Innovative Assessments that Support Students' STEM Learning

Phonraphee Thummaphan

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Min Li, Chair

Jimmy de la Torre

Katherine Lewis

Program Authorized to Offer Degree:

College of Education

University of Washington

Abstract

Innovative Assessments that Support Students' STEM Learning

Phonraphee Thummaphan

Chair of the Supervisory Committee:

Professor Min Li

College of Education

The present study aimed to represent the innovative assessments that support students' learning in STEM education through using the integrative framework for Cognitive Diagnostic Modeling (CDM). This framework is based on three components, cognition, observation, and interpretation (National Research Council, 2001). Specifically, this dissertation demonstrates how this framework combines psychometrics and cognitive psychology, and utilizes the integrative nature of cognition, observation, and interpretation in science and math assessments.

At present, STEM assessments do not fully support students' learning (National Research Council, 2001; Songer & Ruiz-Primo, 2012). We need innovative well-defined assessments that respond to students' and educators' needs, particularly for formative purposes. Using the three components of the assessment triangle, cognition, observation, and interpretation (National Research Council, 2001), this study articulated an integrative framework grounded in both a

psychometric model and cognitive theory. This framework can both validate learning theory and provide assessment information with diagnostic and formative implications. Guided by this framework, the CDM approach can uniquely support any innovative assessment, as it combines psychology of learning and statistical methods to make inferences about students' specific knowledge structures and processing skills (Alvers, 2012; de la Torre & Minchen, 2014). Nevertheless, this framework must be carefully applied to assure that the assessment procedures are cohesive, thus providing good support for students' learning. Specifically, the research questions are: *What is the integrative framework for CDM?* and *How can this CDM integrative framework be applied to develop and validate STEM assessments?*

To answer these research questions, this dissertation includes three publications that demonstrate what an integrative framework for CDM is and how this framework can be applied. The first paper, "*Models for cognitive diagnostic modeling,*" focuses on fundamental knowledge about the models used for CDM. Grounded in a systematic review of the literature, the integrative framework for assessments is defined based on the components of cognition, observation, and interpretation. This integrative CDM framework, based upon cognitive science and statistical techniques, can be used as a guide for performing CDM analysis.

The second and third papers focus on applying CDM in science and math assessments. Specifically, the second paper, "*A cognitive diagnostic modeling approach to instructional sensitivity*," emphasizes using CDM for elementary science data analysis in relation to instructional sensitivity. To answer the research questions, *Are assessments sensitive enough to detect student learning differences due to instruction? If so, do they have formative value for teachers and students?*, we examined the formative value of instructional sensitivity of assessment items from two elementary science modules. To determine whether items varying in

instructional sensitivity yield different formative values for diagnosing student learning, we created booklets with items of different instructional proximity (from *close* to *far proximal*), and administered them in 38 classrooms (824 students) using a pretest-posttest design. Incorporated into the CDM analysis, the item and test indices show that the data fit well with the specified Q-matrix. Attributes with higher gain had been heavily addressed in the intended curriculum (i.e., a greater number of learning activities) compared to those with relatively smaller gain.

The third paper, "Examining the relationship of characteristics of word problems and item parameters in the context of an online math game," demonstrates the application of CDM for math word problem game data. The main research question was *How are item characteristics of word problems associated with item parameters*? The sample included 225 Grade 4-6 players and their performance on 22 items across two booklets. We performed a correlation to investigate the relationship of item characteristics and CDM item parameters. Results showed that consistency and model type were significantly correlated with item difficulty. The sequence analysis with students' action log data provided visualization of their modeling strategies that further validated the results of CDM and correlation.

This framework consists of two elements—CDM as theory-driven model and statistical procedures—and tends to improve the capacity of CDM by increasing the formative values of assessment and validate any cognitive/learning theory that guides the design of assessments. This dissertation illustrates how such a framework can be applied to *develop and validate STEM assessments*. Specifically, it can be used to generate indices of the instructional sensitivity of assessment items which allow score interpretations at a finer grid. Moreover, it can be used to examine the relationship between item characteristics and item parameters that can in turn guide assessment development and interpretation.

TABLE OF CONTENTS

List of Figures…… ii

List of Tables……. iii

Acknowledgements…… iv

Introduction…… 1

Section 1 Model for Cognitive Diagnostic Modeling…… 6

Overview of CDM……...……... 6

Commonly Used Cognitive Diagnostic Models……. 9

Steps Involved in CDM…….....**.....** 12

An Integrative Framework ……...……. 18

Section 2 A Cognitive Diagnostic Modeling Approach to Instructional Sensitivity…….. 28

Introduction…… 30

Methods……...……... 35

Results…….……... 41

Conclusions…… 49

Section 3 Examining the Relationship of Characteristics of Word Problems and Item Parameters in the Context of an Online Math Game……...……...…… 53

Introduction…… 54

Methods…….……. 60

Results…….……... 69

Conclusions…… 77

Conclusion……..…… 80

Future Directions of Research…….……... 82

References……... 84

Appendix..……... 93

LIST OF FIGURES

Figure 1: Examples of DEISA Items for the LF Module..... 38

Figure 2: Attribute Mastery Level by Attribute (in percentage)..... 44

Figure 1: Examples of Word Problem Items from Different Sources Included in the Game..... 62

Figure 2: Examples of Setting up the Model and Equation Steps in the Game..... 63

Figure 3: Attribute Mastery Level by Attribute (in percentage)..... 71

Figure 4: Sequence Analysis Results of the Cluster 2 of Item Group 3..... 75

Figure 5: Sequence Analysis Results of the Cluster 1 of Item Group 2 and 3 of Item Group 4 for Booklet..... 77

LIST OF TABLES

Table 1: Characteristic of CDMs……..……... 7

Table 2: Comparison of Commonly Used Models…….……..... 12

Table 1: Attribute Definitions for the Items of the Two Science Modules…… 36

Table 2a: Q-matrix for the LF Module…….. 40

Table 2b: Q-matrix for the LF Module…….. 40

Table 3: Information about the School Districts…….……..……..…… 40

Table 4: Item-level Fit Statistics……. 42

Table 5: Item Parameters from the CDM Analysis…… 48

Table 6: Item Parameters Related to Instructional Sensitivity……. 49

Table 1: Attribute Definitions for the Items of the Two Booklets …….. 61

Table 2: Q-matrix for Modeling……... 65

Table 3: Groups of Items ……. 68

Table 4: Item-level Fit Statistics……. 70

Table 5: Item Parameters…….. 73

Table 6: Correlation Coefficients between Item Characteristics and Item Difficulty …….. 73

ACKNOWLEDGEMENTS

There are many people I would like to thank for their support throughout the writing of this dissertation and during my lengthy term in graduate school. First and foremost, I give my heartfelt thanks to Dr. Min Li, my committee chair and academic advisor. Working with her has been an amazing opportunity. She has not only dedicated an enormous amount of time and energy providing academic and professional guidance and support, but has also been a great role model as a teacher and researcher. She has consistently encouraged me to pursue my academic life purpose and passion, and to fulfill my potential in regard to social contribution. I cannot find the words to fully express my gratitude for her encouragement and help.

Second, I would especially like to thank Dr. Jimmy de la Torre, who is always available to provide support and who introduced me to one of my interest, Cognitive Diagnostic Modeling. Similarly, my sincere thanks go to Dr. Katherine Lewis for her greatly help and advice from math education perspectives. Also, I appreciate the support and dedication of Dr. Dagmar Amtmann for joining my committee despite her many other commitments and obligations. Moreover, I would like to express deep appreciation to Dr. Cathy Taylor, who has always provided full support for my journey. Her passion for rigorous assessment practices and policy continues to inspire my own passion for state-of-the-art assessment research. In addition, a million thanks go out to Dr. Zoran Popović for giving me the opportunity to fully participate in his research in game development and leaning, which became one of my passions. Many thanks also to Roy Szeto who has always generously provided data support when I need it.

There are many others I would like to thank, including my many fellow graduate students, the research scientists at Center for Assessment, and friends all over the world. In

particular, Drs. Linda Liaw, Ting Wang, and Dongsheng Dong, my M&S classmates, I thank you for the technical discussions and for helping me navigate through the program. I also thank Pao Baylon and Daniel Yoo for being not only great roommates but also brothers who support me in everything. They are atmospheric and political scientists, yet they patiently listened to me talk about measurement research for hours on end. Special thanks to Alec Kennedy in the UW Evans School who is always available in offering his help in statistical programming. With his help, I have been much more comfortable with finishing my dissertation.

Finally, I am forever indebted to my family for their lifelong encouragement, patience, and sacrifice, which have allowed me to study abroad in order to achieve my dreams. Without their unwavering love and support, I never would have made it to this point.

INTRODUCTION

Various assessments have been developed in STEM education, but many do not fully support students' learning (National Research Council, 2001; Songer & Ruiz-Primo, 2012). Assessments should effectively provide useful and timely information for individual and personalized learning support, responding to both formal and informal learning contexts, as well as increasingly diverse student populations, in terms of achievement, special needs, and so on. However, the commonly used assessments such as most multiple-choice tests rarely serve these diverse instructional contexts because they mostly focus on the summative assessment purpose, suggesting the need for innovative assessments.

Educators have stressed the formative purposes of assessment (Pellegrino, 2014). In addition to using assessment as summative and evaluative tools, many have pointed to the growing need for enhancing the formative value of assessments. Formative assessment offers teachers and students the opportunity to make decisions and actions based upon timely information that meets students' learning needs during the instructional process (Black et al., 2003; Chappuis & Chappuis, 2008). Specifically, teachers need information regarding individual students' learning progression and status of mastery in order to evaluate learning progress and figure out how to enhance student learning (Stiggins, 2008). Heritage (2007) also indicates that teachers need to understand students' prior knowledge in a specific content area in order to build on students' previous learning. With such useful information, instruction can more effectively promote students' learning. As mentioned by Wiggins (1998), the ultimate goal of assessment should be "to educate and improve student performance, not merely to audit it" (p. 7).

Innovations in assessment should aim to not only provide useful assessment information, but also aim to support students' deep and active learning, high motivation and engagement, knowledge transfer, and building of skills, as well as to provide a meaningful

learning experience for students and teachers (Mowl, 2006). Specifically, innovation is needed for STEM assessment to support students' learning for several reasons. First, current assessment has not yet adequately addressed the important constructs stressed by the Next Generation Science Standards (NGSS) and Common Core State Standards (CCSS) thus, innovation needs to broaden the types of constructs that can be assessed, for example, modeling, problem-solving strategies, and understanding and application of cross-cutting concepts. Second, because STEM learning situations involve modern and complex settings such as game-based learning and tablet-based learning, innovation needs to provide just-in-time diagnostic and formative assessment to support students' learning. Third, the interpretation and use of assessment information will be valid only if the assessment is grounded in learning or cognition models; thus, innovation needs to incorporate cognitive science and measurement methodology.

