributes' APCs did not display logical order in these models.

Results generated from the level 1 attributes and the level 2 attributes were consistent and close to each other. Analyses of 28 models for Booklets 2 and 3 found that the $R^2$ values explained by the level 2 attributes were higher than those by the level 1 attributes; however, their adjusted $R^2$ values were equal to or close to each other, except for three QM4 models, which suggested that the same variance was explained by the two levels of attributes after penalizing the $R^2$ values for adding the extra variables.

The LSDs of the models with the same type of attributes, but at the different levels, were compared (QM1 *vs* QM6, QM1 *vs* QM4, QM3 *vs* QM5, QM3-2 *vs* QM5-2). For the corresponding models with the same type of attributes, their LSDs were close to each other. But, overall, the models with the level 2 attributes had relatively smaller LSDs than those with the level 1 attributes.

With respect to the APC, the pattern of the level 1 attributes' probabilities would affect the pattern of the corresponding level 2 attributes' probabilities. For example, for Booklet 2, the probability of "a4-data and chance" was a constant of 100% across ability levels in the seven models; the probabilities of two sub-attributes of a4 (b7-data and b8-chance) were also 100% across ability levels. For Booklet 3, similar issues were also found for "a2-algebra" and its sub-attribute "b4-algebraic expressions and equations," and "a6-applying" and its sub-attribute "appl_a1/a2/a3."

Research question four: *Are the attributes of the two levels appropriate for recovering the students' mathematics achievement?* The LSDM is based on the assumption that successful response on an item requires mastery of all underlying knowledge and cognitive processes. As a result, the correct item response probability, ideally, would be equal to the product of the probabilities of correct performance on all relative attributes. Thus, students' performance on each item would be based on their performance on the required attributes. The APCs reflect the performance on attributes of students at the different ability levels. The APCs are expected to show a logical pattern in terms of monotonicity, relative difficulty, and discrimination.

Most APCs of QM1 ~ QM8 exhibited a reasonable pattern. Thus, most proposed attributes of both levels could be used to recover students' performance. In the following models, almost all APCs were acceptable: QM1 (contL1) for Booklet 3 and five models with the new process attributes—QM4 (cogL2), QM5 (congL2+IT), QM5-2 (cogL2+IT+b2), QM9 (cogL1), and QM10 (cogL1+IT) for both booklets.

However, still some attributes' probabilities showed unreasonable patterns, being a constant of 100% across ability levels or being a unexpected curve, such as "a2-algebra," "a4-data and chance," "b7-data," "b8-chance," "a6-applying," "appl_a1/a2/a3," "reas_a1/a2," "d7-judging_operation rule," and "IT2-complexity." The content attributes a2, a4, b7, and b8 were necessary knowledge topics measured by the test; however, their APCs did not display logically in many models. The problem APCs indicated that (1) the TIMSS data may not completely fit the LSDM; (2) there is still space for improving the Q-matrices. Moreover, many models indicated that the attribute IT2 was the most difficult attribute. But, at the lowest ability levels, the probability of IT2 did not rise with an increase in ability level, suggesting that students at the lowest ability levels might implement different cognitive strategies with the difficult attributes, such as guessing, and as a result, the performance on the difficult attributes would not match the ability at the lowest ability levels.

Research question five: *What attributes combined into a Q-matrix can adequately explain the TIMSS mathematics test?* According to Borsboom and Mellenbergh (2007), "A test is valid for measuring a theoretical attribute if and only if variation in the attribute causes variation in the measurement outcomes through the response process that the test elicits" (p. 93). In the present study, the Q-matrices were cross-validated by multiple regression and the LSDM. With respect to multiple regression, a higher variance in item difficulty was expected to be explained by the required attributes and a good Q-matrix. With respect to the LSDM, for an adequate Q-matrix, small LSDs and MAD were expected to produce well-recovered ICCs; also, the APCs were anticipated to be meaningful curves.

Based upon a series of analyses with the eleven Q-matrix models, the Q-matrices including the level 2 new process attributes and the IT attributes, with or without b2 (QM5 and QM5-2), were found to be reliable Q-matrices that adequately explained the TIMSS mathematics test. First, most of the variance in item difficulty was explained by these two Q-matrices ($R^2 \geq .78$ and adjusted $R^2 \geq .64$ for Booklet 2, $R^2 \geq .62$ and adjusted $R^2 \geq .40$ for Booklet 3). Second, the mean LSDs for these two models were the smallest among the eight models. Third, almost all APCs were acceptable, although the probabilities of "IT2-complexity" and "d7-judging_operationRule" at the lowest ability levels were slightly higher than those at higher ability levels. Fourth, these Q-matrices contained the information of the three types of attributes. All APCs of QM4 (cogL2) with the new process attributes were also acceptable, but QM4 did not include the IT attributes. Also, for QM4, the variances explained were relatively small ($R^2 = .60$ and adjusted $R^2 = .42$ for Booklet 2, $R^2 = .25$ and adjusted $R^2 = -.04$ for Booklet 3). So, QM4 was not as good as QM5 and QM5-2. The recovered ICC also indicated that QM5-2 was better than QM4.

Research question six: *Do the two booklets hold the same construct validity in mathematics assessment?* Reliability and validity are main concerns for a high quality measurement. The designers of the TIMSS assessments implemented a series of strict and comprehensive procedures to ensure that the tests are reliable and the test validity is supported. To maximize assessment coverage while keeping student assessment burden to a minimum, the TIMSS assessment applied a rotated block design, that is, 14 blocks of mathematics items were assembled rotationally in the 14 booklets (Mullis et al., 2005b).

To effectively measure and compare the achievement of the students taking different booklets, the same construct validity is expected to hold across the booklets.

To address the construct validity across the booklets, this study compared the results of Booklet 2 and Booklet 3 through the 14 Q-matrix models. The regression analyses showed that for all Q-matrix models, the explained variances for Booklet 3 were much less than those for Booklet 2. According to the LSDM analyses, 11 models' mean LSDs of Booklet 3 were higher than those of Booklet 2. Overall, Booklet 3 had more problem APCs than Booklet 2. With respect to the MAD for QM5-2, the ICC recovery of the 15 common items for Booklet 3 were a little worse than those for Booklet 2. Thus, all results indicated that the items of Booklet 3 were not explained by the Q-matrices as well as the items of Booklet 2. It seems that construct validity was not held constant across the two booklets. However, analysis of the shared assessment block M03 does not support this statement.

When the data of students' response to the items were calibrated using the Rasch model for each booklet, only one common item had the same difficulty across the booklets. Fourteen common items of Booklet 2 had clearly higher item difficulties than the corresponding items of Booklet 3, with a mean difference in item difficulty of .40. A regression analysis found that the Q-matrix for the 15 common items accounted for a higher variance in the item difficulties for Booklet 2 than that for Booklet 3. Moreover, the LSDM analyses employed fixed ability levels. The variance in item probabilities estimated with the Rasch model mainly came from the variance in the item difficulties. Thus, comparison of the explained variances and the ICC recovery could not demonstrate if the same construct validity was supported, or not, across the assessment booklets.

## Limitations

This current study implemented a comprehensive analysis to ensure the validation of the proposed Q-matrices. However, some limitations were found during the analyses.

The first limitation is the data used in this study. Retrofitting CDMs to existing tests is likely to generate unsatisfactory results (Close, 2012; Gierl, Alves, & Majeau, 2010; Lee & Sawaki, 2009). This study retrofitted the LSDM to the TIMSS data. The probabilities of two content attributes ("a2-algebra" and "a4-data and chance") were not found to be well displayed for many models, suggesting that the data may be not completely appropriate for the LSDM. For some attributes (e.g., reasoning), there were not enough items measuring them. In addition, the largest MAD between the ICC recovery and the estimated ICC indicated that the LSDM could not recover well the items requiring only one attribute.

The second limitation is the number of the random Q-matrices. To compare the specified Q-matrices to the random Q-matrices, only one random Q-matrix was generated for each specified Q-matrix. Results indicated that the eight specified Q-matrices were more reliable than the eight random Q-matrices. But, this did not demonstrate that the specified Q-matrices were always better than the random Q-matrices. To demonstrate that the superiority of the specified Q-matrices over the random Q-matrices did not depend on chance, it is better to generate or simulate more random Q-matrices for the comparison analyses.

The third limitation to be addressed is that the analyses did not consider guessing and slip factors. Guessing is known to be a major threat to the validity of a test score (Royal & Hedgpeth, 2013), as well as slip. For instance, for the most difficult attribute

“IT2-complexity,” the relatively high probabilities at the lowest ability levels suggested that those students at the lowest ability levels possibly applied different cognitive strategies for IT2, and so, their performance on IT2 was not displayed in a reasonable pattern. Moreover, it is possible that the items were calibrated with the Rasch model without guessing and slip parameters, which resulted in the large differences in the item difficulties of the 14 common items between Booklet 2 and for Booklet 3. This study did not consider the effects of guessing and slip and so it is recommended that future work include guessing and slip estimates.

The fourth limitation is the cross-validation between Booklet 2 and Booklet 3. One objective of this study was to investigate whether the same validity held across the two booklets. All analysis results were compared between Booklet 2 and Booklet 3. However, analysis of the 15 shared items found that the regression analysis and the LSDM did not indicate that the two booklets had the same construct validity.

**Recommendations for Further Study**

As researchers pointed out, although statistical methods are useful tools for studying CDMs and the Q-matrix, no approach is perfect and so, it is not appropriate to rely on one or two methods (Close, 2012; de la Torre, 2008; Tatsuoka, 2009). Also, it is impossible to completely reveal complex cognitive mechanisms just using several parameters. Therefore, validation of the Q-matrix needs to integrate findings from different approaches, as well as relative theories, to generate final conclusions. For further study, the Q-matrix could be tested by other statistical methods, such as the deterministic inputs, noisy ‘‘and’’ gate (DINA) model (de la Torre & Douglas, 2004; de la Torre, 2011; Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1977) and the noisy inputs, deterministic ‘‘and’’ gate (NIDA) model (Junker & Sijtsma, 2001; Maris, 1999), which include guessing and slipping parameters at the item level or the attribute level. Also, the relationship between the TIMSS mathematics items and the required attributes could be further examined with the data in different assessment cycles.

Second, it may be profitable to invite experts in cognitive psychology and cognitive science to aid in building the Q-matrix. Compared to the content attributes, the cognitive process attributes are harder to identify. Studies in cognitive science would provide useful methods to enhance the construction of the Q-matrix. For example, through investigating information processing, protocol analysis was found to be a valid approach to illustrate thought sequences via verbal reports (Ericsson, 2002; Newell & Simon, 1972). Future study of the Q-matrix would be better to invite experts in mathematics teaching, mathematics content, and cognitive psychology to analyze students’ protocols of solving mathematics items.

Third, studies could further investigate the mathematical attribute profile of each student or the students at the same ability level. One important function of DCMs is to diagnose the attribute profiles of test-takers with empirical data (DeCarlo, 2011; Hartz & Roussos, 2008; Rupp, Templin, & Henson, 2010; Yang & Embretson, 2007). An attribute profile refers to a combination of attributes mastered and attributes not mastered by a student. For the TIMSS test, students’ performance on attributes or their skills profile could be compared among the participating countries.

The LSDM chooses a set of fixed ability levels to analyze the ICC recovery, which may be different from the estimated ability levels based on the data. Thus, fourth, future study of the Q-matrix could explore the use of the estimated ability levels, which may generate more useful information with respect to the analyzed items and the test-takers. Also, the relationship between individual students' performance on each item and the attributes is expected to be investigated in future study.

In addition, the present study examined the relationship of items and required attributes in the mathematics assessment. However, the Q-matrix method can be applied to assessment in other fields, such as language testing, medical examination, and clinical diagnosis (e.g., attention-deficit/hyperactivity disorder or autism). Future study could further explore validity of the Q-matrix through analyzing assessments in different disciplines.