What we require are well-defined innovative assessments that can respond to the aforementioned needs of students and educators. However, the way these assessments have been defined in previous studies needs refining if we are to address the well-grounded components of assessments. For example, Mowl (2006) defines innovative assessments as follows: "Innovative assessment … is a term we use which encompasses a whole range of different techniques and methods, not all of which are new inventions. What unites them is a common goal: to improve the quality of student learning" (p. 2). Looney (2009) defines them by approaches that refer to methodology and technology: "Innovative approaches [to testing]… may include new methodologies for assessment, as well as new technologies that can measure complex skills and reasoning processes" (p. 18). Both definitions provide good groundwork for defining this term, but they fail to characterize the complete elements of the assessment. Therefore, by using the assessment triangle, which includes three components, cognition, observation, and interpretation (National Research Council, 2001), I take into

consideration that innovations ought to incorporate state-of-art cognitive models, involve novel methods of eliciting observable knowledge and skills, and utilize powerful psychometrics techniques for generating interpretable results. Simply speaking, innovative assessments should provide valid inferences about student performance based on test scores (Cronbach, 1988; Messick, 1989; Kane, 1992, 2013; Haertel, 2013). So that assessment information can be effectively used to support students' learning needs.

Applying my notion of innovative assessment, current assessment innovations typically fall into several formats or methods. In the 1990s, there were significant improvements in performance-based assessment (Ananda & Rabinowitz, 2001). Recent examples include the edTPA for teacher candidates developed by Stanford University and the American Association of Colleges for Teacher Education (AACTE), and the new Advanced Placement Computer Science Principles Test (AP CSP)developed by College Board. The distinctive characteristics of performance-based assessment are the use of tasks or situations that provide subtle indicators of students' performance. Nonetheless, these indicators are simply observable, not solely based on the underlying cognitive/learning process, and thus accommodate limited interpretations. The more recent innovative assessment is technology-based assessment (Rabinowitz & Brandt, 2001), for example, computer adaptive testing (CAT) that tailors test items based on the examinee's ability level (See Linacre, n.d.; Davey, 2011). However, CAT per se does not automatically make an assessment innovative. The lack of emphasis on using a cognition model for assessment development results in limited usefulness of assessment information despite the fact that technology can strengthen the observations of construct and increase interpretation capacity of analytics tools. In short, these assessment innovations lack components of the assessment triangle, in particular the cognition component, necessary to yield a truly innovative assessment.

An integrative framework grounded in both psychometric and assessment models and cognitive theory is needed (National Research Council, 2001). Such a framework can be used for developing and validating STEM assessments so that they can provide assessment information that has diagnostic and formative implications. The present framework is based on a cognition model for constructing assessment items or tasks, and applied vigorous measurement techniques that can provide specific information about students' knowledge/skills that can be readily used by teachers and learners. The unique characteristics of this framework systemize the process of assessment development and validation in order to offer promising opportunities for both instructional improvement and theory development.

Guided by this integrative framework, the potentials of any innovative assessment can be uniquely supported by the characteristics of the Cognitive Diagnostic Modeling (CDM) approach, as it combines psychology of learning and statistical methods to make inferences about students' specific knowledge structures and processing skills (Alvers, 2012; de la Torre & Minchen, 2014). Nevertheless, the integrative framework for CDM needs to be carefully designed and implemented to make sure that the assessments are cohesive throughout the procedures, thus providing good support for students' learning. The research focus of this dissertation is how to develop and validate innovative assessments based on this framework. Specifically, the research questions are: *What is the integrative framework for CDM?* and *How can this CDM integrative framework be applied to develop and validate STEM assessments?*

To answer these research questions, this dissertation includes three publications that demonstrate what the integrative framework is and how this integrative framework can be applied. The following publications focus on these innovative assessment issues: (1) *Models for cognitive diagnostic modeling*, (2) *A cognitive diagnostic modeling approach to instructional sensitivity*, and (3) *Examining the relationship of characteristics of word*

problems and item parameters in the context of an online math game. These papers, representing my interest around *innovative assessments*, explore ways to ensure the diagnostic value of assessments by combining applied measurement and learning sciences in science and mathematics education.

These three papers are coherent in terms of concept and content. They represent the critical theoretical and practical knowledge of CDM in STEM assessment. The first paper, "*Models for cognitive diagnostic modeling*," focuses on the fundamental knowledge about the models used for CDM. It presents an integrative CDM framework based upon cognitive science and statistical techniques that can be used as a guideline for performing CDM analysis. The second and third papers focus on the application of CDM in science and math assessments. Specifically, the second paper, "*A cognitive diagnostic modeling approach to instructional sensitivity*," emphasizes the application of CDM for elementary science data analysis in relation to the instructional sensitivity aspects. The findings demonstrate the utility of the CDM item parameters to quantify the level of instructional sensitivity and present the students' attribute mastery level. The third paper, "*Examining the Relationship of Characteristics of Word Problems and Item Parameters in the Context of an Online Math Game*" demonstrates the application of CDM for math word problem game data. It not only examines students' modeling performance that can be used for instructional planning but also evaluates the relationship of item features and CDM estimated item parameters. The details of each paper are presented in the following sections.

SECTION 1

MODELS FOR COGNITIVE DIAGNOSTIC MODELING

Phonraphee Thummaphan[1], Min Li[1], Jimmy de la Torre[2]

[1]University of Washington, Seattle,[2] Rutgers, The State University of New Jersey

Thummaphan, P., Li, M., & de la Torre, J. (2016). Models for cognitive diagnostic modeling. In V. Kijtorntham (Ed.). *Methodological and Theoretical Articles for Behavioral Science Research in Community and School*. Bangkok: Behavioral Science Research Institute.

Psychometric techniques have been developed to measure cognitive constructs and provide practical implications for assessment. Cognitive diagnostic modeling (CDM) is a measurement method that is useful for diagnosing the cognitive mastery of examinees. CDM involves various models and steps. We start by providing an overview of CDM. Then we describe four psychometric models for cognitive diagnosis that are most commonly encountered in the literature by discussing their assumptions, strengths, and weaknesses, and the extent to which they relate to each other. Also, we discuss the different steps involved in CDM (i.e., attribute definition, different types of validation, model fitting, and different approaches to model selection). Finally, we propose an integrative framework with which one may carry out CDM.

Overview of CDM

Definition of CDM

CDM which involves a cognitive diagnosis model (also known as diagnostic classification model), is a psychometric modeling approach that can be used to measure students' cognitive skills or knowledge required to answer items correctly. Cognitive Diagnostic Models are "latent variable models developed primarily for assessing student mastery and nonmastery on a set of finer-grained skills" (de la Torre, 2011, p. 179). The goal of CDM is to measure student mastery of the required skills, and to evaluate the diagnostic

power of item measuring such skills (Hartz, 2002). The main advantage of CDM is that they provide a detailed profile on student understanding by specifying strengths and weaknesses of individual learners rather than an overall score. Thus, assessment results can be used for diagnosis and improvement of student learning (Pellegrino, 2013).

Various CDM approaches have been developed. In what follows, the word "and" refers to models that are non-compensatory and "or" refers to models that are compensatory. Models include the *deterministic inputs, noisy and-gate* and *noisy-input, deterministic 'and' gate* (DINA and NIDA; e.g., Junker & Sijtsma, 2001), *multiple-choice DINA* (MC-DINA; e.g. de la Torre, 2009a), *multi-strategy DINA* (MS-DINA; de la Torre & Douglas, 2008), *deterministic inputs, noisy 'or' or-gate* and *noisy-input, deterministic 'or' or-gate* (DINO and NIDO; e.g., Templin & Henson, 2006), *reparametrized unified model* (RUM/fusion; e.g., DiBello, Stout, & Roussos, 1995; Hartz, 2002), *compensatory reparameterized unified model* (C-RUM; Hartz, 2002) *noncompensatory reparameterized unified model* (NC-RUM; e.g., DiBello et al., 1995; Hartz, 2002), *reduced reparameterized unified model* (rRUM; Hartz, 2002), *multiple classification latent class model* (MCLCM; e.g. Maris, 1999), *general diagnostic model* (GDM; e.g., von Davier, 2005; Xu & von Davier, 2006), *log-linear cognitive diagnostic model* (LCDM; e.g., von Davier, 2005, 2014), *nominal response LCDM* (NR-LCDM; e.g., Templin et al., 2008), and *generalized DINA* (G-DINA; e.g., de la Torre, 2011).

Each CDM model has different characteristics. Rupp and Templin (2008) classified CDMs by jointly considering three characteristics: (1) measurement scales of observed response variables (dichotomous vs. polytomous), (2) measurement scales of latent predictor variables (dichotomous vs. polytomous), and (3) model type based on the combination rule of latent predictor variables (compensatory vs. noncompensatory). Gu (2011) made the classification more complete by adding the Nominal Response and Polytomous, but not

Ordered models to the observed response category. Characteristics of each model (Table 1) were determined by summarizing Rupp and Templin (2008) and Gu (2011), as well as by adding conjunctivity information.

Table 1. *Characteristic of CDMs*

Model	Observed Response Variables			Latent Predictor Variables		Combination Rule		Conjunctivity	
	dichotomous	Polytomous	Nominal	dichotomous	polytomous	Compensatory	non-compensatory	Conjunctive	disjunctive
DINA	✓			✓			✓	✓	
MS-DINA	✓			✓			✓	✓	
NIDA	✓			✓			✓	✓	
MCLCM	✓	✓		✓	✓	✓	✓	✓	✓
NC-RUM	✓	✓		✓	✓	✓	✓	✓	
rRUM	✓			✓			✓	✓	
DINO	✓			✓		✓			✓
NIDO	✓			✓		✓			✓
C-RUM	✓	✓		✓	✓	✓		✓	
GDM	✓	✓		✓	✓	✓		✓	
LCDM	✓	✓		✓	✓	✓		✓	
MC-DINA			✓	✓			✓	✓	
NR-LCDM			✓	✓		✓		✓	
G-DINA	✓			✓		✓	✓	✓	✓

As shown in Table 1, in addition to the types of observed response and latent predictor variables, the combination rule of latent predictor variables is used to classify the models. Rupp and Templin (2008, p. 240) note that "the difference between *compensatory* and *noncompensatory* models reflects how the latent predictor variables are combined across the different skills to produce the observed responses." That is, compensatory models assume that a lack of one skill can be compensated for by a possession of another skill. In contrast, a noncompensatory model assumes that a lack of one skill cannot be compensated for by having another skill.