## Conclusion

One of the experts who assisted identifying the Q-matrix for this study has 27 years of teaching experience in mathematics for Grades 6 to 12. She pointed out that every year her students took different mathematics assessments, including TIMSS; the students' scores were often used as a key standard to measure the performance of students, teachers, and schools; however, many teachers thought the test scores were not very useful for better instruction and improving students' leaning. Study of CDA and CDMs have the potential to address this dilemma through producing richer feedback about students' performance. Constructing a reliable Q-matrix is a critical step for effectively implementing CDA and CDMs.

Moreover, development of assessments with high quality is a very challenging task, one that consumes numerous resources in time, funding, expertise, and researchers' efforts. Studying the relationship between test items and required knowledge and cognitive skills could greatly improve the efficiency in producing valuable test items. For example, Gierl and his colleagues explored automatic item generation in multiple languages based on CDM studies (Fung, Lai, & Gierl, 2013; Gierl & Lai, 2013a, 2013b; Gierl, Lai, Fung, & Latifi, 2013), which could substantially advance the construction of assessment items. This study explored the relationship between the items and the attributes based on the mathematics test of the largest international assessment TIMSS. The analyses would generate some useful findings for further enhancing this most influential assessment.

# References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716 – 723. doi:10.1109/TAC.1974.1100705

Almond, R. G., Mislevy, R. J., Williamson, D. M., & Yan, D. (2011, April). *Bayesian networks in educational assessment*. Presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Baker, F. B. (1993). Sensitivity of the linear logistic test model to misspecification of the weight matrix. *Applied Psychological Measurement*, *17*(3), 201–210.

Birenbaum, M., Tatsuoka, C., & Xin, T. (2005). Large-scale diagnostic assessment: Comparison of eighth graders' mathematics performance in the United States, Singapore and Israel. *Assessment in Education: Principles, Policy & Practice*, *12*(2), 167–181.

Birenbaum, M., Tatsuoka, C., & Yamada, T. (2004). Diagnostic assessment in TIMSS-R: Between-countries and within-country comparisons of eighth graders' mathematics performance1. *Studies in Educational Evaluation*, *30*(2), 151–173.

Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 85–118). New York: Cambridge University Press.

Chang, H.-H., Boughton, K., Wang, C., & Zhang, L. (2010). Implementing cognitive diagnosis in large scale assessment. Retrieved from http://www.ctb.com/ctb.com/control/researchArticleMainAction?articleId=25555&p=ctbResearch

Chen, C., & Zhang, J. (2012, April). *Q-matrix optimization in the cognitive diagnostic assessment*. Presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Chen, J., De la Torre, J., & Zhang, Z. (2012, April). *Relative and absolute fit evaluation in cognitive diagnosis modeling*. Presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Chen, Y.-H. (2006). *Cognitively diagnostic examination of Taiwanese mathematics achievement on TIMSS-1999*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI Number: 3220288).

Choi, H.-J., Templin, J. L., & Cohen, A. S. (2010, May). *The impact of model misspecification on estimation accuracy in diagnostic classification models*. Presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Close, C. (2012). *An exploratory technique for finding the Q-matrix for the DINA model in cognitive diagnostic assessment: Combining theory with data*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI Number: 3498498).

Close, C., Davison, M., & Davenport, E. C., Jr. (2012, April). *An exploratory technique for finding the Q-matrix in cognitive diagnostic assessment: Combining theory with data*. Presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Corter, J. E., & Tatsuoka, K. K. (2002). *Diagnostic assessments for mathematics tests grades 6-12* (Technical Report). New York: College Board.

Cui, Y., Gierl, M. J., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*(1), 19–38. doi:10.1111/j.1745-3984.2011.00158.x

Dallas, A., Furter, R., Luo, X., & Ma, J. (2012, April). *Using assessment engineering to build effective blueprints and task models for state end of grade assessments*. Presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Dallas, A., & Store, D. (2010, May). *Highly correlated: An examination of model misspecification across correlations*. Presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

De la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*(4), 343–362. doi:10.1111/j.1745-3984.2008.00069.x

De la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115–130. doi:10.3102/1076998607309474

De la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199. doi:10.1007/s11336-011-9207-7

De la Torre, J., & Chiu, C.-Y. (2009, July). *Q-matrix validation under the generalized DINA model framework*. Presented at the 74th Annual and 16th International Meeting of the Psychometric Society, Cambridge, England.

De la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353. doi:10.1007/BF02295640

De la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*(4), 595–624. doi:10.1007/s11336-008-9063-2

De la Torre, J., Henson, R. A., & Templin, J. L. (2009, April). *Skills diagnosis with latent variable models*. Presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

De la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement*, *47*(2), 227–249. doi:10.1111/j.1745-3984.2010.00110.x

De la Torre, J., & Lee, Y.-S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement*, *47*(1), 115–127. doi:10.1111/j.1745-3984.2009.00102.x

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*(1), 8 –26. doi:10.1177/0146621610377081

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics* (pp. 979–1030). Amsterdam, Netherlands: Elsevier.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–389). Hillsdale, NJ: Erlbaum.

Dimitrov, D. M. (2007). Least squares distance method of cognitive validation and analysis for binary items using their item response theory parameters. *Applied Psychological Measurement*, *31*(5), 367–387. doi:10.1177/0146621606295199

Dimitrov, D. M., & Atanasov, D. V. (2012). Conjunctive and disjunctive extensions of the least squares distance model of cognitive diagnosis. *Educational and Psychological Measurement*, *72*(1), 120–138. doi:10.1177/0013164411402324

Dogan, E. (2006). *Establishing construct validity of the mathematics subtest of the University Entrance Examination in Turkey: A Rule Space application*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI Number: 3237082).

Dogan, E., & Tatsuoka, K. K. (2008). An international comparison using a diagnostic testing model: Turkish students’ profile of mathematical skills on TIMSS-R. *Educational Studies in Mathematics, 68*(3), 263–272.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179–197.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika*, *49*(2), 175–186.

Embretson, S. E. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125–150). Hillsdale, NJ: Lawrence Erlbaum Associates.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*(3), 380–396.

Embretson, S. E. (2010). Cognitive design systems: A structural modeling approach applied to developing a spatial ability test. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 247–273). Washington, DC: American Psychological Association.

Ericsson, K. A. (2002). Protocol analysis and verbal reports on thinking. Retrieved from http://www.psy.fsu.edu/faculty/ericsson/ericsson.proto.thnk.html

Fall, E. (2009). *Applications of exploratory Q-matrix discovery procedures in diagnostic classification models*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI Number: 1488811).

Fall, E., & Templin, J. L. (2009, April). *Probabilistic Q-matrix specification in an uncertain world: An examination of cognitive diagnosis models with reading comprehension tests*. Presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Feng, Y., & Habing, B. (2012, April). *Q-matrix validation method for the reduced RUM*. Presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374.

Foy, P., Galia, J., & Li, I. (2008). Scaling the data from the TIMSS 2007 mathematics and science assessments. In J. F. Olson, M. O. Martin, & I. V. S. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 225–279). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Foy, P., & Olson, J. F. (Eds.). (2009). *TIMSS 2007 user guide for the international database: Released items, mathematics─eighth grade*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Fu, J. (2005). *A polytomous extension of the fusion model and its Bayesian parameter estimation*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI Number: 3175493).

Fu, J., & Li, Y. (2007, March). *Cognitively diagnostic psychometric models: An integrative review*. Presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Fung, K., Lai, H., & Gierl, M. J. (2013, April). *Evaluating the translations on item models in automatic item generation*. Presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Galindo-Garre, F., & Vermunt, J., K. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. *Behaviormetrika*, *33*(1), 43–59.

Gierl, M. J., Alves, C., & Majeau, R. T. (2010, May). *Using principled test design to develop and evaluate a diagnostic mathematics assessment in Grades 3 and 6*. Presented at the annual meeting of the American Educational Research Association, Denver, CO.

Gierl, M. J., & Lai, H. (2013a). Using automated processes to generate test items. *Educational Measurement: Issues and Practice*, *32*(3), 36–50.

Gierl, M. J., & Lai, H. (2013b). Evaluating the quality of medical multiple-choice items created with automated generation processes. *Medical Education*, *47*, 726-733.

Gierl, M. J., Lai, H., Fung, K., & Latifi, F. (2013, April). *Developing and evaluating methods to automatically generate items in multiple languages*. Presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Gierl, M. J., & Leighton, J. P. (2007). Linking cognitively-based models and psychometric methods. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics* (pp. 1103–1106). Amsterdam, Netherlands: Elsevier.

Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 242–274). New York: Cambridge University Press.

Gierl, M. J., Zheng, Y., & Cui, Y. (2008). Using the attribute hierarchy method to identify and interpret cognitive skills that produce group differences. *Journal of Educational Measurement*, *45*(1), 65-89.

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., & Brenwald, S. (2008). *Highlights from TIMSS 2007: Mathematics and science achievement of U.S. fourth- and eight-grade students in international context*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*(4), 301–321.

Hambleton, R. K. (1993). Principles and selected application of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). Phoenix, AZ: The Oryx Press.

Hanson, F. A. (1993). *Testing testing: Social consequences of the examined life*. Berkeley, CA: University of California Press.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI Number: 3044108).

Hartz, S. M., & Roussos, L. A. (2008). *The fusion model for skills diagnosis: Blending theory with practicality* (No. ETS RR-08-71). Retrieved from www.ets.org/Media/Research/pdf/RR-08-71.pdf

Henson, R. A., & Templin, J. L. (2009). *Implications of Q-matrix misspecification in cognitive diagnosis*. Unpublished paper.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191–210. doi:10.1007/s11336-008-9089-5

Huba, M. E., & Freed, J. E. (2000). *Learner-centered assessment on college campuses: Shifting the focus from teaching to learning*. Boston: Allyn and Bacon.

Im, S. (2007). *Statistical consequences of attribute misspecification in the Rule Space Model*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI Number: 3266604).

Im, S., & Corter, J. E. (2011). Statistical consequences of attribute misspecification in the rule space method. *Educational and Psychological Measurement*, *71*(4), 712 – 731. doi:10.1177/0013164410384855

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI Number: 3182288).

Joncas, I. V. S. (2008). TIMSS 2007 Sample Design. In J. F. Olson, M. O. Martin, & I. V. S. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 77–92). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Jones, L. V., & Thissen, D. (2007). A history and overview of psychometrics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics* (pp. 1–28). Amsterdam, Netherlands: Elsevier.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258.

Klein, M. F., Birenbaum, M., Standiford, S. N., & Tatsuoka, K. K. (1981). *Logical error analysis and construction of tests to diagnose student "bugs" in addition and subtraction of fractions* (No. CERL-RR-81-6). Urbana-Champaign, IL: University of Illinois at Urbana-Champaign. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED212496

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*(1), 59–81. doi:10.1111/j.1745-3984.2011.00160.x

Lee, Y.-W., & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly*, *6*(3), 172–189. doi:10.1080/15434300902985108

Leighton, J. P., & Gierl, M. J. (2007). Why cognitive diagnostic assessment? In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 3–18). New York: Cambridge University Press.

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. doi:10.1111/j.1745-3984.2004.tb01163.x

Linacre, J. M. (2013). Winsteps (Version 3.80.0) [Computer Software]. Beaverton, Oregon: Winsteps.com.

Linn, R. L. (1993). Current perspectives and future directions. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 1–10). Phoenix, AZ: The Oryx Press.

Linn, R. L. (2006). *Educational accountability systems* (No. CSE Technical Report 687). Boulder, CO: CRESST/University of Colorado at Boulder.

Linn, R. L. (2010). A new era of test-based educational accountability. *Measurement: Interdisciplinary Research & Perspective*, *8*(2-3), 145–149.

Liu, J., Xu, G., & Ying, Z. (2011a). Theory of self-learning Q-matrix. Retrieved from http://arxiv.org/abs/1010.6120

Liu, J., Xu, G., & Ying, Z. (2011b). Learning item-attribute relationship in Q-matrix based diagnostic classification models. Retrieved from http://arxiv.org/abs/1106.0721

Luecht, R. (2002, April). *From design to delivery: Engineering the mass production of complex performance assessments*. Presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Luecht, R. (2012, April). *Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications*. Presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Luecht, R., & Dallas, A. (2010, May). *Developing assessment engineering task models: A new way to develop test specifications*. Presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Ma, L., Çetin, E., & Green, K. E. (2009, April). *Cognitive assessment in mathematics with the least squares distance method*. Presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational and Behavioral Statistics*, *2*(2), 99–120.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212.

Masters, J. (2010). *A comparison of traditional test blueprinting and item development to assessment engineering in a licensure context*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI Number: 3403696).

MathWorks, Inc. (2005). MATLAB (Version 7.1) [Computer program]. Natick, MA: Author.

Mislevy, R. J. (2010). Some implications of expertise research for educational assessment. *Research Papers in Education*, *25*(3), 253–270. doi:10.1080/02671522.2010.498142

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. Retrieved from http://www.ets.org/Media/Research/pdf/RR-03-16.pdf

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20. doi:10.1111/j.1745-3992.2006.00075.x

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research & Perspective*, *1*(1), 3–62. doi:10.1207/S15366359MEA0101_02

Miyazaki, I. (1976). *China's examination hell: The civil service examinations of imperial China*. (C. Schirokauer, Trans.). New Haven, CT: Yale University Press.

Mullis, I. V. S., & Martin, M. O. (2008). Overview of TIMSS 2007. In J. F. Olson, M. O. Martin, & I. V. S. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 1–12). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., & Foy, P. (with O., J. F., Preuschoff, C., Erberber, E., Arora, A., & Galia, J.). (2008). *TIMSS 2007 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2005). *IEA's TIMSS 2003 international report on achievement in the mathematics cognitive domains: Findings from a* *developmental project*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. (Mullis et al., 2005a)

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. (Mullis et al., 2005b)

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

National Center for Education Statistics. (2011). Trends in International Mathematics and Science Study. Retrieved from http://nces.ed.gov/timss/index.asp

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Office. Retrieved from http://www2.ed.gov/pubs/NatAtRisk/index.html

Neo, T. Y. L. (2011). Effective cognitive diagnostic assessment with computerized adaptive testing. Retrieved from https://www.ideals.illinois.edu/handle/2142/26324

Neo, T. Y. L., & Chang, H.-H. (2012, April). *Attribute-level discrimination indices for cognitive diagnostic computerized adaptive testing*. Presented at the annual meeting of the American Educational Research Association, Vancouver, BC, Canada.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research*, *64*(4), 575–603. doi:10.2307/1170588

U.S. Department of Education. (2001). Public Law print of PL 107-110, the No Child Left Behind Act of 2001. Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/index.html

Organization for Economic Co-operation and Development. (2010). *PISA 2009 assessment framework: Key competencies in reading, mathematics and science*. Paris: OECD Publishing.

Pellegrino, J. W. (2004). The evolution of educational assessment: Considering the past and imagining the future. Retrieved from zotero://attachment/969/

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practices. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education* (Vol. 24, pp. 307–353). Retrieved from http://www.jstor.org/stable/1167273

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institut.

Rho, Y. J. (2010). *Cognitive skill diagnosis in the presence of differential strategy choice: A Bayesian approach*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI Number: 3420712).

Rosenthal, J. A. (2001). *Statistics and data interpretation for the helping professions*. Belmont, CA: Thomson Learning.

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion model skills diagnosis system. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275–318). New York: Cambridge University Press.

Royal, K. D., & Hedgpeth, M.-W. (2013). Investigating guessing strategies and their success rates on items of varying difficulty levels. *Rasch Measurement Transactions*, *27*(1), 1407–1408.

Rupp, A. A., & Templin, J. L. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*(1), 78 –96. doi:10.1177/0013164407301545

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

Schmeiser, C. B. (2007). Practical challenges to psychometrics driven by increased visibility of assessment. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics* (pp. 1117–1119). Amsterdam, Netherlands: Elsevier.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi:10.1214/aos/1176344136

Sheehan, K., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement*, *27*(3), 255–272.

Shu, Z., Henson, R. A., & Willse, J. (2010, May). *Q-matrix validation with the DINA and DINO: Implications of incorrectly adding or omitting attributes*. Presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Snow, R. E., & Lohman, D. F. (1993). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–331). Phoenix, AZ: The Oryx Press.

Suppes, P. (1969). Stimulus-response theory of finite automata. *Journal of Mathematical Psychology*, *6*(3), 327–355.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 51*(3), 337–350. doi:10.1111/1467-9876.00272

Tatsuoka, C. (2005). Corrigendum: Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 54*(2), 465–467. doi:10.1111/j.1467-9876.2005.00494.x

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345–354.

Tatsuoka, K. K. (1984). *Analysis of errors in fraction addition and subtraction problems* (No. NIE-G-81-0002). Urbana-Champaign, IL: University of Illinois at Urbana-Champaign. Retrieved from http://www.eric.ed.gov/PDFS/ED257665.pdf

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics*, *10*(1), 55–73. doi:10.2307/1164930

Tatsuoka, K. K. (1987). Validation of cognitive sensitivity for item response curves. *Journal of Educational Measurement*, *24*(3), 233–245.

Tatsuoka, K. K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–360). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tatsuoka, K. K. (1996). Use of generalized person-fit indexes, zetas for statistical pattern classification. *Applied Measurement in Education*, *9*(1), 65.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York: Taylor & Francis Group.

Tatsuoka, K. K., Birenbaum, M., & Arnold, J. (1989). On the stability of students' rules of operation for solving arithmetic problems. *Journal of Educational Measurement*, *26*(4), 351–361.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal*, *41*(4), 901.

Tatsuoka, K. K., Linn, R. L., Tatsuoka, M. M., & Yamamoto, K. (1988). Differential item functioning resulting from the use of different solution strategies. *Journal of Educational Measurement*, *25*(4), 301–319.

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement*, *20*(3), 221–30.

Tatsuoka, K. K., & Tatsuoka, M. M. (1997). Computerized cognitive diagnostic adaptive testing: Effect on remedial instruction as empirical validation. *Journal of Educational Measurement*, *34*(1), 3–20. doi:10.1111/j.1745-3984.1997.tb00504.x

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305. doi:10.1037/1082-989X.11.3.287

Templin, J. L., & Henson, R. A. (2009, April). *Quantifying reliability in diagnostic classification models*. Presented at the National Council on Measurement in Education, San Diego, CA.

Trow, M. (2006). Reflections on the transition from elite to mass to universal access: Forms and phases of higher education in modern societies since WWII. *International Handbook of Higher Education, 18*(1), 243–280.

Tu, D.-B., Cai, Y., & Dai, H.-Q. (2012). A new method of Q-matrix validation based on DINA model. *Acta Psychologica Sinica*, *44*(4), 558–568.

U.S. Department of Education. (2003). Code of Federal Regulations. Retrieved from http://edocket.access.gpo.gov/cfr_2003/julqtr/34cfr200.8.htm

Von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287–307.

Wang, C., & Gierl, M. J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement*, *48*(2), 165–187. doi:10.1111/j.1745-3984.2011.00142.x

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, *45*(4), 479–494.

Wikipedia (2013). *Coefficient of determination*. Retrieved from http://en.wikipedia.org/wiki/Coefficient_of_determination

Wikipedia. (n.d.). Formal concept analysis. Retrieved from http://en.wikipedia.org/wiki/Formal_concept_analysis

Wilbrink, B. (1997). Assessment in historical perspective. *Studies in Educational Evaluation*, *23*(1), 31–48.

Wille, R. (1984). Restructuring lattice theory: An approach based on hierarchies of concepts. In I. Rival (Ed.), *Ordered sets* (pp. 445–470). Dordrecht-Boston: Reidel.

Yang, X., & Embretson, S. E. (2007). Construct validity and cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 119–145). New York: Cambridge University Press.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145. doi:10.1177/014662168400800201

Zhang, T., & Rupp, A. A. (2009, April). *The impact of prior specification on the estimation of item and respondent parameters under Q-matrix misspecification in the DINA model*. Presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

# Appendix A

## Studies of the Item-Attribute Matrix

| Study | Main Topics | CDM or Index | Data | Number of Attributes | Number of Items | Sample Size |
|---|---|---|---|---|---|---|
| 1 | *Chen, J., de la Torre, J., & Zhang, Z. (2012, April). Relative and absolute fit evaluation in cognitive diagnosis modeling.* | | | | | |
| | Sensitivity of six model fit statistics for absolute or relative fit under different CDM settings with Q-matrix and model missepcification. | saturated G-DINA, DINA, DINO, additive CDM, linear logistic model, and reduced RUM | K.K. Tatsuoka's fraction subtraction data | 8 | 20 | 536 |
| | | DINA and additive CDM | simulated data | 5 | 15 and 30 | 500 and 1,000 |
| 2 | *Chen, C. & Zhang, J. (2012, April). Q-matrix optimization in the cognitive diagnostic assessment.* | | | | | |
| | Establish a refined iterative framework of Q-matrix optimization. | fusion model | Test of Practical Chinese | 4 | 30 | 857 |
| 3 | *Feng, Y., & Habing, B. (2012, April). Q-matrix validation method for the reduced RUM.* | | | | | |
| | The Q-matrix validation based on the difference of correct response probabilities between the examinees who mastered all attributes and those who did not. | reduced RUM | simulated data | 4 | 19 | 1,000 |
| | | | Examination for the Certificate for Proficiency in English | 3 | 28 | 2,922 |
| 4 | *Wang, W., Ding, S., Song, L., & Liu, Y. (2012, April). The application of FCA for aiding identifying attributes in cognitive diagnostic assessment.* | | | | | |
| | Assess the effectiveness of the formal concept analysis in identifying cognitive attributes. | DINA | simulated data | 6 | 6 | 10, 20 and 30 |
| | | | K.K. Tatsuoka's fraction subtraction data | 8 | 8 and 20 | 536 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 5 | *Close, C. (2012). An exploratory technique for finding the Q-matrix for the DINA model in cognitive diagnostic assessment: Combining theory with data. (also: Close & Davison, 2012)* | | | | | |
| | Evaluate a potential exploratory technique that could be used to supplement theory in finding the Q-matrix. | DINA | simulated data | 3 | 21 | 572 and 4,000 |
| | | | K.K. Tatsuoka's fraction subtraction data | 8 | 20 | 536 |
| | | | NAEP 2003 grade 8 math | (no interpretable skills) | 10, 14 and 21 | 31,588, 31,420 and 31,542 |
| 6 | *Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models.* | | | | | |
| | Effects of Q-matrix misspecification (under- and over-specification, and a combination of both) and interaction effect misspecification on item and person parameters, classification accuracy, two item-fit statistics (MAD and RMSEA), and relative model fit | log-linear diagnostic classification models | simulated data (a complex simulation study with 32 data-generation conditions) | 3 and 5 | 25 and 50 | 1,000 and 10,000 |
| 7 | *Tu, D.-B., Cai, Y., & Dai, H.-Q. (2012). A new method of Q-matrix validation based on DINA model.* | | | | | |
| | A modification method of the Q-matrix in the DINA model. | DINA | simulated data | 5 | 31 | 1,000 |
| 8 | *DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix.* | | | | | |
| | The relationship between the Q-matrix and classification of examinees. | reparameterized DINA, higher order DINA, and restricted higher order DINA; AIC and BIC | K.K. Tatsuoka's fraction subtraction data | 8 | 20 | 536 |
| | | | simulated data | 4 | 15 | |
| 9 | *Liu, J., Xu, G., & Ying, Z. (2011a). Theory of self-learning Q-matrix.* | | | | | |
| | Provide theoretical analyses on the learnability of the underlying Q-matrix. | DINA | (no data analysis) | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 10 | *Liu, J., Xu, G., & Ying, Z. (2011b). Learning item-attribute relationship in Q-matrix based diagnostic classification models.* | | | | | |
| | Mathematical framework of estimating the Q-matrix and the model parameters when the slipping and guessing parameters were unknown. | DINA and DINO | (no data analysis) | | | |
| 11 | *Shu, Z., Henson, R. A., & Willse, J. (2010, May). Q-matrix validation with the DINA and DINO: Implications of incorrectly adding or omitting attributes.* | | | | | |
| | Use Q3 to detect misspecification of skills in the Q-matrix. | DINA and DINO | simulated data | 4 | 30 | 1,000 |
| 12 | *Choi, H.-J., Templin, J. L., & Cohen, A. S. (2010). The impact of model misspecification on estimation accuracy in diagnostic classification models.* | | | | | |
| | Effect of Q-matrix misspecification and model misspecification on structural parameters and classification accuracy. | DINA, NIDA, compensatory RUM, and log-linear CDM; AIC and BIC | simulated data | 4 | 40 | 100, 200, 500, 1,000, 2,000 and 4,000 |
| | | | college chemistry test (each item with only one attribute) | 3 | 25 | 1,465 |
| 13 | *Rho, Y. J. (2010). Cognitive skill diagnosis in the presence of differential strategy choice: A Bayesian approach.* | | | | | |
| | Effect of multiple-strategy use on cognitive skill diagnosis and the cognitive diagnosis model for the presence of multiple-strategy use. | Mix-NIDA | K.K. Tatsuoka's fraction subtraction data | 7, 7 and 8 | 40 | 536 |
| 14 | *Zhang, T., & Rupp, A. A. (2009, April). The impact of prior specification on the estimation of item and respondent parameters under Q-matrix misspecification in the DINA model.* | | | | | |
| | The impact of prior specification for item parameters on the estimation of item and respondent parameters under Q-matrix misspecification. | Bayesian DINA | simulated data | 3 and 4 | (no) | 250 and 500 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| 15 | *de la Torre, J., & Chiu, C.-Y. (2009, July). Q-matrix validation under the generalized DINA model framework.* |  |  |  |  |  |
|  | A discrimination index to validate Q-matrix of the generalized DINA model. | generalized DINA: DINA, DINO, additive CDM, and the combined models of the additive CDM with DINA and with DINO | simulated data | 5 | 30 | 1,000 |
| 16 | *Fall, E. (2009). Applications of exploratory Q-matrix discovery procedures in diagnostic classification models. (also: Fall & Templin, 2009)* |  |  |  |  |  |
|  | Explore a probabilistic estimation procedure that allows for uncertainty in the construction of the Q-matrix. | DINA | reading test of the Comprehensive Adult Student Assessment System | 6 | 39 | 312 |
|  |  | DINO | reading comprehension subtest of the National Assessment of Education Progress | 2 | 24 | 249 |
|  |  | DINO | reading test of the Graduate Equivalency Degree | 2 | 20 | 312 |
| 17 | *Henson, R. A., & Templin, J. L. (2009). Implications of Q-matrix misspecification in cognitive diagnosis.* |  |  |  |  |  |
|  | Effects of Q-matrix misspecifications (under-specification, over-specification, and a combination of both) on RUM item and examinee parameters. | RUM | simulated data | 7 | 40 | 3,000 |
| 18 | *Tatsuoka, K. K. (2009). Cognitive assessment: An introduction to the rule space method. (book)* |  |  |  |  |  |
|  | Validation of attributes; validation of a Q-matrix | RSM |  |  |  |  |