One more assumption related to CDM is conjunctivity, or the condensation rule, which makes assumptions about how a correct response occurs. Conjunctive models assume

that correct responses occur when the examinee masters *all* "required" attributes, while disjunctive models assume that mastery of one or more "required" attributes is sufficient to produce the correct response (Broaddus & Shaftel, 2012; Junker, 1999; Rupp, Templin, & Henson, 2008). However, Junker (1999) mentions that it remains unclear how the attributes combine to produce a response. Moreover, conjunctivity can be related to the combination rule as well. For example, disjunctive models can be viewed as compensatory because one attribute can compensate for others and produce the correct response, particularly for the completely compensatory case (Junker, 1999; de la Torre & Douglas, 2004).

It is important to note that the *attribute hierarchy method* (AHM; e.g., Leighton, Gierl, & Hunka, 2004) and *rule-space methodology* (RSM; e.g., Tatsuoka, 1983, 1995) which are usually found in the literature, are not really CDMs, but are frameworks. To illustrate, while models involve a direct statistical link between observed response variables and latent predictor variables, RSM and AHM (an extension of the RSM) do not. Thus, they "are essentially classification algorithms and not unified statistical models that are completely embedded within a fully probabilistic framework" (Rupp & Templin, 2008, p. 238). Moreover, *Bayesian inference networks* (BINs; e.g., Yan, Mislevy, & Almond, 1993), are not single models, as are the others, but is instead "a general modeling framework for representing different kinds of latent variable models" (Rupp & Templin, 2008, p. 241). In addition, G-DINA is a general model in the sense that it encompasses models such as DINA, DINO,or rRUM, depending on the constraints. As such, G-DINA can represent any model types classified as a compensatory or conjunctive.

Commonly Used Cognitive Diagnostic Models

A number of commonly used CDMs are related to each other (de la Torre & Douglas, 2004; de la Torre & Chiu, 2010; de la Torre, 2011; Rojas, de la Torre, & Olea, 2012). In this

paper, we present four such models: DINA, NIDA, DINO, and G-DINA, with DINA as the focus of the paper and the model used as a point of comparison for the other models.

The DINA Model

The DINA model (deterministic inputs, noisy ‘‘and’’ gate; e.g., Haertel, 1989; Junker & Sijtsma, 2001; de la Torre & Doulas, 2004; de la Torre, 2009b) is considered one of the most parsimonious and interpretable CDMs, as it requires only two parameters for each item regardless of the number of attributes involved (de la Torre, 2009; de la Torre, 2011; Lee, Park, & Taylan, 2011). The DINA model is a noncompensatory and conjunctive model. Taking the deterministic perspective, the DINA assumes that examinees who only have all the required attributes can answer items correctly. It also assumes that an examinee cannot compensate with any other attribute that he/she has, regardless of its high magnitude (Rupp et al., 2008). Thus, the DINA model classifies examinees into two classes on each item: (1) item masters, comprising those who have all required attributes to answer the item correctly, and (2) item nonmasters, comprising those who are otherwise (Rupp et al., 2008). However, if one takes the probabilistic view, examinees can still answer items correctly or incorrectly, based on what Rupp et al. (2008) call “lucky guesses” or “careless errors.” Original terms are “guess” and “slip” (Junker & Sijtsma, 2001).

The DINA formula model includes three important components: the latent variable, slipping parameter, and guessing parameter (Rupp et al., 2008), as shown in the probability of a correct response of Examinee i who has the skills vector or latent class α_i for Item j as follows:

$$P_j(\alpha_i) = P\left(X_{ij} = 1 \,\middle|\, \alpha_i\right) = {g_j}^{1-\eta_{ij}}(1 - s_{ij})^{\eta_{ij}},$$

where $P_j(\alpha_i)$ is the probability of a correct response for item j in skills vector α_i, $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$ is the binary indicator of examinee mastery status of all required attributes

(K is the number of attributes), X_{ij} is the observed response for Item j, $\alpha_i$is the skills vector …,$(1 - s_{ij})$ is the probability of not slipping, and g_j is the probability of guessing.

It is important to understand the term *deterministic inputs, noisy''and''gate*, which describes the characteristics of this model. In this formula, η_{ij} is the latent variable that is the *deterministic input* of the DINA model used to indicate or classify examinee mastery for the item. The *and-gate* is the conjunctive process, indicating whether or not all required attributes are present, which creates the variable η_{ij}. The *and-gate* "functions like an output summary that takes the individual elements and condenses them much as a gate in a park forces people to enter and exit in a particular way" (Rupp et al., 2008, p. 117). *Noisy* refers to slip and guessing parameters – slip indicates answering the item incorrectly due to a careless error despite the examinee having all required attributes, whereas guessing refers to answering the item correctly by guessing even though the examinee lacks the required attributes (Rupp et al., 2008; de la Torre, 2009b).

Therefore, the DINA model provides one slip and one guessing parameter for each item, regardless of the number of attributes involved. That is, the total number of parameters estimated in the DINA model is determined by the number of items instead of the number of attributes (Rupp et al., 2008). This makes the DINA model parsimonious, but it requires that examinees must master all the measured attributes in order to answer the items correctly. Again, this model does not discriminate between those who lack different numbers of required attributes, thus making it very restrictive (de la Torre, 2013).

NIDA, DINO, and G-DINA

Table 2 compares NIDA, DINO, and Generalized DINA (G-DINA) against DINA, demonstrating the various strengths and weaknesses of each model. Briefly, while being parsimonious, DINA is very restrictive. The G-DINA model is the more flexible since it permits examinees with different attribute patterns to have different probabilities to answer

the item correctly. However, it involves more parameters, and estimation takes a longer time, and also requires a larger sample size (de la Torre & Lee, 2013). NIDA is more flexible than DINA; however, parameters are still constrained to be equal across items. DINO is equally restrictive to DINA, but in different ways. DINO model requires possession of at least one required attribute to produce correct answer, whereas in DINA, all attributes are required for providing the correct answer.

Table 2. *Comparison of Commonly Used Models*

Model	Number of Parameter Estimates	Level of Parameter Estimation	Strengths	Weaknesses
DINA	2 per item	Item	- Parsimonious, interpretable - Small sample size	- Restrictive: all attributes require the correct answer; cannot differentiate examinees who lack different attributes
NIDA	2 per attribute	Attribute	- Parsimonious, interpretable - Small sample size - Can distinguish examinees who lack different attributes	- Parameters are constrained across items; may not be flexible enough for some analyses.
DINO	2 per item	Item	- Parsimonious, interpretable - Small sample size - Allows compensation of attributes	- Restrictive: cannot differentiate examinees who lack different attributes; may not be flexible enough for some analyses
G-DINA	Can be more than 2 per item[a]	Item	-General, more flexible	- Requires higher number of parameters - Requires larger sample size - Takes time to converge

Note: a. The number of attributes depends on the number of attributes required for each item.

Steps Involved in CDM

Overall, CDM applies the latent class model approach by classifying examinees according to the attribute patterns – the combinations of attributes possessed or not possessed. Based on students' responses on items, CDM iteratively fits the model with a fixed number of classes, given the number of attributes set prior to the analysis, and evaluates the model fit.

CDM is different from CDA[1] (Cognitive Diagnostic Assessment) as the former is the model, the latter is the assessment. The important steps involved and presented in this section are about attribute definition, Q-matrix, model fitting, and validation.

Attribute Definition

One of the most important steps for CDMs is attribute definition, which is the process of determining the specification and definition of the skill(s) of interest. Roussos et al. (2010, p. 38) explained that "in general, this component of the implementation process considers how many and what kinds of skills or attributes are involved, at what level(s) of difficulty, and in what form of interaction." So this process requires a clear assessment purpose and contemporary cognitive or leaning theory that informs the knowledge structures and processes that a learner would use in answering/successfully completing the questions/tasks.

The attribute definition, which is the document with the definition details of all attributes, is developed based on the cognitive model or relevant framework of diagnostic interest. For example, Lee et al. (2011) developed the attributes of TIMSS mathematics knowledge and skills from the 2007 TIMSS Mathematics Framework (Mullis et al., 2005). The decision about the fine-grained sizes of attributes should be made by considering both theoretical (i.e. construct representativeness), and practical factors (i.e. the purposes and context of using assessment information) (Jang, 2009). The definitions of attributes are usually developed by researchers and verified by experts' judgment. For each attribute, a clear description of knowledge and skills should be provided in order to effectively guide the development or selection of assessment items.

Below is an example of the attribute definition developed by Thummaphan, Dong, and Li (2015). This definition is used to analyze mathematics word problems at Grade 4 level based on the CCSS. The definition includes the label and description of the underlying understanding as well as a tip for coding word problems used in the study.

C1 (Whole Numbers)

A student who has mastered this attribute should be able to understand and apply basic concepts and operations in whole numbers:

- Including addition, subtraction, multiplication, division, exponentiation, sign, absolute value, place value, rounding of 2- or 3-digit integer values, number patterns, comparison of magnitude of numbers, greatest common divisor, least common multiple, etc.

The Q-matrix

After the assessment items are developed or selected, the next step, which is unique for CDM, is the production of the Q-matrix, an item-by-attribute binary matrix. Providing the correct Q-matrix specification is not an easy task as it must take into account the combination and structure of skills (DeCarlo, 2011). In general, the rows are the items in the test and columns are the attributes required to answer items correctly. The matrix must cover all items. Each item will be analyzed and coded based on the attribute definition to determine which attributes are required for answering each item correctly. The numbers indicating whether an attribute is required or not to answer an item correctly appear in the Q-matrix as 1 or 0, respectively. The Q-matrix elements are usually determined by experts' judgment and verified by researchers to examine inter-rater reliability. Therefore, the coders must be trained to fully understand the attribute definition. Low reliability among coders requires reexamination of the attribute definition and reconsidering the coders' understanding of the attribute definition.

Model Fitting

Methods of Estimation

There are two common model estimation methods: expectation maximization (EM) and Markov chain Monte Carlo (MCMC). EM uses the marginal maximum likelihood estimation (MMLE) method with an iterative procedure to reach the maximum likelihood parameter values (von Davier & Yamamoto, 2004; Rupp et al., 2010). In contrast, MCMC

applies the Bayesian estimation method to estimate the posterior distributions for all parameters using sampling methods (Rupp et al., 2010).

The EM algorithm is a common method for estimating latent variable models and latent class models, and takes less time compared to MCMC (DiBello, Roussos, & Stout, 2007; Rupp et al., 2010; von Davier & Yamamoto, 2004). However, if the number of latent classes and the set of constraints increase, EM estimation becomes computationally intensive and complex (DiBello et al., 2007; Rupp et al., 2010). Therefore, the EM algorithm has been used to estimate more straightforward models, whereas MCMC has been used to handle more models with more complex formulations (de la Torre, 2009b; Liu, Douglas, & Henson, 2009).