| | | | | | | |
|---|---|---|---|---|---|---|
| 19 | *de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications.* | | | | | |
| | Developing an empirically based method of validating a Q-matrix for the DINA model. | DINA | simulated data | 5 | 30 | 5,000 |
| | | | K.K. Tatsuoka's fraction subtraction data | 5 | 15 | 2,144 |
| | | | NAEP 2003 grade 8 math | 9 | 90 | 3,823 |
| 20 | *Rupp, A. A., & Templin, J. L. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model.* | | | | | |
| | Effect of Q-matrix misspecification (underfiting, overfitting, balanced misfit, and incorrect dependency relationships between attributes) on parameter estimates and misclassification rates. | DINA | simulated data | 4 | 15 | 10,000 |
| 21 | *Dimitrov, D. M. (2007). Least squares distance method of cognitive validation and analysis for binary items using their item response theory parameters.* | | | | | |
| | Use the LSDM to validate the cognitive structures for a test. | least squares distance method (LSDM) | a mathematics test | 8 | 29 | 287 |
| | | | a reading comprehension test | 5 | 10 | 234 |
| 22 | *Im, S. (2007). Statistical consequences of attribute misspecification in the Rule Space Model. (also: Im & Corter, 2011)* | | | | | |
| | Statistical consequences of attribute misspecification (exclusion of an essential attribute, inclusion of a superfluous attribute, and order relations between attributes) on classification consistence and examinees' attribute mastery probability. | RSM | simulated data (The Q-matrix is from K.K. Tatsuoka's fraction subtraction data.) | 7 | 20 | 1,800 ~ 3,300 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 23 | *Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries.* | | | | | |
| | Compare the mathematics achievement of eighth-grade students at the attribute level across a sample of 20 countries in the TIMSS-R. | RSM | TIMSS-R 1999 | 23 | 163 | 51,435 |
| 24 | *Hartz, S. M. (2002). A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* | | | | | |
| | Evaluate the effectiveness and robustness of RUM parameter estimation and examinee classification with different Q-matriices. | RUM | simulated data | 7 | 40 | 1,500 |
| | | | 1992 ACT math section | 8 | 60 | 1,585 |
| 25 | *Baker, F. B. (1993). Sensitivity of the linear logistic test model to misspecification of the weight matrix.* | | | | | |
| | Effect of degree of misspecification in Q-matrix, sample size, and density of Q-matrix on item difficulty parameters and basic parameters for the contribution of attribute to item difficulty. | linear logistic test model | simulated data | 8 | 21 | 20, 50, 100 and 1,000 |

# Appendix B

## Examples of TIMSS Mathematics Items

**Example 1: Item M042055 (M02-04)**
There are 30 students in a class. The ratio of boys to girls in the class is 2:3. How many boy are there in the class?

(A) 6 (B) 12 (C) 18 (D) 20

*Note*: The attributes measured by this item are as follows:

a. the content attribute: a1-number and b2-ratio

b. the TIMSS cognitive process attributes: a5-knowing, a6-applying, and applying_a1

c. the new cognitive process attributes: c2-computing, d3-formulating, and d4-computing_number

**Example 2: Item M042263 (M02-08)**
Joe knows that a pen costs 1 zed more than a pencil.
His friend bought 2 pens and 3 pencils for 17 zeds.
How many zeds will Joe need to buy 1 pen and 2 pencils?
Show you work.

*Note*: The attributes measured by this item are as follows:

a. the content attribute: a1-number, a2-algebra, b1-whole number, and b4-algebraic expressions and equations/formulas functions

b. the TIMSS cognitive process attributes: a5-knowing, a6-applying, a7-reasoning (=c4), applying_a1, applying_a2, and reasoning_a2 (=d10)

c. the new cognitive process attributes: c2-computing, c3-judging, c4, d3-formulating, and d4-computing_number, d5-computing_algebra, d7-judge_operationRule, and d10

d. the IT attributes: IT2-complexity, and IT3-constructed-response

# Appendix C
# Informed Consent Form

**Research Project:** Validation of the Item-Attribute Matrix in TIMSS–Mathematics Using Multiple Regression and the LSDM

Dear ___________________:

You are invited to participate in a study that will explore validation of the item-attribute matrix of the mathematics assessment in the Trends in International Mathematics and Science Study (TIMSS). The study is conducted by Lin Ma to fulfill the dissertation requirements for a PhD in the Research Methods and Statistics Program at University of Denver. Results will be used (1) to validate the item-attribute matrix using multiple regression and the least squares distance method and (2) to evaluate the construct validity of the mathematics assessment in TIMSS 2007. Lin Ma can be reached at 720-224-5031/ lma3@du.edu. This project is supervised by Dissertation Director, Dr. Kathy Green, Research Methods and Statistics Program, Morgridge College of Education, University of Denver, Denver, CO 80208 (303-871-2490, kgreen@du.edu).

Your consent to participate is highly valued. Participation in this study will take about 4-10 hours of your time. Participation will involve specifying the 22 cognitive attributes at Level 2 and the maybe revised attributes to the 49 mathematics items. Participation in this project is strictly voluntary, and the risks associated with this project are minimal. If, however, you experience discomfort, you may discontinue participation at any time. Refusal to participate or withdrawal from participation will involve no penalty or loss of benefits to which you are otherwise entitled. However, as an expression of my appreciation for your participation, you will be rewarded with $200 cash if you complete the entire coding and also discuss your coding to agreement with three other experts.

Your responses are not anonymous. The discrepancies between your specified attribute matrices and three other persons' attribute matrices (including the researcher) will be discussed until agreement is reached.

If you have any concerns or complaints about how you were treated during the interview, please contact Paul Olk, Chair, Institutional Review Board for the Protection of Human Subjects, at 303-871-4531, or you may email du-irb@du.edu, Office of Research and Sponsored Programs or call 303-871-4050 or write to either at the University of Denver, Office of Research and Sponsored Programs, 2199 S. University Blvd., Denver, CO 80208-2121.

You may keep this page for your records. Please sign the next page if you understand and agree to the above. If you do not understand any part of the above statement, please ask the researcher any questions you have. Your participation is greatly appreciated!

I have read and understood the foregoing descriptions of the study called *Validation of the Item-Attribute Matrix in TIMSS–Mathematics Using Multiple Regression and the LSDM*. I have asked for and received a satisfactory explanation of any language that I did not fully understand. I agree to participate in this study, and I understand that I may withdraw my consent at any time. I have received a copy of this consent form.

Signature _____________________ Date _________________

## Appendix D
## The Q-Matrix with the Attributes Based on the TIMSS Assessment Framework and Item Type

| No. | Block | Block Seq | Item ID | Content—L1 (a5-Knowing) * | | | | Content—L2 | | | | | | | | Cognitive Process—L1 | | | Cognitive Process—L2 | | | | | | | | Complex Cognitive Process | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | a1 | a2 | a3 | a4 | a1-number | | a2-algebra | | a3-geometry | | a4-dataChance | | a5 | a6 | a7 | a6-Applying | | | | a7-Reasoning | | | | | | |
| | | | | | | | | b1 | b2 | b3 | b4 | b5 | b6 | b7 | b8 | | | | a1 | a2 | a3 | a4 | a1 | a2 | a3 | a4 | IT1 | IT2 | IT3 |
| 1 | M02 | 1 | M042003 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 2 | M02 | 2 | M042079 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 3 | M02 | 3 | M042018 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 | | 1 | 0 | 1 |
| 4 | M02 | 4 | M042055 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 5 | M02 | 5 | M042039 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 6 | M02 | 6 | M042199 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 7 | M02 | 07A | M042301A | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | | 1 | 0 | 1 |
| 8 | M02 | 07B | M042301B | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | | 1 | 0 | | 0 | 0 | 1 |
| 9 | M02 | 07C | M042301C | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | | 1 | 0 | | 0 | 1 | 1 |
| 10 | M02 | 8 | M042263 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | | 1 | 0 | | 1 | 1 | 1 |
| 11 | M02 | 9 | M042265 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 | 1 | | 0 | 0 | 0 |
| 12 | M02 | 10 | M042137 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 13 | M02 | 11 | M042148 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 14 | M02 | 12 | M042254 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | | 0 | 0 | | 0 | 0 | 0 |
| 15 | M02 | 13 | M042250 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 1 |
| 16 | M02 | 14 | M042220 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | | 0 | 0 | | 1 | 1 | 1 |
| 17 | M03 | 1 | M022097 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | M03 | 2 | M022101 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | M03 | 3 | M022104 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20 | M03 | 4 | M022105 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 21 | M03 | 5 | M022106 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 22 | M03 | 6 | M022108 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 23 | M03 | 7 | M022110 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 24 | M03 | 8 | M022181 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25 | M03 | 9 | M032307 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