Model Fit Statistics

Model fit statistics vary according to the model estimation methods used. The absolute fit statistics determine how well the model fits the actual data, whereas the relative fit statistics compare the fits between competing models. The absolute fit statistics include the residuals between the observed and predicted proportion of correct individual items, between the observed and predicted Fisher-transformed correlation of item pairs (referred to as transformed correlation), and between the observed and predicted log-odds ratios of item pairs. Computation of both the statistics and the standard errors for each statistics is required to perform the test, which determines whether the residuals differ significantly from zero (Chen, de la Torre, & Zhang, 2013).

Similarly, the posterior predictive model (PPM) can be used to compare model-predicted data with observed data statistics when the Bayesian approach is used to estimate model parameters (Roussos et al., 2010). The mean absolute difference (MAD) between predicted and observed statistics should be small to confirm that the model fits well. A number of studies have successfully used PPM (e.g., Jang, 2008; Templin & Henson, 2006).

Relative fit statistics which include –2 log-likelihood (–2LL), akaike information criterion (AIC), and Bayesian information criterion (BIC), are computed as a function of the maximum likelihood (ML) (Chen et al., 2013), which is similar to methods using log-likelihood fit statistics that compare the fits of competing models (Roussos et al., 2010). Von Davier (2005) used these statistics by comparing the fit of the compensatory MCLCM with that of the unidimensional two-parameter logistic model in the analysis of TOEFL data.

The Wald Test is also used in CDM to determine the item level fit statistics for the purpose of item-by-item model comparison. Specifically, it is used to statistically compare the fit between the saturated (G-DINA) and reduced (i.e., DINA, DINO, Additive CDM) models in the G-DINA model context in terms of *the Type I error and power of the test* (see de la Torre, 2011; de la Torre & Lee, 2013, for details). Then, the Wald test allows researchers to determine the best fit model for each item, rather than the best fit model for the whole test.

Validation

Validation is the process of gathering evidence to evaluate validity claim(s). While, evidence for the validity of any inference is a very important aspect of CDMs, Roussos et al. (2010) stated that validity research showing the statistical evidence of diagnostic instrument has not progressed well. Two types of validity evidence need to be considered: internal and external. Internal validity evidence is tested by using data from the test itself to evaluate the validity of mastery classification, whereas external evidence for validity is obtained using external data (Roussos et al., 2010) such as scores from other, related tests.

One key step for gathering internal validity evidence is specification testing of the Q-matrix. Several studies investigated this issue (Rupp & Templin, 2008; DeCarlo, 2011; de la Torre, & Zhang, 2013). Rupp and Templin simulated data with different misspecifications such as underfitting the Q-matrix (i.e., specifying 0s where there should be 1s), overfitting the

Q-matrix (i.e., specifying 1s where there should be 0s), and providing a balanced misfit for the Q-matrix (i.e., exchanging 0s and 1s while controlling for the overall number of changes), to examine the effects of Q-matrix misspecifications on the slip and guessing parameter estimates and misclassification rates. They found that misspecification of a particular item overestimated the parameters of that item.

DeCarlo (2011) used the posterior mode estimation (PME) with the DINA model in the fraction subtraction data. The results revealed some classification problems such as examinees who have zero total scores being classified as mastering most skills. He concluded that the Q-matrix misspecification can heavily affect the classification accuracy.

In addition to considering Q-matrix misspecification, IMstats and EMstats (Hartz & Roussos, 2008) can be computed for items and examinees, and used for evaluating the evidence for internal validity of the RUM model. To illustrate the process for IMstats, examinees would be classified into two groups: item master, indicating the examinee possess all required skills for the item, and item nonmaster if otherwise. Next, for each item, IMstats compare the proportions of examinees answering the item correctly across the two groups. If the performance is similar, then the researcher should revisit the item, the Q matrix and the skills coding of that item. For EMstats, a similar procedure is followed, but for examinees instead of items.

Examining differential item functioning (DIF) is another way to investigate the validity of score interpretations. In CDM, "an item exhibits DIF in the context of CDMs if the probabilities of success on the item are different for examinees who have the same attribute mastery profile but are from different groups" (Hou, de la Torre, & Nandakumar, 2014). Hou et al. used the Wald test to detect both uniform and nonuniform DIF in the DINA model and found that Wald test performance was comparable to or outperformed the Mantel-Haenszel (MH) and SIBTEST.

In addition, Ayers, Rabe-Hesketh, and Nugent (2013) performed MH tests using the predicted skill set profiles as matching criterion from the logistic and latent regression and standard DINA model. They found some DIF items with different models. Thus, DIF analysis can be performed with allowing the differences of skill mastery probabilities between groups.

With respect to external evidence for validity, one method for obtaining evidence is to use comparable test data with the same model to compare the results. If the classification results (i.e., percentage of the classification) are consistent between two assessment tests, this provides evidence for the validity claim of the test. For example, Templin and Henson (2006) compared the classification results of the DINO model using MCMC between two psychological tests: South Oak Gambling Screen (SOGS; developed by Lesieur & Blume, 1987) and Gambling Research Instrument (GRI; developed by Feasel, Henson, & Jones, 2004). After classifying the pathological gamblers, the results found that 89.2% of the classifications resulting from those two tests were consistent, supporting the validity claim of the instrument.

Related to these steps, an integrative framework by which one may carry out CDM from start to finish, including examples of both general and specific models, is presented in the next section.

An Integrative Framework

In this section, we propose an integrative framework for implementing CDM in practice. In this framework, we specifically highlight several theoretical considerations.

CDM as a Theory-Driven Model

CDM is a substantively-grounded model by nature (DiBello & Stout, 2003). It requires the relevant learning/cognitive theory as a basis for guiding instrument development and score interpretation and use. Hence, applying the theory-driven approach for CDM assists building of a validity argument or claim for enhancing the formative value of the assessment

information (Kane, 2006). Considering CDM as a theory-driven model includes two parts: assessment purpose and cognitive model.

Assessment purpose should determine the score interpretation and use. CDM can be used for multiple purposes, e.g., to validate the construct definition, to diagnose student knowledge and skills, and to validate the attributes of items. With the selected assessment purpose, researchers can construct or select assessment instruments that reflect the cognitive theory so that support the score interpretation and use. For example, if one wants to diagnose student knowledge and skills, the score interpretation should focus on the interpretation of student mastery level, rather than the assessment psychometrics properties.

In terms of the assessment items used in the CDM, one can decide whether one wants to (1) choose existing test data that matches the cognitive model of interest, or (2) refine the test items to make them more sensitive to the chosen cognitive model (Roussos et al., 2010). However, it is important for the CDM that the items should be highly sensitive to the cognitive construct so that the linkage between the cognitive construct and CDM is well-matched. Choosing from available items may be limited in terms of matching the construct of interest. Therefore, designing and developing good test items is the most appropriate approach. In this case, Henson and Douglas (2005) showed that using a heuristic, which is a simple algorithm used for test construction, in developing the item bank based on the CDM Information Index (CDI) results in higher classification accuracy rates than randomly constructed tests in both DINA and RUM models.

After deciding the assessment purpose, the attribute space can be defined and guided by the cognitive model. Cognitive models are "theoretical maps of how people learn and organize content knowledge" (Broaddus & Shaftel, 2012). There are several issues related to attribute definition: model and inferences, availability of the cognitive model, and numbers of attributes.

First, it is important to know that each cognitive model has its own capacity to make inferences about examinees' cognitive strengths and weaknesses. Leighton and Gierl (2007) examined three cognitive models in educational measurement: Test specifications/large-scale assessment model (LSM), domain mastery/curriculum-based assessment model (DMM), and task performance/cognitive diagnostic model (TPM). The researchers looked at each model's capacity to (a) inform test item development for measuring cognitive processes of interest, and (b) support the construct validity of inferences made about students' cognitive processing. Of the three models, the TPM was found to be most appropriate for the CDM because its goal is to assess cognitive strengths and weaknesses on the cognitive domain with the capacity to provide high psychological evidence, although, it can measure low range of skills. In contrast, although DMM can cover a large range of content, it mostly focuses on behavior strengths and weaknesses, whereas the LSM focuses only overall mastery of a large collection of knowledge and skills.

Secondly, attribute definition requires availability of a cognitive model. One challenge is the limitations of existing cognitive and learning theories, especially the cognitive models based on information processing theory, a topic of current interest in some disciplines such as problem solving in math and science. However, there are ways to develop a cognitive model or process, including verbal reports and protocol studies in which respondents reveal their mental/cognitive processes by thinking aloud while performing the assessment tasks, eye-tracking research that utilizes appropriate equipment to track the respondents' eye movement as the evidence to infer about their cognitive processes, and expert panels to invite domain experts to describe the cognitive processes behind item responses (de la Torre, 2014; Li, & Suen, 2013; Rupp, et al., 2010). Using evidence from these types of studies, the cognitive model can be used for the development of attribute lists and items. Such work must be well-established in order to get productive diagnostic results (Roussos, Templin, & Henson, 2007).

Finally, number of attributes can impact the attribute definition. As mentioned earlier, the attribute definition process involves the specification and definition of the attributes. The number of attributes then somewhat influences the level of specificity that researchers may choose to delineate and to statistically estimate the attributes involved in the items. In other words, researchers need to justify their decisions of defining attributes at a finer-grained size or collapsing attributes at a broader sense, given practical considerations (e.g., computer program capacity, testing time). To illustrate, researchers might choose to consider a small number of attributes to measure a specific construct, such as proportional reasoning. In contrast, they are interested in assessing student learning of broad content, they might decide to define attributes in a much broader sense.

Statistical Procedures

In this section, we describe the four critical statistical procedures in CDM: model selection, model estimation, model parameters interpretation, and Q-matrix validation.

Choosing the Right Model

It is important to run CDMs with a model that matches the attribute definition, particularly whether the attributes can be compensated or not. For example, item 2+3-1=?, which is shown in Rupp et al. (2008), requires both addition and subtraction skills. If one of the required attributes is lacking, it is impossible to answer this item correctly. Therefore, a model with a noncompensatory rule, such as DINA as opposed to DINO, should be used.

Several other factors should to be considered when selecting the model. For detailed review, see de la Torre, Hong, and Deng (2010) and Rojas, de la Torre, and Olea (2012). The first factor is whether the true model is known or not (always not known in practice; may be reasonably surmised). When the true model is known, using that model for estimation would be the suitable solution. But if the true model is unknown, G-DINA is a

good choice since it provided the best attribute classification accuracy (ACA) compared to other models.

The second consideration is sample size. De la Torre et al. (2010) studied factors affecting the item parameter estimation and classification accuracy of the DINA model. They found that a sample size of 1,000 is sufficient to provide accurate (in terms of bias) parameter estimates, but the precision (in terms of SD) is improved as the sample size increased. Rojas et al. (2012) found that G-DINA is quite good (in terms of classification rates), compared to DINA, DINO, and A-CDM, when the sample size is small (e.g., 100 or 200), although this does not hold true in all scenarios. For example, the DINO model can provide a proportion of correctly classified individual attributes given the scenario that the true model is DINO, item quality is low, and sample size is small. Moreover, a sample size of 400 can provide relatively good results, compared to a larger sample size. So overall, the increase of the sample size reduces the variation or bias in item parameter estimation.

Model Estimation

As described in an earlier section, it is important to know what kind of model estimation is appropriate for the model. For example, von Davier & Yamamoto (2004) used EM instead of MCMC because they said that EM works well for their models of interest, has much lower computational cost, can apply the IRT toolbox of methods (e.g., weighting, restrictions, and fitting), and can reuse available script languages and libraries efficiently. This is empirically confirmed by Feng (2013) who compared these two model estimation methods with data from the Examination for the Certificate for Proficiency in English (ECPE), a test developed by the University of Michigan, using the rRUM. The results found that the EM algorithm presents accurate estimates, with significantly lower computational time compared to estimation by MCMC.

In practice, model estimation is embedded in the program mechanism. To run most programs, researchers need to provide the Q-matrix as an input of the program along with the dataset in a format that is readable by the program. They also need to write the syntax. However, some programs provide easily modified usable codes. For example, the codes for DINA and G-DINA in Ox require only changing the numbers of the sample size, attributes, and items, and also the name of the Q-matrix and response data files (in space- or tab-delimited format). Given these inputs, Ox can run the model estimation.

Moreover, different estimation methods require different software for analyses. While some models are used with a variety of software (e.g., the DINA model can be analyzed by Ox console, and CDM in R), some models need specific software and require research licenses (e.g., Full NC-RUM and Reduced NC-RUM in the Arpeggio program). Hence, researchers need to consider the model, the availability of necessary software, and in some cases, access to software expertise.

Interpretation of Model Parameters

After running the model and getting the results, all model parameters should be evaluated and interpreted. As stated by Roussos et al. (2010, p. 44), "The estimates for the ability distribution and item parameters should be evaluated for internal consistency, reasonability, and concurrence with substantive expectations." Basically, the goal of CDM is to make inferences about the attribute vector α_i or attribute pattern (de la Torre, 2014). Thus, student mastery should be reported. If a skill was much harder or easier than the expected, based on the hypothesized cognitive model or the targeted content, the researcher may need to reconsider the Q-matrix, the item difficulty, or even the attribute definition.

To illustrate, the outputs of DINA and G-DINA include estimates of item parameters with the corresponding standard errors, posterior distribution of the attributes, examinee attribute classification, and item- and test-level fit statistics. The estimates of item

parameters with the corresponding standard errors in DINA and G-DINA are different and require different interpretations. The interpretation of the DINA model is straightforward because there are only two parameter estimates for each item: the guessing and slip parameters, which stand in for difficulty (1-slip) and for discrimination (1-guessing-slip).

The numbers of parameter estimates for each item in the G-DINA model depends on the attributes required for each item. Therefore, it requires interpretation of each attribute combination as well, and this information demonstrates the probability of success associated with each attribute pattern. The posterior distribution of the attributes from G-DINA shows the estimated proportion of each attribute pattern in the population. The examinee attribute classification informs the profile of individual examinees, specifically, which attributes he/she mastered. From this output, we can know which and how many examinees are in the same group or have the same attribute pattern, and which attribute pattern should be provided for remediation or improvement. The item- and test-level fit statistics provide the absolute and relative fit statistics for model fit evaluation. The item-level fit statistics show the expected values of the item statistics (i.e., proportion correct, correlation, and log-odds ratio). These absolute fit statistics indicate how well the model fits the data; if it fits well, the fit statistics should be small (de la Torre, 2014). The test-level fit statistics, -2LL, AIC, and BIC, are used for the relative fit evaluation, which is important for comparing models and choosing the best fitting model. Again, the smaller values of the absolute fit statistics, the better the fit of the model.

Q-matrix Validation

Various validation methods can be used in CDM such as DIF analysis and comparison between pretest and posttest scores, as mentioned earlier in the CDM steps section. In this section, we focus on Q-matrix validation. The Q-matrix is normally developed based on expert judgment, which can be subjective and involve possible misspecifications.

Misspecifications of the Q-matrix impact the parameter estimation and classification accuracy (de la Torre, 2014). For example, if attributes are mistakenly omitted from the Q-matrix, it impacts the estimated slipping parameter (Rupp & Templin, 2008). Thus, it is very important to validate the Q-matrix.

Several methods have been developed for validating the Q-matrix. One is the sequential EM-based δ-method for Q-matrix validation developed by de la Torre (2008) for selecting the optimal q vectors to improve model-data fit. This method is based solely on the available data. In the DINA model, this method can both identify and correct miss-specified q vectors, and retain those that were appropriately specified simultaneously. The other method, index ς^2 (de la Torre & Chiu, 2015) can be used without assuming the specific CDMs involved, only that they are subsumed by the G-DINA model, which is quite general. So $\varsigma^2$is a generalization of the discrimination index specifically for the DINA model proposed by de la Torre (2008) and is used to give the model more applicability and generality.

Another way of Q-matrix validation is the use of statistical models such as multiple regression and the linear logistic test model (LSDM; Dimitrov, 2007). A higher variance explained of item difficulty in multiple regression is an indicator of a valid Q-matrix (Tatsuoka, 2009). Under the LSDM, a small least squares distances (LSDs), small mean absolute difference (MAD), and meaningful attribute probability curves (APCs) suggest a correct Q-matrix (Dimitrov, 2007). One empirical study using these methods is Ma (2014) who validated the proposed Q-matrices of TIMSS–Mathematics and found that the proposed Q-matrices were acceptable, albeit with a few unreasonable APCs.

Q-matrix validation should not be solely based on statistical methods; Q-matrix revision with respect to the underlying cognitive model is also critical. Tatsuoka (2009) pointed out, “it is dangerous to rely on a single optimization technique for attribute selection” (p. 273). Consideration of the underlying cognitive model or assessment theory can be

employed to determine if the results from the proposed Q-matrix are consistent with the theoretical cognitive model. For example, Li and Suen (2013) refined the Q-matrix after getting the results of RUM model by taking a closer look at the item by using their substantive knowledge and think-aloud verbal reports, and reanalyzed the model with the refined Q-matrix. They concluded that "it is important that Q-matrix modification decisions based 17 on statistical modeling receive substantive support."

In sum, CDM is a model that can be used to estimate examinees' latent cognitive constructs to provide individual students' learning profiles, guide teachers' instruction, and provide the items' psychometric properties. Various models have different characteristics, strengths, and weaknesses. In short, the DINA model is parsimonious and very restrictive, whereas the G-DINA is the most flexible but involves more parameters, and requires a longer estimation time and a larger sample size.

Generally, there are related multiple steps involved in CDM such as attribute definition, Q-matrix validation, model fitting, and model selection. While other steps are highly statistics-based, the attribute definition is based on conceptual work that requires to deliberately clarify the assessment purposes and specify the analysis framework in order to guide the analysis. Also, the output interpretation is a reflective work that should relate back to the attribute definition. Specifically, the interpretation should determine whether the outputs make sense or not based on the definition framework.

The learning/cognitive sciences should guide the attribute definitions. Although CDM is very useful for assessment data analysis, it is more beneficial if the assessment design is based on a learning/cognitive sciences theory. A well-defined construct and items based on available cognitive and learning theory are needed to guide attribute definition. Thus, the researcher needs to consider several things when implementing CDM, specifically, cognitive

and learning theory and statistical procedures. With those considerations, CDM can yield highly useful benefits for diagnostic purposes.

SECTION 2

A COGNITIVE DIAGNOSTIC MODELING APPROACH TO INSTRUCTIONAL SENSITIVITY

Min Li[1], Phonraphee Thummaphan[1], & Maria Araceli Ruiz-Primo[2]

[1]Univeristy of Washington at Seattle, [2]University of Colorado Denver

Abstract

Assessments have been pervasively used in classrooms by educators and researchers for various purposes from studying the effects of educational innovations to examining the instructional outcomes in a classroom. Scholars have recurrently brought up a concern about the interpretation and use of assessment scores: Are assessments sensitive enough to detect student learning differences due to instruction? If so, do they have formative value for teachers and students? In this paper we examined the formative value of instructional sensitivity of assessment items from two elementary science modules. In order to determine whether items varying in instructional sensitivity yield different formative values for diagnosing student learning we created booklets with items of different instructional proximity (from *close* to *far proximal*), and administered them in 38 classrooms (824 students) using a pretest-posttest design. Incorporated the cognitive diagnostic modeling analysis, the item and test indices show that the data fit well the specified Q-matrix. The attributes with the higher gain have been heavily addressed in the intended curriculum (i.e., a greater amount of learning activities) compared to those with relatively smaller gain.

Assessments have been pervasively used in classrooms by educators and researchers for various purposes, ranging from accountability of teacher effectiveness and formative use to monitor where students are in relation to the learning goals, to detection and study of educational innovations. A concern has been repeatedly brought up by scholars (see Airasian & Madaus, 1983; Madaus, Airasian, & Kellaghan, 1980; Ruiz-Primo, Shavelson, Hamilton,

& Klein, 2002) about the interpretation and use of assessment scores: Are assessments sensitive enough to detect student achievement differences?

If we assume that one of the important purposes of assessments is to make inferences about the variables of the learning environment based on the status and progress of student achievement, then one essential piece of validating assessments must address instructional sensitivity. Assessments lacking instructional sensitivity cannot adequately monitor the instructional quality students receive or evaluate the effectiveness of educational reforms.

Over decades, researchers have proposed different methods to examine the instructional sensitivity, mostly in the form of: (1) judgmental evidence with teachers, students, or experts on their self-reported ratings of sensitivity, or (2) empirical evidence with student test scores to link the item statistics to indicators of instruction implemented (e.g., topics covered, instructional time allocated, or quality of instruction). Unlike prior research that involves approaches of classical test theory and/or item response theory, this paper aims to incorporate the *cognitive diagnostic modeling* (CDM) approach to investigate the instructionally sensitivity of assessments, using two elementary science modules. We ask: *how is the instructional sensitivity of the items related to their cognitive diagnostic estimates?* Both quantitative and qualitative evidence is provided to evaluate the technical properties of items under the CDM framework and examine the validity claim that items with greater instructional sensitivity offer more diagnostic value for student learning. We hypothesize that the items close to the actual classroom instruction provide more accurate values of CDM estimates and attribute mastery levels compared to the ones less aligned to the classroom instruction.