(Continued)

(Continued)

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 26 | M03 | 10 | M032523 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 27 | M03 | 11 | M032701 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 28 | M03 | 12 | M032704 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 29 | M03 | 13 | M032525 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30 | M03 | 14 | M032579 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 31 | M03 | 15 | M032691 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 32 | M04 | 1 | M042001 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 |
| 33 | M04 | 2 | M042022 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 |
| 34 | M04 | 3 | M042082 | 1 | 1 | 0 | 0 | 1 | 0 | | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 1 | 0 | 0 |
| 35 | M04 | 4 | M042088 | 0 | 1 | 0 | 0 | 0 | 0 | | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 |
| 36 | M04 | 05A | M042304A | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 1 |
| 37 | M04 | 05B | M042304B | 1 | 1 | 0 | 0 | 1 | 1 | | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 | 0 | 1 | 1 | 1 |
| 38 | M04 | 05C | M042304C | 1 | 1 | 0 | 0 | 1 | 1 | | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 | 0 | 1 | 1 | 1 |
| 39 | M04 | 05D-1 | M042304D-1 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 | 0 | 1 | 0 | 1 |
| 40 | M04 | 05D-2 | M042304D-2 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | | 0 | 0 | 0 | 0 | 1 |
| 41 | M04 | 6 | M042267 | 1 | 1 | 0 | 0 | 1 | 0 | | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 |
| 42 | M04 | 7 | M042239 | 1 | 1 | 0 | 0 | 1 | 0 | | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 1 | 0 | 0 |
| 43 | M04 | 8 | M042238 | 1 | 1 | 1 | 0 | 1 | 0 | | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 |
| 44 | M04 | 9 | M042279 | 0 | 0 | 1 | 0 | 0 | 0 | | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | | 1 | 0 | 0 | 0 | 0 |
| 45 | M04 | 10 | M042036 | 1 | 0 | 1 | 0 | 1 | 0 | | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | | 1 | 0 | 1 | 0 | 0 |
| 46 | M04 | 11 | M042130 | 1 | 0 | 1 | 0 | 1 | 0 | | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | | 0 | 0 | 1 | 1 | 1 |
| 47 | M04 | 12A | M042303A | 1 | 0 | 0 | 1 | 1 | 1 | | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | | 0 | 0 | 1 | 1 | 1 |
| 48 | M04 | 12B | M042303B | 1 | 0 | 1 | 1 | 1 | 0 | | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | | 0 | 1 | 1 | 1 | 1 |
| 49 | M04 | 13 | M042222 | 1 | 0 | 0 | 1 | 1 | 1 | | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 1 | 0 |
| number of related items-booklet2 | | | | 28 | 6 | 9 | 6 | 28 | 14 | 2 | 5 | 7 | 5 | 4 | 2 | 31 | 20 | 4 | 14 | 3 | 4 | 3 | 0 | 3 | 1 | 0 | 4 | 5 | 11 |
| number of related items-booklet3 | | | | 29 | 9 | 9 | 6 | 29 | 13 | 0 | 9 | 6 | 7 | 3 | 5 | 33 | 22 | 4 | 17 | 2 | 6 | 3 | 2 | 0 | 2 | 1 | 9 | 8 | 12 |

* The four content attributes at level one (a1, a2, a3, and a4) are the same as the four knowing attributes at level two (know_a1, know_a2, know_a3, and know_a4).

# Appendix E
## The Q-Matrix with the New Cognitive Process Attributes

| No. | Block | Block Seq | Item ID | Cognitive Process—L1 | | | | Cognitive Process—L2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | c1 | c2 | c3 | c4 | c1-Identifying | | c2-Computing | | | c3-Judging | | | c4-Reasoning | | |
| | | | | | | | | d1 | d2 | d3 | d4 | d5 | d6 | d7 | d8 | d9 | d10 | d11 |
| 1 | M02 | 1 | M042003 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 |
| 2 | M02 | 2 | M042079 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 | 0 |
| 3 | M02 | 3 | M042018 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | | 0 | 0 |
| 4 | M02 | 4 | M042055 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | | 0 | 0 |
| 5 | M02 | 5 | M042039 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | | 0 | 0 |
| 6 | M02 | 6 | M042199 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 |
| 7 | M02 | 07A | M042301A | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 |
| 8 | M02 | 07B | M042301B | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | | 1 | 0 |
| 9 | M02 | 07C | M042301C | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | | 1 | 0 |
| 10 | M02 | 8 | M042263 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | | 1 | 0 |
| 11 | M02 | 9 | M042265 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 1 |
| 12 | M02 | 10 | M042137 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | | 0 | 0 |
| 13 | M02 | 11 | M042148 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 |
| 14 | M02 | 12 | M042254 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | | 0 | 0 |
| 15 | M02 | 13 | M042250 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | | 0 | 0 |
| 16 | M02 | 14 | M042220 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | | 0 | 0 |
| 17 | M03 | 1 | M022097 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 18 | M03 | 2 | M022101 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | M03 | 3 | M022104 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20 | M03 | 4 | M022105 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 21 | M03 | 5 | M022106 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 22 | M03 | 6 | M022108 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 23 | M03 | 7 | M022110 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 24 | M03 | 8 | M022181 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 25 | M03 | 9 | M032307 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

(Continued)

(Continued)

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 26 | M03 | 10 | M032523 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 27 | M03 | 11 | M032701 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 28 | M03 | 12 | M032704 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 29 | M03 | 13 | M032525 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 30 | M03 | 14 | M032579 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 31 | M03 | 15 | M032691 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 32 | M04 | 1 | M042001 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 |
| 33 | M04 | 2 | M042022 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 |
| 34 | M04 | 3 | M042082 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | | 0 |
| 35 | M04 | 4 | M042088 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | | 0 |
| 36 | M04 | 05A | M042304A | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 |
| 37 | M04 | 05B | M042304B | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | | 0 |
| 38 | M04 | 05C | M042304C | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | | 0 |
| 39 | M04 | 05D-1 | M042304D-1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 |
| 40 | M04 | 05D-2 | M042304D-2 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | | 0 |
| 41 | M04 | 6 | M042267 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | | 0 |
| 42 | M04 | 7 | M042239 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | | 0 |
| 43 | M04 | 8 | M042238 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | | 0 |
| 44 | M04 | 9 | M042279 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 1 |
| 45 | M04 | 10 | M042036 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | | 1 |
| 46 | M04 | 11 | M042130 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | | 0 |
| 47 | M04 | 12A | M042303A | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 |
| 48 | M04 | 12B | M042303B | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | | 0 |
| 49 | M04 | 13 | M042222 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 |
| number of related items-booklet2 | | | | 11 | 24 | 18 | 4 | 4 | 8 | 17 | 22 | 5 | 11 | 3 | 5 | 0 | 3 | 1 |
| number of related items-booklet3 | | | | 11 | 25 | 24 | 4 | 7 | 7 | 11 | 25 | 9 | 15 | 3 | 8 | 2 | 0 | 2 |

# Appendix F

## The Random Q-Matrix with the Attributes Based on the TIMSS Assessment Framework and Item Type (Booklet 2)

| No. | Block | Block Seq | Item ID | Content—L1 (a5-Knowing) | | | | Content—L2 | | | | | | | | Cognitive Process—L1 | | | Cognitive Process—L2 | | | | | | | | Complex Cognitive Process | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | a1 | a2 | a3 | a4 | a1-number | | a2-algebra | | a3-geometry | | a4-dataChance | | a5 | a6 | a7 | a6-Applying | | | | a7-Reasoning | | | | | | |
| | | | | | | | | b1 | b2 | b3 | b4 | b5 | b6 | b7 | b8 | | | | a1 | a2 | a3 | a4 | a1 | a2 | a3 | a4 | IT1 | IT2 | IT3 |
| 1 | M02 | 1 | M042003 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | | 0 | 0 | | 0 | 0 | 1 |
| 2 | M02 | 2 | M042079 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | | 0 | 1 | 1 |
| 3 | M02 | 3 | M042018 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 4 | M02 | 4 | M042055 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | | 0 | 1 | | 0 | 1 | 0 |
| 5 | M02 | 5 | M042039 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 6 | M02 | 6 | M042199 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 | | 0 | 1 | 0 |
| 7 | M02 | 07A | M042301A | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | | 1 | 0 | | 0 | 0 | 0 |
| 8 | M02 | 07B | M042301B | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 1 |
| 9 | M02 | 07C | M042301C | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | | 0 | 0 | | 0 | 0 | 0 |
| 10 | M02 | 8 | M042263 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 11 | M02 | 9 | M042265 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | | 0 | 0 | | 1 | 0 | 0 |
| 12 | M02 | 10 | M042137 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 13 | M02 | 11 | M042148 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | | 0 | 0 | | 0 | 0 | 0 |
| 14 | M02 | 12 | M042254 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 1 |
| 15 | M02 | 13 | M042250 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 16 | M02 | 14 | M042220 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 17 | M03 | 1 | M022097 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 | 1 | | 0 | 0 | 1 |
| 18 | M03 | 2 | M022101 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | | 1 | 0 | | 1 | 1 | 0 |
| 19 | M03 | 3 | M022104 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | | 0 | 0 | | 1 | 1 | 1 |
| 20 | M03 | 4 | M022105 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 21 | M03 | 5 | M022106 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 22 | M03 | 6 | M022108 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 | 0 | | 1 | 0 | 1 |
| 23 | M03 | 7 | M022110 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 24 | M03 | 8 | M022181 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 25 | M03 | 9 | M032307 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 26 | M03 | 10 | M032523 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 27 | M03 | 11 | M032701 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 1 |
| 28 | M03 | 12 | M032704 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | | 0 | 0 | 0 |
| 29 | M03 | 13 | M032525 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | | 1 | 0 | 0 |
| 30 | M03 | 14 | M032579 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | | 0 | 0 | | 0 | 1 | 0 |
| 31 | M03 | 15 | M032691 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | | 0 | 1 | 0 |
| number of related items-booklet2 | | | | 28 | 7 | 6 | 8 | 28 | 7 | 6 | 3 | 7 | 7 | 5 | 4 | 31 | 20 | 4 | 10 | 1 | 7 | 6 | 0 | 2 | 2 | 0 | 5 | 7 | 8 |

# Appendix G

## The Random Q-Matrix with the Attributes Based on the TIMSS Assessment Framework and Item Type (Booklet 3)

| No. | Block | Block Seq | Item ID | Content—L1 (a5-Knowing) | | | | Content—L2 | | | | | | | | Cognitive Process—L1 | | | Cognitive Process—L2 | | | | | | | | Complex Cognitive Process | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | a1 | a2 | a3 | a4 | a1-number | | a2-algebra | | a3-geometry | | a4-dataChance | | a5 | a6 | a7 | a6-Applying | | | | a7-Reasoning | | | | | | |
| | | | | | | | | b1 | b2 | b3 | b4 | b5 | b6 | b7 | b8 | | | | a1 | a2 | a3 | a4 | a1 | a2 | a3 | a4 | IT1 | IT2 | IT3 |
| 17 | M03 | 1 | M022097 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 | 0 | 1 | 1 | 0 |
| 18 | M03 | 2 | M022101 | 1 | 1 | 1 | 1 | 1 | 0 | | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 |
| 19 | M03 | 3 | M022104 | 0 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | | 0 | 0 | 1 | 0 | 0 |
| 20 | M03 | 4 | M022105 | 1 | 0 | 0 | 0 | 1 | 0 | | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 |
| 21 | M03 | 5 | M022106 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | | 1 | 0 | 0 | 0 | 0 |
| 22 | M03 | 6 | M022108 | 0 | 0 | 0 | 0 | 0 | 1 | | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 1 | 1 | 1 |
| 23 | M03 | 7 | M022110 | 1 | 1 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | | 1 | 0 | 0 | 0 | 1 |
| 24 | M03 | 8 | M022181 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 |
| 25 | M03 | 9 | M032307 | 1 | 0 | 1 | 0 | 1 | 1 | | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 1 |
| 26 | M03 | 10 | M032523 | 1 | 0 | 0 | 1 | 1 | 0 | | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | | 0 | 0 | 1 | 0 | 1 |
| 27 | M03 | 11 | M032701 | 1 | 0 | 0 | 0 | 1 | 0 | | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 |
| 28 | M03 | 12 | M032704 | 0 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 |
| 29 | M03 | 13 | M032525 | 1 | 0 | 0 | 0 | 1 | 1 | | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 1 |
| 30 | M03 | 14 | M032579 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 |
| 31 | M03 | 15 | M032691 | 1 | 0 | 0 | 1 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 |
| 32 | M04 | 1 | M042001 | 1 | 1 | 0 | 0 | 1 | 0 | | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | | 0 | 0 | 1 | 1 | 0 |
| 33 | M04 | 2 | M042022 | 1 | 1 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | | 0 | 0 | 0 | 1 | 0 |
| 34 | M04 | 3 | M042082 | 1 | 1 | 0 | 0 | 0 | 0 | | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 1 | 1 |
| 35 | M04 | 4 | M042088 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | | 0 | 0 | 0 | 0 | 1 |
| 36 | M04 | 05A | M042304A | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 1 | 1 | 0 | 0 |
| 37 | M04 | 05B | M042304B | 0 | 0 | 1 | 0 | 1 | 0 | | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 1 | 1 | 0 |
| 38 | M04 | 05C | M042304C | 1 | 1 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 1 |
| 39 | M04 | 05D-1 | M042304D-1 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 1 | 1 |
| 40 | M04 | 05D-2 | M042304D-2 | 1 | 0 | 1 | 0 | 1 | 0 | | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | | 0 | 0 | 1 | 0 | 0 |
| 41 | M04 | 6 | M042267 | 1 | 1 | 0 | 0 | 1 | 1 | | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 1 |
| 42 | M04 | 7 | M042239 | 1 | 1 | 0 | 0 | 1 | 1 | | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | | 0 | 0 | 0 | 0 | 0 |
| 43 | M04 | 8 | M042238 | 1 | 1 | 1 | 0 | 1 | 0 | | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 1 | 0 |
| 44 | M04 | 9 | M042279 | 1 | 0 | 0 | 1 | 1 | 0 | | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 |
| 45 | M04 | 10 | M042036 | 1 | 0 | 0 | 0 | 1 | 0 | | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 |
| 46 | M04 | 11 | M042130 | 1 | 0 | 1 | 1 | 1 | 0 | | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 1 | 0 |
| 47 | M04 | 12A | M042303A | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 |
| 48 | M04 | 12B | M042303B | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 1 | 0 |
| 49 | M04 | 13 | M042222 | 1 | 1 | 1 | 0 | 1 | 1 | | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 1 | 0 |
| number of related items-booklet3 | | | | 29 | 10 | 9 | 5 | 29 | 6 | 0 | 8 | 8 | 9 | 5 | 7 | 33 | 22 | 4 | 8 | 4 | 7 | 9 | 2 | 0 | 2 | 1 | 8 | 11 | 10 |

# Appendix H
# The Random Q-Matrix with the New Cognitive Process Attributes (Booklet 2)

| No. | Block | Block Seq | Item ID | Cognitive Process—L1 | | | | Cognitive Process—L2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | c1 | c2 | c3 | c4 | c1-Identifying | | c2-Computing | | | c3-Judging | | | c4-Reasoning | | |
| | | | | | | | | d1 | d2 | d3 | d4 | d5 | d6 | d7 | d8 | d9 | d10 | d11 |
| 1 | M02 | 1 | M042003 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | | 0 | 0 |
| 2 | M02 | 2 | M042079 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | | 0 | 0 |
| 3 | M02 | 3 | M042018 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | | 0 | 0 |
| 4 | M02 | 4 | M042055 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | | 0 | 0 |
| 5 | M02 | 5 | M042039 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | | 1 | 0 |
| 6 | M02 | 6 | M042199 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 |
| 7 | M02 | 07A | M042301A | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 |
| 8 | M02 | 07B | M042301B | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | | 0 | 0 |
| 9 | M02 | 07C | M042301C | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 |
| 10 | M02 | 8 | M042263 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | | 0 | 0 |
| 11 | M02 | 9 | M042265 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 |
| 12 | M02 | 10 | M042137 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 |
| 13 | M02 | 11 | M042148 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | | 0 | 1 |
| 14 | M02 | 12 | M042254 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 |
| 15 | M02 | 13 | M042250 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 | 0 |
| 16 | M02 | 14 | M042220 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | | 0 | 1 |
| 17 | M03 | 1 | M022097 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | | 0 | 0 |
| 18 | M03 | 2 | M022101 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 |
| 19 | M03 | 3 | M022104 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 |
| 20 | M03 | 4 | M022105 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 | 1 |
| 21 | M03 | 5 | M022106 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | | 0 | 0 |
| 22 | M03 | 6 | M022108 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | | 0 | 0 |
| 23 | M03 | 7 | M022110 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | | 0 | 0 |
| 24 | M03 | 8 | M022181 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 |
| 25 | M03 | 9 | M032307 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 |
| 26 | M03 | 10 | M032523 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | | 0 | 0 |
| 27 | M03 | 11 | M032701 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | | 0 | 0 |
| 28 | M03 | 12 | M032704 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | | 0 | 0 |
| 29 | M03 | 13 | M032525 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | | 0 | 0 |
| 30 | M03 | 14 | M032579 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | | 0 | 0 |
| 31 | M03 | 15 | M032691 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | | 0 | 0 |
| number of related items-booklet2 | | | | 11 | 24 | 18 | 4 | 8 | 4 | 9 | 12 | 17 | 7 | 5 | 7 | 0 | 1 | 3 |

# Appendix I
## The Random Q-Matrix with the New Cognitive Process Attributes (Booklet 3)

| No. | Block | Block Seq | Item ID | Cognitive Process—L1 | | | | Cognitive Process—L2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | c1 | c2 | c3 | c4 | c1-Identifying | | c2-Computing | | | c3-Judging | | | c4-Reasoning | | |
| | | | | | | | | d1 | d2 | d3 | d4 | d5 | d6 | d7 | d8 | d9 | d10 | d11 |
| 17 | M03 | 1 | M022097 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | | 0 |
| 18 | M03 | 2 | M022101 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | | 0 |
| 19 | M03 | 3 | M022104 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 |
| 20 | M03 | 4 | M022105 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | | 0 |
| 21 | M03 | 5 | M022106 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | | 0 |
| 22 | M03 | 6 | M022108 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 |
| 23 | M03 | 7 | M022110 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | | 0 |
| 24 | M03 | 8 | M022181 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | | 0 |
| 25 | M03 | 9 | M032307 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 |
| 26 | M03 | 10 | M032523 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 |
| 27 | M03 | 11 | M032701 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 |
| 28 | M03 | 12 | M032704 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | | 1 |
| 29 | M03 | 13 | M032525 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | | 0 |
| 30 | M03 | 14 | M032579 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 |
| 31 | M03 | 15 | M032691 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 |
| 32 | M04 | 1 | M042001 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | | 0 |
| 33 | M04 | 2 | M042022 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 |
| 34 | M04 | 3 | M042082 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | | 0 |
| 35 | M04 | 4 | M042088 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 |
| 36 | M04 | 05A | M042304A | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | | 0 |
| 37 | M04 | 05B | M042304B | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 |
| 38 | M04 | 05C | M042304C | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | | 0 |
| 39 | M04 | 05D-1 | M042304D-1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | | 0 |
| 40 | M04 | 05D-2 | M042304D-2 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 |
| 41 | M04 | 6 | M042267 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | | 1 |
| 42 | M04 | 7 | M042239 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 |
| 43 | M04 | 8 | M042238 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | | 0 |
| 44 | M04 | 9 | M042279 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | | 0 |
| 45 | M04 | 10 | M042036 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | | 0 |
| 46 | M04 | 11 | M042130 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | | 0 |
| 47 | M04 | 12A | M042303A | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 |
| 48 | M04 | 12B | M042303B | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 |
| 49 | M04 | 13 | M042222 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | | 0 |
| number of related items-booklet3 | | | | 11 | 25 | 24 | 4 | 7 | 7 | 12 | 19 | 14 | 14 | 4 | 8 | 2 | 0 | 2 |

# Appendix J
## The Revised Q-Matrix with the Attributes Based on the TIMSS Assessment Framework and Item Type

| No. | Block | Block Seq | Item ID | Content—L1 (a5-Knowing) | | | | Content—L2 | | | | | | | | Cognitive Process—L1 | | | Cognitive Process—L2 | | | | | | Complex Cognitive Process | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | a1 | a2 | a3 | a4 | a1-number | | a2-algebra | | a3-geometry | | a4-dataChance | | a5 | a6 | a7 | a6-Applying | | | a7-Reasoning | | | | |
| | | | | | | | | b1 | b2 | b3 | b4 | b5 | b6 | b7 | b8 | | | | a1 | a2 | a3 | a1 | a2 | a3 | IT2 | IT3 |
| 1 | M02 | 1 | M042003 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 |
| 2 | M02 | 2 | M042079 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 |
| 3 | M02 | 3 | M042018 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | | 0 | 0 | 0 | 1 |
| 4 | M02 | 4 | M042055 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | | 0 | 0 | 0 | 0 |
| 5 | M02 | 5 | M042039 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | | 0 | 0 | 0 | 0 |
| 6 | M02 | 6 | M042199 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 |
| 7 | M02 | 07A | M042301A | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 1 |
| 8 | M02 | 07B | M042301B | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | | 1 | 0 | 0 | 1 |
| 9 | M02 | 07C | M042301C | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | | 1 | 0 | 1 | 1 |
| 10 | M02 | 8 | M042263 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | | 1 | 0 | 1 | 1 |
| 11 | M02 | 9 | M042265 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 1 | 0 | 0 |
| 12 | M02 | 10 | M042137 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | | 0 | 0 | 0 | 0 |
| 13 | M02 | 11 | M042148 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 |
| 14 | M02 | 12 | M042254 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | | 0 | 0 | 0 | 0 |
| 15 | M02 | 13 | M042250 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | | 0 | 0 | 0 | 1 |
| 16 | M02 | 14 | M042220 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | | 0 | 0 | 1 | 1 |
| 17 | M03 | 1 | M022097 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | M03 | 2 | M022101 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | M03 | 3 | M022104 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20 | M03 | 4 | M022105 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 21 | M03 | 5 | M022106 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 22 | M03 | 6 | M022108 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 23 | M03 | 7 | M022110 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 24 | M03 | 8 | M022181 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25 | M03 | 9 | M032307 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

(Continued)

(Continued)

| | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 26 | M03 | 10 | M032523 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 27 | M03 | 11 | M032701 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 28 | M03 | 12 | M032704 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 29 | M03 | 13 | M032525 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30 | M03 | 14 | M032579 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 31 | M03 | 15 | M032691 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 32 | M04 | 1 | M042001 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 |
| 33 | M04 | 2 | M042022 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 |
| 34 | M04 | 3 | M042082 | 1 | 1 | 0 | 0 | 1 | 0 | | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 |
| 35 | M04 | 4 | M042088 | 0 | 1 | 0 | 0 | 0 | 0 | | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 |
| 36 | M04 | 05A | M042304A | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 | 1 |
| 37 | M04 | 05B | M042304B | 1 | 1 | 0 | 0 | 0 | 1 | | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 1 | 1 |
| 38 | M04 | 05C | M042304C | 1 | 0 | 0 | 0 | 0 | 1 | | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 1 | 1 |
| 39 | M04 | 05D-1 | M042304D-1 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 | 1 |
| 40 | M04 | 05D-2 | M042304D-2 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | | 0 | 0 | 1 |
| 41 | M04 | 6 | M042267 | 1 | 1 | 0 | 0 | 1 | 0 | | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | | 0 | 0 | 0 |
| 42 | M04 | 7 | M042239 | 1 | 1 | 0 | 0 | 1 | 0 | | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 |
| 43 | M04 | 8 | M042238 | 1 | 1 | 1 | 0 | 1 | 0 | | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | | 0 | 0 | 0 |
| 44 | M04 | 9 | M042279 | 0 | 0 | 1 | 0 | 0 | 0 | | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | | 1 | 0 | 0 |
| 45 | M04 | 10 | M042036 | 1 | 1 | 1 | 0 | 1 | 0 | | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | | 1 | 0 | 0 |
| 46 | M04 | 11 | M042130 | 1 | 1 | 1 | 0 | 1 | 0 | | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | | 0 | 0 | 1 |
| 47 | M04 | 12A | M042303A | 1 | 0 | 0 | 1 | 0 | 1 | | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 | 1 |
| 48 | M04 | 12B | M042303B | 1 | 0 | 1 | 1 | 1 | 0 | | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | | 0 | 1 | 1 |
| 49 | M04 | 13 | M042222 | 1 | 0 | 0 | 1 | 0 | 1 | | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 | 0 |
| number of related items-booklet2 | | | | 28 | 7 | 9 | 6 | 14 | 14 | 2 | 5 | 7 | 5 | 5 | 1 | 31 | 21 | 4 | 17 | 3 | 4 | 0 | 3 | 1 | 5 | 11 |
| number of related items-booklet3 | | | | 29 | 11 | 9 | 6 | 16 | 13 | 0 | 9 | 6 | 7 | 5 | 3 | 33 | 23 | 4 | 18 | 2 | 6 | 2 | 0 | 2 | 5 | 12 |

*Note*. The changed elements and the re-specified attribute b1 were highlighted.