Introduction

Defining Instructional Sensitivity

We (Ruiz-Primo & Li, 2012) adapt and extend the definition of instructional assessment as referring to the extent to which the assessments or assessment items: (1) represent the enacted curriculum (the material covered by the test has actually been taught), and (2) reflect the quality of the enacted curriculum (the quality of instruction) (also see Ruiz-Primo, 2015, for the rationale of connecting the instructional sensitivity to the three types of curriculum).

Consistent with this definition, we conceptualize the instructional sensitivity as a continuum from more sensitive to less sensitive: assessments then can appear as immediate, close, proximal, and distal (Ruiz-Primo & Li, 2012). At the *immediate level*, assessments are artifacts from the enacted curriculum such as notebooks, worksheets, or tasks for group work, which are often the learning tasks planned by teachers or suggested by the curriculum that can provide assessment information about student achievement. At a *close level*, assessments are similar to the content and activities as described in the curriculum and implemented in the classroom and, therefore, are more curriculum-sensitive. At a *proximal level*, assessments consider the knowledge and skills relevant to the curriculum, but context (e.g., scenarios) differs from what is studied in the module. At a *distal level*, assessments are based on state or national standards in a particular domain, and thus are highly probable to be only minimally related to what students learned in the module.

DEISA Approach to Developing Instructionally Sensitive Assessments

This paper used tests and data collected from the parent project, *Developing and Evaluating Instructionally Sensitive Assessment (DEISA)*. The DEISA approach to developing instructionally sensitive assessments and evaluating their technical quality is guided by three theoretical underpinnings: transfer of learning, big ideas, and type of

knowledge (for a detailed review, see Ruiz-Primo & Li, 2012). These foundations are used to define the construct, operationalize instructional sensitivity, generate the item prototypes, and derive the interpretative and validity arguments. We have experimented and refined this assessment approach over four tryouts with five different modules. The purpose of including different modules from different school districts is to test whether the approach is robust given the variation of curriculum and teacher variables.

We see close assessments as those that help to measure whether the original targeted learning took place or not, the proximal items as the means to measure how much of what was learned transfers to new contexts, and the distal assessments to capture the learning in the domain of science in general that to some degree is accumulated over many modules (Briddle, 1997; Darling-Hammond, 1999, 2007; Linn & Harnisch, 1981; Popham, 2007a, 2007b; Toutkoushian & Curtis, 2005; White, 1982). Therefore, we reason that the gain of student performance from learning a particular module, if well taught, should be the largest in the close items and decline as distance increases. In order to validate the DEISA assessments, "at a minimum, we want some evidence that students' performance on these measures improves with teaching" (Baker, 1994, p. 60).

In this paper, we employ the CDM approach to analyze and compare students' test scores on the assessment items at different proximities to determine whether this expected pattern on the change of student performance is observed.

Statistical Procedures to Evaluate Instructional Sensitivity

A range of statistical approaches has been used to measure the instructional sensitivity of assessments or assessment items (for a thorough review of additional approaches, see Polikoff, 2010). One common method is pre-to-post difference index (PPDI) introduced by Cox and Vargas in 1966. This method works with raw scores by comparing the performance of a taught group with an untaught group (Haladyna & Roid, 1981). Here the taught group

and untaught group can be two different samples or can be one sample which take the test before the instruction and after the instruction. In the latter case, PPDI is the simple difference in item difficulty (percentage of correct responses) between the pretest and posttest of the same sample. PPDI is suggested by Polikoff (2010) in his recent review, as a robust, sound, and easily computed statistic of instructional sensitivity.

Item Response Theory (IRT) modeling is another popular method used to examine the instructional sensitivity. For instance, Muthén (1988) and Muthén, Kao, and Burnstein (1988) extended the IRT modeling to estimate the difficulty parameter for items by allowing it to vary with the level of instructional coverage and/or the level of instructional quality. Two recent studies by Albano and Rodriguez (2013) and Naumann, Hochweber, and Hartig (2014) demonstrated the uses of two differential item functioning (DIF) approaches. Albano and Rodriguez (2013) used hierarchical generalized linear modeling to investigate how the mathematics DIF items with respect to gender could be accounted for by opportunity to learn, a student self-report measure of whether or not they had studied the topics. Naumann and his colleagues (2014) employed a longitudinal multilevel-DIF model to estimate the change in classroom-level item difficulties between two time points of testing as the item statistics for instructional sensitivity.

Furthermore, statistical procedures to estimate the variance components have been used in the instructional sensitivity research as well. One such example is Jarjoura and Brennan's 1982 study in which Generalizability theory was used to calculate the estimated variance components. The authors included item and category facets in their model in order to examine variance components of a weighted composite of category universe scores relative to the category weights assigned from the table of content specifications. Other studies that can be loosely included in this method camp are those in which researchers studied percentage of variance of student scores that can be explained by instructional variables using

either analysis of variance (e.g., Brutten, Mouw, & Perkins, 1992) or hierarchical linear modeling (e.g., D'Agostino, Davis, & Megan, 2007; Ing, 2008; Welsh, 2009).

Cognitive diagnostic modeling (CDM) has received little attention as a statistical procedure to study instructional sensitivity. Educators have recognized the importance of formatively using assessment results as information for enhancing student learning and as means for communicating with students, teachers, and parents (Pellegrino, 2013; Popham et al., 2005). CDM has been considered as an effective technical solution to meet these needs. CDM, is a psychometric modeling approach used to measure students' cognitive skills or knowledge required to answer items correctly. "Cognitive diagnosis models (CDMs) are latent variable models developed primarily for assessing student mastery and non-mastery on a set of finer-grained skills" (de la Torre, 2011, p. 179). What remains unclear is how item parameters generated from CDM can provide empirical evidence to validate instructional sensitivity claims.

Cognitive Diagnostic Modeling Approach

Research on formative assessment points out that the key of formative assessment lies in the use of assessment information to alter learning and teaching processes: "…Only by keeping a very close eye on emerging learning through formative assessment can teachers be prospective, determining what is within the students' reach, and providing them experiences to support and extend learning through which students can then incorporate new learning into their developing schema" (Heritage, 2010, p. 8). CDM produces detailed diagnostic information from student responses on assessments that can be readily available for teachers, students, and parents.

Hartz (2002) stated that the goal of doing CDM is to measure student mastery on their required skills or underlying competencies and to evaluate the item diagnostic power of measuring required skills/competencies. By expressing learning outcomes as a set of basic

skills or attributes required to respond in an assessment, CDM can estimate a profile of the skills an individual has mastered. The main advantage of CDM is to obtain a detailed profile on student understanding that provides strengths and weaknesses of individual students rather than an overall score. Thus, assessment results can be used for diagnosing and remediating to improve student learning (Pellegrino, 2013).

There are a number of common addressed CDMs such as *deterministic inputs, noisy and-gate* and *noisy-input, deterministic and-gate*(DINA and NIDA; e.g., Junker & Sijtsma, 2001), *deterministic inputs, noisy 'or' gate* and *noisy-input, deterministic or-gate*(DINO and NIDO; e.g., Templin & Henson, 2006), the *reparametrized unified model* (RUM) / *fusion* model (e.g., DiBello et al., 1995; Hartz, 2002), and *Generalized DINA* (G-DINA) (de la Torre, 2011). Different CDMs have different characteristics, statistical assumptions, and procedures (Rupp & Templin, 2008). For example, DINA model is the model that involves dichotomous variables for both latent predictor and observed response, and has the assumption of non-compensatory requiring that all measured skills are presented to answer item correctly. [1]

A CDM analysis starts with the attribute definitions. It is the process of determining specifications and definitions of the skill(s) of interest. Then, the assessment items are analyzed and coded based on the attribute definitions to produce the Q-matrix – an item-by-attribute binary matrix that identifies which attributes are required to answer each item correctly. After producing Q-matrix, item and person ability parameters can be estimated based on selected models for the data analysis. Roussos et al. (2010) recommend marginal maximum likelihood (MML) estimation as it is a useful algorithm for parameter estimation from data that provide only imperfect measurement of the construct of interest. Finally, the

[1] The "difference between *compensatory* and *noncompensatory* models reflects how the latent predictor variables are combined across the different skills to produce the observed responses" (Rupp & Templin, 2008). Non-compensatory model assumes that a low value or lack of one skill cannot be compensated for by a high value on another skill latent variable.

results will be interpreted to decide whether or not they are reasonable. If the results do not make sense, the attribute definitions and Q-matrix should be re-considered and revised from which another run of data analysis and interpretation need to be carried out.

We claim that the CDM is an appropriate psychometric procedure to directly test whether instructionally sensitive assessments have any diagnostic values to the instruction. By offering assessment information at a finer-grain level for student learning, CDM results allow researcher to examine the extent to which patterns of student performance can be directly tied to the instruction they have received and/or the intended curriculum only, especially if the goals and outcomes of instruction can be specified into and aligned with the attribute definitions. With the same reasoning, if items are sensitive to instruction, CDM analysis can produce the assessment reports with more elaborated information about students' performance so that teachers and/or students are able to use such diagnostic information to adjust their instructional and learning processes. This is exactly the benefits of performing any CDM analysis. Thus, CDM offers a psychometric procedure to substantiate the possible link between student mastery of hypothesized attributes and items with varying proximity. In this paper we report an empirical study in which students responses were analyzed with the CDM approach using data from two science modules at grade 5. In what follows, we describe the research methods and then report the findings of the CDM analysis to investigate how the instructional sensitivity of the items is associated with their cognitive diagnostic estimates.

Methods

The methods of this paper followed the steps of CDM analysis: (1) attribute definition to guide the assessment instrument and Q-matrix construction, (2) item development to use for data collection, (3) Q-matrix development to use for analysis, (4) model fitting to estimate the parameters, and (5) validation to verify the results.

Attribute definition

Attributes used to define skills/understanding needed to respond a particular item in this study were developed based on the big ideas of each module identified by the research team and teachers, using the curriculum mapping procedure (Ruiz-Primo et al., 2011). To determine the ultimate set of attributes or skills required, researchers reviewed the list of big ideas and evaluated item specifications of how items tapped into the big ideas generated by item writers, and finalized the attribute definitions (see Table 1 for the attribute definitions).

Table 1. *Attribute Definitions for the Items of the Two Science Modules* [a]

Label of Attribute	Explanation	# of Times Assessed by Items
LF Module		
Processes of erosion and deposition	Understanding the processes and outcomes of erosion and depositions	6
Factors that influence erosion and deposition	Understanding how factors of water amount and velocity influences the processes of erosion and depositions	5
Reading/interpreting maps	Making sense of the information presented in different kinds of maps, such as contour lines	3
MS Module		
Definitions of solution/mixture	Defining the terms of solution and mixture and providing examples for solution or mixture	2
Making solutions/ separation procedures	Applying the procedures to making solutions or separate mixtures	7
Properties of matter in solutions	Understanding the properties of substance in a solution maintain the same	3
Reading images in an item	Reading and decoding the graphic information presented	2

Note a. Attributes that corresponded to the unique items but not the common items were omitted from this table. These attributes include two LF attributes (i.e., types of landforms, modeling) and two MS attributes (i.e., chemical/physical changes, concentration).