# Appendix K
## The Revised Q-Matrix with the New Cognitive Process Attributes

| No. | Block | Block Seq | Item ID | Cognitive Process—L1 | | | | Cognitive Process—L2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | c1 | c2 | c3 | c4 | c1-Identifying | | c2-Computing | | | c3-Judging | | c4-Reasoning | | |
| | | | | | | | | d1 | d2 | d3 | d4 | d5 | d7 | d8 | d9 | d10 | d11 |
| 1 | M02 | 1 | M042003 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 |
| 2 | M02 | 2 | M042079 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 |
| 3 | M02 | 3 | M042018 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | | 0 | 0 |
| 4 | M02 | 4 | M042055 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | | 0 | 0 |
| 5 | M02 | 5 | M042039 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | | 0 | 0 |
| 6 | M02 | 6 | M042199 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | | 0 | 0 |
| 7 | M02 | 07A | M042301A | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 |
| 8 | M02 | 07B | M042301B | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | | 1 | 0 |
| 9 | M02 | 07C | M042301C | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | | 1 | 0 |
| 10 | M02 | 8 | M042263 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | | 1 | 0 |
| 11 | M02 | 9 | M042265 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 1 |
| 12 | M02 | 10 | M042137 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | | 0 | 0 |
| 13 | M02 | 11 | M042148 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 |
| 14 | M02 | 12 | M042254 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 |
| 15 | M02 | 13 | M042250 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | | 0 | 0 |
| 16 | M02 | 14 | M042220 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | | 0 | 0 |
| 17 | M03 | 1 | M022097 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | M03 | 2 | M022101 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | M03 | 3 | M022104 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20 | M03 | 4 | M022105 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 21 | M03 | 5 | M022106 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 22 | M03 | 6 | M022108 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 23 | M03 | 7 | M022110 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 24 | M03 | 8 | M022181 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25 | M03 | 9 | M032307 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

(Continued)

(Continued)

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 26 | M03 | 10 | M032523 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 27 | M03 | 11 | M032701 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 28 | M03 | 12 | M032704 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 29 | M03 | 13 | M032525 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 30 | M03 | 14 | M032579 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 31 | M03 | 15 | M032691 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 32 | M04 | 1 | M042001 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 |
| 33 | M04 | 2 | M042022 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | | 0 |
| 34 | M04 | 3 | M042082 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | | 0 |
| 35 | M04 | 4 | M042088 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 |
| 36 | M04 | 05A | M042304A | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | | 0 |
| 37 | M04 | 05B | M042304B | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | | 0 |
| 38 | M04 | 05C | M042304C | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | | 0 |
| 39 | M04 | 05D-1 | M042304D-1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | | 0 |
| 40 | M04 | 05D-2 | M042304D-2 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | | 0 |
| 41 | M04 | 6 | M042267 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | | 0 |
| 42 | M04 | 7 | M042239 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | | 0 |
| 43 | M04 | 8 | M042238 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | | 0 |
| 44 | M04 | 9 | M042279 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 1 |
| 45 | M04 | 10 | M042036 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | | 1 |
| 46 | M04 | 11 | M042130 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | | 0 |
| 47 | M04 | 12A | M042303A | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 |
| 48 | M04 | 12B | M042303B | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | | 0 |
| 49 | M04 | 13 | M042222 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 |
| number of related items-booklet2 | | | | 11 | 24 | 8 | 4 | 4 | 8 | 16 | 22 | 6 | 3 | 5 | 0 | 3 | 1 |
| number of related items-booklet3 | | | | 11 | 25 | 12 | 4 | 7 | 7 | 10 | 25 | 11 | 4 | 8 | 2 | 0 | 2 |

*Note*. The changed elements and the re-specified attribute c3 were highlighted.

# Appendix L

## Correlation Coefficients among the Item Difficulties and All Attributes (1) ~ (8)

(1) Correlation Coefficient among the Item Difficulty and the Attributes for QM1 ~ QM5-2 of Booklet 2 (with TIMSS Process Attributes)

| | Variables | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ItemDifficulty | 1.00 | | | | | | | | | | | | | | | |
| 2 | b1integer | -.03 | 1.00 | | | | | | | | | | | | | | |
| 3 | b2decimal | -.09 | -.82$^{**}$ | 1.00 | | | | | | | | | | | | | |
| 4 | know_a1 | -.21 | .30 | .30 | 1.00 | | | | | | | | | | | | |
| 5 | know_a2 | .22 | .44$^{*}$ | -.34 | .18 | 1.00 | | | | | | | | | | | |
| 6 | know_a3 | .14 | .28 | -.58$^{**}$ | -.51$^{**}$ | -.01 | 1.00 | | | | | | | | | | |
| 7 | know_a4 | -.36$^{*}$ | -.12 | .21 | .16 | -.26 | -.31 | 1.00 | | | | | | | | | |
| 8 | appl_a1 | .23 | -.22 | .43$^{*}$ | .36$^{*}$ | .03 | -.56$^{**}$ | .44$^{*}$ | 1.00 | | | | | | | | |
| 9 | appl_a2 | .46$^{**}$ | .36$^{*}$ | -.30 | .11 | .61$^{**}$ | .03 | -.16 | .30 | 1.00 | | | | | | | |
| 10 | appl_a3 | .09 | .23 | -.35 | -.20 | .02 | .60$^{**}$ | -.19 | -.42$^{*}$ | -.13 | 1.00 | | | | | | |
| 11 | reas_a2 | .46$^{**}$ | .36$^{*}$ | -.30 | .11 | .61$^{**}$ | .03 | -.16 | .30 | 1.00$^{**}$ | -.13 | 1.00 | | | | | |
| 12 | reas_a3 | .06 | -.17 | -.17 | -.56$^{**}$ | -.10 | .29 | -.09 | -.20 | -.06 | -.07 | -.06 | 1.00 | | | | |
| 13 | a6applying | .31 | -.07 | .21 | .24 | .04 | -.17 | .34 | .76$^{**}$ | .23 | .27 | .23 | -.26 | 1.00 | | | |
| 14 | a7reasoning | .44$^{*}$ | .23 | -.35 | -.20 | .48$^{**}$ | .18 | -.19 | .16 | .85$^{**}$ | -.15 | .85$^{**}$ | .47$^{**}$ | .06 | 1.00 | | |
| 15 | IT2complexity | .62$^{**}$ | -.05 | .13 | .14 | .18 | -.28 | .01 | .40$^{*}$ | .45$^{*}$ | -.17 | .45$^{*}$ | -.08 | .30 | .35 | 1.00 | |
| 16 | IT3constructed | .54$^{**}$ | .14 | -.13 | .01 | .08 | -.03 | -.02 | .27 | .44$^{*}$ | -.08 | .44$^{*}$ | -.14 | .22 | .32 | .41$^{*}$ | 1.00 |

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

(2) Correlation Coefficient among the Item Difficulty and the Content Attributes for QM6 of Booklet 2

| | Variables | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ItemDifficulty | 1.00 | | | | | | | | |
| 2 | b1integer | -.03 | 1.00 | | | | | | | |
| 3 | b2decimal | -.09 | -.82$^{**}$ | 1.00 | | | | | | |
| 4 | b3pattern | .37 | .29 | -.24 | 1.00 | | | | | |
| 5 | b4formula | .15 | .31 | -.22 | .24 | 1.00 | | | | |
| 6 | b5shape | .19 | .29 | -.49$^{**}$ | .17 | -.03 | 1.00 | | | |
| 7 | b6geomMeas | -.04 | .31 | -.40$^{*}$ | -.12 | .05 | .39$^{*}$ | 1.00 | | |
| 8 | b7dataRepre | -.31 | -.05 | .13 | -.12 | -.19 | -.24 | -.19 | 1.00 | |
| 9 | b8chance | -.15 | -.17 | .20 | -.05 | -.08 | -.10 | -.08 | -.08 | 1.00 |

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

(3) Correlation Coefficient among the Item Difficulty and the Attributes for QM2 ~ QM3-2 of Booklet 2 (with New Process Attributes)

| | Variables | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ItemDifficulty | 1.00 | | | | | | | | | | | | |
| 2 | b1integer | -.03 | 1.00 | | | | | | | | | | | |
| 3 | b2decimal | -.09 | -.82$^{**}$ | 1.00 | | | | | | | | | | |
| 4 | know_a1 | -.21 | .30 | .30 | 1.00 | | | | | | | | | |
| 5 | know_a2 | .22 | .44$^{*}$ | -.34 | .18 | 1.00 | | | | | | | | |
| 6 | know_a3 | .14 | .28 | -.58$^{**}$ | -.51$^{**}$ | -.01 | 1.00 | | | | | | | |
| 7 | know_a4 | -.36$^{*}$ | -.12 | .21 | .16 | -.26 | -.31 | 1.00 | | | | | | |
| 8 | c1identifing | -.41$^{*}$ | .00 | .00 | .01 | -.40$^{*}$ | -.03 | .66$^{**}$ | 1.00 | | | | | |
| 9 | c2computing | .10 | -.13 | .49$^{**}$ | .61$^{**}$ | .29 | -.50$^{**}$ | .07 | -.41$^{*}$ | 1.00 | | | | |
| 10 | c3judging | .28 | .35 | -.54$^{**}$ | -.31 | .21 | .43$^{*}$ | -.29 | -.28 | -.03 | 1.00 | | | |
| 11 | c4reasoning | .44$^{*}$ | .23 | -.35 | -.20 | .48$^{**}$ | .18 | -.19 | -.29 | -.02 | -.01 | 1.00 | | |
| 12 | IT2complexity | .62$^{**}$ | -.05 | .13 | .14 | .18 | -.28 | .01 | -.14 | .24 | -.06 | .35 | 1.00 | |
| 13 | IT3constructed | .54$^{**}$ | .14 | -.13 | .01 | .08 | -.03 | -.02 | -.13 | .08 | .02 | .32 | .41$^{*}$ | 1.00 |