Item development

The test materials were situated in a larger study, the DEISA project (Ruiz-Primo & Li, 2008). This study included information from two science modules: the *Landforms* (LF) and *Mixtures and Solutions* (MS) modules published by Full Option Science System (FOSS, 2005 edition). The LF module introduces students to the Earth Science concept of landforms changing over time, largely through stream table investigations of erosion and deposition. The module also has a heavy focus on modeling through topographic maps, aerial

photographs, and three-dimensional models, all through interaction with those materials. The big ideas identified for this module are: (1) the Earth has a wide variety of landforms that have different characteristics; (2) erosion and deposition are processes that shape and reshape the Earth's land surface, and water is the predominant agent of change; and (3) a model can represent structures, processes, and mechanisms at a manageable scale. The module consists of six lessons, including three with experiments and the other three requiring students to manipulate different models. In all the six lessons, the guiding questions for drawing conclusions are closely tied back to the big ideas of the module.

In the MS module, students work with various substances in four investigations to study the properties of mixtures and solutions, experiment with the procedures of mixing and separating substances, explore the saturation and concentration of a solution, and learn about the evidence of chemical reaction, evaporation, and crystal formation. The big ideas of this module include: (1) mixtures are combinations of substances (liquids, gases, solids) that can be physically separated into component parts; (2) solubility is a property of matter; and (3) some mixtures result in chemical reactions that transform the component parts (reactants) and create new products.

For each of the LF and MS modules, we developed close, near proximal, and far proximal items based on the DEISA approach.[2] MC items were used as it is less costly in item development and scoring. We constructed the DEISA items in bundles, starting with the close item that most resembled what students were asked to do in their investigations. We then increased the amount of manipulation of item characteristics, such as familiarity of objects or organisms, familiarity of setting, cognitive demands (see Li et al., 2012, for a detailed description of the item manipulation and empirical evidence for viability of this item

[2]We selected distal items from released large-scale tests of NAEP and state assessments. Although distal items were included in the booklets, they were excluded in this paper as they tended to assess underlying understanding only loosely related to the LF big ideas.

development approach). Figure 1 presents example items from the LF module—that is, *one bundle* of close, near proximal, and far proximal items, all of which assess student understanding of the deposition process.

Close item

Sue's stream table erodes more earth material than Tom's.

What is probably true?

A. Sue's stream channel is less deep than Tom's.
B. Sue's stream channel is narrower than Tom's.
C. Sue's delta is smaller than Tom's.
D. Sue's delta is larger than Tom's.

Near Proximal Item

The average speed of water in Cherry Creek was faster in 2008 than in 2010. Because of this, the creek eroded more material in 2008.

What else was probably true in 2008?

A. The creek deposited less material downstream.
B. The creek deposited more material downstream.
C. The creek channel was less steep than in 2010.
D. The creek level was lower than in 2010.

Far Proximal Item

One form of erosion takes place when rocks break off from a cliff.

What is the deposition of these rocks?

A. They land at the base of the cliff.
B. They get picked up by a river.
C. They break off other rocks when they hit them.
D. They are not deposited.

Figure 1. *Examples of DEISA Items for the LF Module*

Within each module, we organized the items into two assessment booklets, ensuring that each booklet had items varying in proximity from close to far-proximal. Because of the way the booklet was designed, we considered items included in just one of the two booklets

as unique items; items appearing in both booklets at the exact same locations were labeled common items (or anchor items). Using a standardized testing procedure the booklets were administered to students before and after instruction in the target module.

In this study, we used only the common items in the data analysis to ensure that the sample size of students per item would be sufficient for the proposed statistical methods; otherwise, inclusion of unique items would reduce the student sample by half. This decision resulted in 9 LF items of three complete bundles, and 8 MS items of two complete bundles as well as one partial bundle.

Q-matrix

Two researchers refined the item specifications and created the Q-matrix (see Table 2 for the Q-matrix). After a thorough review of the items and Q-matrix by two content experts, we judged that there were no misspecifications of attributes in the Q-matrix.

Model fitting

Participants

The sample for this study included 38 fifth-grade teachers and 824 students recruited from three school districts in a Midwest state (427 students for the LF module; 397 students for the MS module). Table 5 provides information about the school districts. School Districts 2 was a relatively large district with a higher percentage of diverse ethnicity compared to the other school districts. Student performance on the state mandatory assessments in District 2 was at or slightly below the state average whereas student performance in Districts 1 and 3 was above the state average.

Table 2a. *Q-matrix for the LF Module*

Item ID	Processes of Erosion and Deposition	Factors that Influence Erosion and Deposition	Reading/interpreting Maps
LF2	1	1	0
LF3	1	1	0
LF5	0	0	1
LF10	0	0	1
LF13	1	1	0
LF14	1	1	0
LF17	0	0	1
LF20	1	0	0
LF21	1	1	0

Table 2b. *Q-matrix for the MS Module*

Item ID	Definitions of Solution/Mixture	Making Solutions /Separation Procedures	Properties of Matter in Solutions	Reading Images in an Item
MS2	0	1	1	0
MS4	0	1	0	0
MS6	0	1	0	0
MS9	0	0	1	0
MS11	1	1	0	1
MS16	0	1	1	1
MS18	0	1	0	0
MS20	1	1	0	0

Note. 0s are used when an attribute is not required to answer an item correctly. 1 indicates the attribute is a needed understanding or skill for students to provide a correct response.

Table 3. *Information about the School Districts*

	School District 1	School District 2	School District 3
Location	Suburban	Suburban	Suburban and rural
Total enrollment (students of Pre-K to Grade 12)	29,160	42,990	15,667
Ethnicity (percentage)			
Caucasian	71.11	57.60	76.10
Hispanic	17.08	32.54	18.80
African American	0.99	2.45	0.90
American Indian	0.52	0.73	0.60
Asian/Pacific Islander	5.98	0.15	1.60
Two or more ethnic groups	4.32	6.53	2.10
Free/reduced lunch (percentage)	18.71	33.94	34.90
English Language Learners (percentage)	8.21	11.56	3.50
Module(s) used	LF and MS	LF and MS	MS

Data analysis

There were four runs of the CDM analyses: using both the pre-test and post-test item level scores for the two modules, based on the two specified Q-matrix tables. DINA (de la Torre, 2009) model seems appropriate for instructional sensitivity examination purpose since it is parsimonious and interpretable. It was used in Ox console (Doornik, 2002) in this analysis. The code estimated the item and person parameters of the DINA model using an expectation-maximization (EM) algorithm (de la Torre, 2011). That is, the outputs provided parameter estimates and standard errors of the parameters based on an identity link function, posterior distribution of the attributes, and examinee attribute classification. The original code provided by de la Torre, was modified by specifying the number of examinees, test length, number of attributes, the Q-matrix file names, and response data file names.

Validation

We used the item level scores to generate *pre-to-post difference index* (PPDI) estimates to index the pretest-posttest gain of student learning. PPDI is calculated by subtracting the percentage of correct responses in the pretest from that of the posttest. Therefore, it refers to the gain of student performance from the pretest to the posttest. A higher PPDI indicates a more sensitive item, assuming that pattern of student performance before and after instruction should correspond to the item sensitivity. Furthermore, we performed correlation analysis between CDM estimates and PPDI to examine whether item parameters from different approaches yield comparable findings.

Results

In this section we present the results of the CDM analysis across the two science modules. To test the robustness of the DEISA approach, we focus on the students' mastery of the specified attributes in the DEISA assessment and evaluate the item parameters by

considering proximity of the item (close, proximal, and distal). Patterns of student performance are interpreted by using the Q-matrix information and comparing against the item statistics based on the PPDIs to explain possible differences in the patterns of student responses.

Model Fit Indices

The item-level fit statistics as absolute fit statistics, are intended to evaluate the relation between items in the test and data (see Table 4). The values of those three statistics for both pre-test and post-test in both modules are quite small, close to zero. For example, Mean Absolute Deviation (Mean Abs. Dev.) of the residuals between the observed and predicted proportion of correct individual items (Prop) was 0.0150 for pre-test and 0.0229 for post-test in LF module; and it was 0.0166 for pre-test and 0.0188 for post-test in MS module. Therefore, we decided that the model, specified by the Q-matrix, fittted the data fairly well.

Table 4. *Item-level Fit Statistics*

	Pre-test			Post-test		
	Prop	Z(Corr)	Log(OR)	Prop	Z(Corr)	Log(OR)
LF						
Mean Abs. Dev.	0.0150	0.0373	0.1576	0.0229	0.0467	0.2027
Max. Abs Dev.	0.0310	0.1079	0.4361	0.0400	0.1042	0.4469
SE(Max. Abs Dev.)	0.0252	0.0505	0.2039	0.0244	0.0505	0.2610
MS						
Mean Abs. Dev.	0.0166	0.0265	0.1595	0.0188	0.0401	0.2319
Max. Abs Dev.	0.0654	0.0699	0.6001	0.0490	0.0965	0.7471
SE(Max. Abs Dev.)	0.0186	0.0524	0.5140	0.0263	0.0558	0.6152

Note. *Mean Abs. Dev.* is Mean Absolute Deviation; *Max. Abs Dev.* is Maximum Absolute Deviation; *SE(Max. Abs Dev.)* is Standard Error of Maximum Absolute Deviation; *Prop* is residuals between the observed and predicted proportion of correct individual items; *Z(Corr)* is residuals between the observed and predicted Fisher-transformed correlation of item pairs (referred to as transformed correlation); *Log(OR)* is residuals between the observed and predicted log-odds ratios of item pairs.

Attribute Mastery

Regarding the mastery level as shown in Figure 2, the magnitudes of the learning gain were mostly greater than 20% except for Attribute: Making solutions/Separation procedures. For the LF module, the two attributes gained the most were 32% as the highest gain for Attribute 2: Factors that influence erosion and deposition, and followed by Attribute 3: Reading/interpreting maps with 29%. These two big ideas were the main focuses of the LF module, each of which was addressed in 40% of the module activities, including teacher demonstration, student investigation, group discussion, and worksheets.

Although with the least gain, the students have mastered the most on Attribute 1: Processes of erosion and deposition for both pre-test (66%) and post-test (88%). This finding suggests that students tended to bring a lot of prior knowledge around this big idea from other related science modules in their previous learning. Therefore, the gain appears less impressive compared to the other two attributes.