$^{*}$ $p < .05$, $^{**}$ $p < .01$

(4) Correlation Coefficient among the Item Difficulty and the Attributes for QM4 and QM5-2 of Booklet 2 (with New Process Attributes)

| | Variables | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ItemDifficulty | 1.00 | | | | | | | | | | | | | | | | |
| 2 | b2decimal | -.09 | 1.00 | | | | | | | | | | | | | | | |
| 3 | know_a1 | -.21 | .30 | 1.00 | | | | | | | | | | | | | | |
| 4 | know_a2 | .22 | -.34 | .18 | 1.00 | | | | | | | | | | | | | |
| 5 | know_a3 | .14 | -.58$^{**}$ | -.51$^{**}$ | -.01 | 1.00 | | | | | | | | | | | | |
| 6 | know_a4 | -.36$^{*}$ | .21 | .16 | -.26 | -.31 | 1.00 | | | | | | | | | | | |
| 7 | d1compareNum | -.36$^{*}$ | .04 | .13 | -.21 | -.25 | .30 | 1.00 | | | | | | | | | | |
| 8 | d2recognizing | -.28 | -.09 | -.06 | -.32 | .11 | .64$^{**}$ | -.01 | 1.00 | | | | | | | | | |
| 9 | d3formulating | .39$^{*}$ | .41$^{*}$ | .24 | .08 | -.47$^{**}$ | -.02 | -.29 | -.13 | 1.00 | | | | | | | | |
| 10 | d4compu_numb | .05 | .58$^{**}$ | .51$^{**}$ | .01 | -.37$^{*}$ | .13 | -.18 | -.27 | .33 | 1.00 | | | | | | | |
| 11 | d5compu_alge | .14 | -.28 | .16 | .91$^{**}$ | -.13 | -.24 | -.19 | -.29 | .15 | -.05 | 1.00 | | | | | | |
| 12 | d7judg_rule | .24 | -.30 | .11 | .35 | -.21 | -.16 | -.13 | -.19 | -.01 | .21 | .39$^{*}$ | 1.00 | | | | | |
| 13 | d8judg_geom | .14 | -.40$^{*}$ | -.45$^{*}$ | -.03 | .69$^{**}$ | -.21 | -.17 | -.06 | -.33 | -.11 | .01 | -.14 | 1.00 | | | | |
| 14 | reas_a2 | .46$^{**}$ | -.30 | .11 | .61$^{**}$ | .03 | -.16 | -.13 | -.19 | .21 | -.03 | .39$^{*}$ | .26 | -.14 | 1.00 | | | |
| 15 | reas_a3 | .06 | -.17 | -.56$^{**}$ | -.10 | .29 | -.09 | -.07 | -.11 | -.14 | -.29 | -.09 | -.06 | -.08 | -.06 | 1.00 | | |
| 16 | IT2complexity | .62$^{**}$ | .13 | .14 | .18 | -.28 | .01 | -.17 | -.06 | .59$^{**}$ | .09 | .23 | .15 | -.19 | .45$^{*}$ | -.08 | 1.00 | |
| 17 | IT3constructed | .54$^{**}$ | -.13 | .01 | .08 | -.03 | -.02 | -.29 | .02 | .30 | .03 | -.02 | .21 | -.14 | .44$^{*}$ | -.14 | .41$^{*}$ | 1.00 |

$^{*}$ $p < .05$, $^{**}$ $p < .01$

(5) Correlation Coefficient among the Item Difficulty and the Attributes for QM1 ~ QM5-2 of Booklet 3 (with TIMSS Process Attributes)

|  | Variables | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ItemDifficulty | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2 | b1integer | -.01 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 3 | b2decimal | .03 | -.78$^{**}$ | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 4 | know_a1 | .02 | .36$^{*}$ | .30 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |
| 5 | know_a2 | .18 | .34 | -.31 | .07 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |
| 6 | know_a3 | .13 | .22 | -.49$^{**}$ | -.40$^{*}$ | .14 | 1.00 |  |  |  |  |  |  |  |  |  |  |
| 7 | know_a4 | -.10 | -.14 | .26 | .18 | -.33 | -.11 | 1.00 |  |  |  |  |  |  |  |  |  |
| 8 | appl_a1 | .14 | -.21 | .49$^{**}$ | .41$^{*}$ | -.26 | -.40$^{*}$ | .43$^{*}$ | 1.00 |  |  |  |  |  |  |  |  |
| 9 | appl_a2 | .21 | .26 | -.20 | .09 | .36$^{*}$ | .13 | -.12 | -.02 | 1.00 |  |  |  |  |  |  |  |
| 10 | appl_a3 | .07 | .33 | -.38$^{*}$ | -.07 | .17 | .77$^{**}$ | -.02 | -.20 | -.12 | 1.00 |  |  |  |  |  |  |
| 11 | reas_a1 | .21 | .26 | -.20 | .09 | -.18 | .13 | .21 | .23 | -.06 | .21 | 1.00 |  |  |  |  |  |
| 12 | reas_a3 | .07 | .01 | -.20 | -.29 | .09 | .41$^{*}$ | -.12 | -.28 | -.06 | .21 | -.06 | 1.00 |  |  |  |  |
| 13 | a6applying | .19 | -.02 | .26 | .36$^{*}$ | -.09 | .11 | .31 | .72$^{**}$ | .17 | .31 | .17 | -.11 | 1.00 |  |  |  |
| 14 | a7reasoning | .20 | .20 | -.30 | -.15 | -.07 | .40$^{*}$ | .07 | -.03 | -.09 | .31 | .68$^{**}$ | .68$^{**}$ | .04 | 1.00 |  |  |
| 15 | IT2complexity | .60$^{**}$ | -.24 | .35$^{*}$ | .16 | -.12 | -.07 | .02 | .39$^{*}$ | -.11 | .02 | .25 | -.11 | .28 | .10 | 1.00 |  |
| 16 | IT3constructed | .39$^{*}$ | -.10 | .16 | .09 | -.27 | -.04 | -.03 | .44$^{*}$ | -.19 | .13 | .34 | -.19 | .36$^{*}$ | .11 | .38$^{*}$ | 1.00 |

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

(6) Correlation Coefficient among the Item Difficulty and the Content Attributes for QM6 of Booklet 3

|  | Variables | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | ItemDifficulty | 1.00 |  |  |  |  |  |  |  |
| 2 | b1integer | -.01 | 1.00 |  |  |  |  |  |  |
| 3 | b2decimal | .03 | -.78$^{**}$ | 1.00 |  |  |  |  |  |
| 4 | b4formula | .29 | .09 | -.08 | 1.00 |  |  |  |  |
| 5 | b5shape | -.07 | .17 | -.38$^{*}$ | -.11 | 1.00 |  |  |  |
| 6 | b6geomMeas | .15 | .38$^{*}$ | -.42$^{*}$ | .02 | .52$^{**}$ | 1.00 |  |  |
| 7 | b7dataRepre | -.02 | .10 | .01 | -.26 | .02 | .19 | 1.00 |  |
| 8 | b8chance | .03 | -.10 | .18 | -.19 | -.15 | .09 | .16 | 1.00 |

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

(7) Correlation Coefficient among the Item Difficulty and the Attributes for QM2 ~ QM3-2 of Booklet 3 (with New Process Attributes)

| | Variables | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ItemDifficulty | 1.00 | | | | | | | | | | | | |
| 2 | b1integer | -.01 | 1.00 | | | | | | | | | | | |
| 3 | b2decimal | .03 | -.78$^{**}$ | 1.00 | | | | | | | | | | |
| 4 | know_a1 | .02 | .36$^{*}$ | .30 | 1.00 | | | | | | | | | |
| 5 | know_a2 | .18 | .34 | -.31 | .07 | 1.00 | | | | | | | | |
| 6 | know_a3 | .13 | .22 | -.49$^{**}$ | -.40$^{*}$ | .14 | 1.00 | | | | | | | |
| 7 | know_a4 | -.10 | -.14 | .26 | .18 | -.33 | -.11 | 1.00 | | | | | | |
| 8 | c1identifing | -.23 | -.04 | .09 | .07 | -.36$^{*}$ | .00 | .67$^{**}$ | 1.00 | | | | | |
| 9 | c2computing | .08 | -.02 | .31 | .44$^{*}$ | .40$^{*}$ | -.29 | -.28 | -.50$^{**}$ | 1.00 | | | | |
| 10 | c3judging | .21 | .53$^{**}$ | -.61$^{**}$ | -.11 | .40$^{*}$ | .67$^{**}$ | -.19 | -.13 | -.01 | 1.00 | | | |
| 11 | c4reasoning | .20 | .20 | -.30 | -.15 | -.07 | .40$^{*}$ | .07 | .13 | -.44$^{*}$ | .11 | 1.00 | | |
| 12 | IT2complexity | .60$^{**}$ | -.24 | .35$^{*}$ | .16 | -.12 | -.07 | .02 | -.12 | .04 | -.14 | .10 | 1.00 | |
| 13 | IT3constructed | .39$^{*}$ | -.10 | .16 | .09 | -.27 | -.04 | -.03 | .00 | -.01 | -.18 | .11 | .38$^{*}$ | 1.00 |

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

(8) Correlation Coefficient among the Item Difficulty and the Attributes for QM4 and QM5-2 of Booklet 3 (with New Process Attributes)

| | Variables | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ItemDifficulty | 1.00 | | | | | | | | | | | | | | | | |
| 2 | b2decimal | .03 | 1.00 | | | | | | | | | | | | | | | |
| 3 | know_a1 | .02 | .30 | 1.00 | | | | | | | | | | | | | | |
| 4 | know_a2 | .18 | -.31 | .07 | 1.00 | | | | | | | | | | | | | |
| 5 | know_a3 | .13 | -.49$^{**}$ | -.40$^{*}$ | .14 | 1.00 | | | | | | | | | | | | |
| 6 | know_a4 | -.10 | .26 | .18 | -.33 | -.11 | 1.00 | | | | | | | | | | | |
| 7 | d1compareNum | -.12 | .19 | .19 | -.37$^{*}$ | -.15 | .72$^{**}$ | 1.00 | | | | | | | | | | |
| 8 | d2recognizing | -.10 | -.11 | -.03 | -.21 | .18 | .52$^{**}$ | .27 | 1.00 | | | | | | | | | |
| 9 | d3formulating | .17 | .41$^{*}$ | .04 | -.05 | -.40$^{*}$ | -.31 | -.34 | -.34 | 1.00 | | | | | | | | |
| 10 | d4compu_numb | .11 | .46$^{**}$ | .66$^{**}$ | .25 | -.29 | -.10 | -.23 | -.40$^{*}$ | .22 | 1.00 | | | | | | | |
| 11 | d5compu_alge | .18 | -.31 | .07 | 1.00$^{**}$ | .14 | -.33 | -.37$^{*}$ | -.21 | -.05 | .25 | 1.00 | | | | | | |
| 12 | d7judg_rule | .10 | -.30 | .14 | .33 | -.23 | -.18 | -.19 | -.19 | -.24 | .21 | .33 | 1.00 | | | | | |
| 13 | d8judg_geom | .16 | -.46$^{**}$ | -.22 | .20 | .92$^{**}$ | -.08 | -.12 | .23 | -.37$^{*}$ | -.18 | .20 | -.21 | 1.00 | | | | |
| 14 | reas_a1 | .21 | -.20 | .09 | -.18 | .13 | .21 | .49$^{**}$ | .18 | -.17 | -.45$^{**}$ | -.18 | -.09 | .15 | 1.00 | | | |
| 15 | reas_a3 | .07 | -.20 | -.29 | .09 | .41$^{*}$ | -.12 | -.13 | -.13 | -.17 | -.15 | .09 | -.09 | .15 | -.06 | 1.00 | | |
| 16 | IT2complexity | .60$^{**}$ | .35$^{*}$ | .16 | -.12 | -.07 | .02 | -.01 | -.01 | .46$^{**}$ | .04 | -.12 | -.16 | -.04 | .25 | -.11 | 1.00 | |
| 17 | IT3constructed | .39$^{*}$ | .16 | .09 | -.27 | -.04 | -.03 | .07 | .07 | .32 | -.01 | -.27 | -.28 | .01 | .34 | -.19 | .38$^{*}$ | 1.00 |

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