Across the four attributes in the MS module, the two with the highest gains were Attribute 3: Properties of matter in solutions with 34%, and followed by Attribute 1: Definitions of solution/mixture with 26%. These two big ideas were covered in the first two lessons in the module, equivalent to 40% of the instructional activities. Attribute 2: Making solutions/ separation procedures seemed to be the easiest for students to master for both pre-test and post-test (85% and 91%, respectively). We speculate that students probably accumulated some naïve ideas about this big idea through their daily observations and experiences. Across the modules, the amount of instructional exposure based on the intended curriculum may help explain the patterns of learning gains of students, reflected by the mastery levels of the attributes in these two modules.

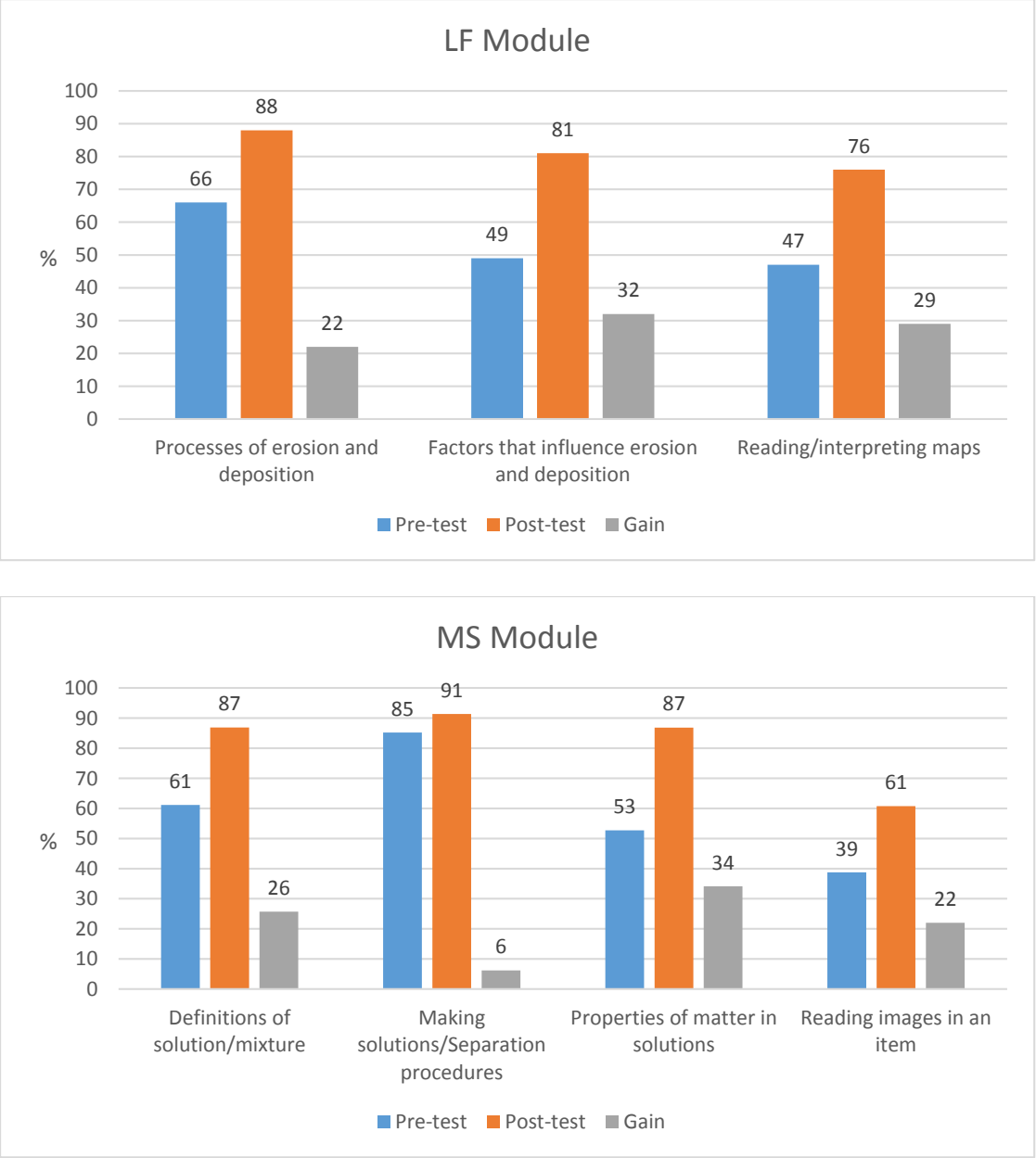


Figure 2. *Attribute Mastery Level by Attribute (in percentage)*

Item Parameters In Relation to Proximity

The DINA item parameters represent the guessing and the slip probabilities for each test item. Guessing refers to the probability of answering an item correctly without mastering the specified attribute(s) whereas slip refers to the probability of selecting a wrong answer even though fully understanding the underlying attribute(s). In addition, the guessing and the slip parameter estimates can be used to calculate the discrimination parameter which refers to

the difference in probabilities of correct responses between two groups of examinees (mastery vs. non-mastery of the specified attribute[s]). Statistically, the discrimination is operationalized as the probability of correctly solving an item without the effect of guessing and slip (i.e., $\delta = 1 - \text{guess}$

Table 5 presents the probabilities of guessing, slip, and discrimination parameters on the items. If the students possess all the required attributes, they are assumed to have higher probability to answer items correctly for more sensitive items, for both modules. Students seemed to have more chance to provide correct answers in post-test than pre-test since they had learned and mastered more of those required attributes (see Figure 2 for the mastery level). To illustrate this, for the LF module, if students master all measured attributes, they have over 80% probability to answer Items LF3, LF5, and LF14 for both pre-test and post-test because their slip parameters are smaller than 0.20. It seems that Items LF17 and LF20 which are far proximal items have much lower opportunity for students to arrive at correct answers in both pre-test and post-test, although examinees have all attributes. Interestingly, most guessing parameters are slightly high indicating that examinees have high likelihood to provide correct answer by guessing, except LF20 which has .019 and .141 guessing parameters in pre-test and post-test.

For the MS module, the slip parameters do not yield any patterns in relation to the item prolixity. If the examinees master all the measured attributes, they will have at least 95% probability to correctly answer Items MS2 and MS6 for the posttest. In contrast, although they possess all the attributes, they still have only less than 50% chance of answering Items MS4, MS9, MS16, and MS20 correctly. Considering the guessing parameters, most of them are also slightly high, except MS9 which has .019 and .079 guessing parameters in the pretest and posttest, respectively. It is interesting to observe that slip parameter of Item MS16 pre-test was 1.000, indicating that examinees who have all the required attributes end up selecting

the wrong answer. But in the posttest, the slip parameter was decreased to .552 which was more reasonable. We need to take a closer look to analyze the content of this item to check if there is a possible misspecification of the Q-matrix for this item.

Regarding item discrimination parameters, MS4 may require additional review because of the negative value (i.e., -.095). This is the only item which failed to discriminate in the post-test performance of students who mastered the specified attributes from those who mastered none of the attributes. This item requires only one attribute which is "Making Solutions /Separation Procedures". Since the guessing and slip values are slightly high (i.e., .578 and .518 respectively), we speculate that the item may involve more attributes that are not covered by this attribute definition.

We take a closer look at the item parameters by comparing them between the pretest and posttest (see Table 6), from the CDM analysis and the difficulty parameters under the classical test theory approach (CTT). In order to easily sort out the patterns, we performed the Pearson correlation between the CDM parameters and PPDI (after excluding the two MS items with low or negative discrimination). The correlation coefficients between Δ_{Guess} and PPDI are .61 for the LF module and .27 for the MS module. The positive correlation can be interpreted as the guess parameters tended to increase from the pretest to the posttest; the amount of the increase is, on average, the least for close items, and the most for the far proximal items. This finding raises a puzzling question: why do students who did not master the attributes have a higher guessing probability at the posttest? Why is this more obvious for close items, but the least for the far proximal items? We speculate the items may involve several other attributes which are not included in the attribute definition and the Q-matrix (note that researchers often cannot specify too many underlying skills or understanding given the number of items, otherwise the model is too saturated). These hidden or unspecified attributes (e.g., memory), although not critical or dominate in the cognitive processes that

students employ to respond the items, may still play certain roles at the posttest.[3] Because close items are more similar to what students have experienced in their lessons, it might be more likely for students to figure out the right answers. Related to the previous explanation, another reason is that students have a higher probably of guessing because they recall key words, if not concepts, from instruction.

The correlation coefficients between Δ_δ and PPDI are -.36 for the LF module. The negative correlation can be interpreted as the discrimination parameters tended to decrease from the pretest to the posttest; in the other words, the close items become less discriminate for the posttest compared to the near and far proximal items. This observation is consistent with how we developed the items in the sense that the close items were intended to use similar activities to the learning activities that students have been exposed whereas the far proximal items were designed to tap the far transfer of student learning. Therefore, the close items may not be the most desirable to fully differentiate students of mastery and non-mastery because both groups of students are familiar with these items. However, the difference in their mastery level of the attributes can be captured in near and far proximal items when the students are supposed to transfer what they have learned. A similar pattern of item statistics has been replicated in another study where we performed the IRT analysis to examine how items of varying proximity corresponded to the extent to which teachers' instructions supported students' transfer of learning (self-citation, 2015).

[3]Examples of the unspecified attributes are vocabulary of landform types, applying the proportional reasoning to compare solubility and concentration levels of solutions, and understanding properties of matters.

Table 5. *Item Parameters from the CDM Analysis*

	Pre-classified	Pre-test					Post-test				
Item ID	Proximity	Guess	SE(Guess)	Slip	SE(Slip)	Disc.	Guess	SE(Guess)	Slip	SE(Slip)	Disc.
LF Module											
LF2	Close	0.370	0.034	0.471	0.057	0.159	0.416	0.059	0.288	0.029	0.296
LF3	Close	0.382	0.035	0.180	0.054	0.438	0.572	0.057	0.121	0.021	0.307
LF5	Close	0.313	0.044	0.194	0.043	0.493	0.239	0.066	0.120	0.025	0.641
LF10	Near prox.	0.224	0.041	0.292	0.047	0.484	0.467	0.063	0.163	0.024	0.370
LF13	Near prox.	0.398	0.034	0.428	0.057	0.174	0.176	0.054	0.273	0.029	0.551
LF14	Near prox.	0.308	0.035	0.180	0.056	0.512	0.304	0.060	0.140	0.024	0.556
LF17	Far prox.	0.241	0.038	0.441	0.046	0.318	0.151	0.051	0.426	0.031	0.423
LF20	Far prox.	0.097	0.089	0.649	0.054	0.254	0.141	0.105	0.650	0.029	0.209
LF21	Far prox.	0.410	0.035	0.256	0.053	0.334	0.333	0.059	0.174	0.025	0.493
MS Module											
MS2	Close	0.173	0.059	0.357	0.060	0.470	0.089	0.161	0.000	0.036	0.912
MS4	Close	0.275	0.097	0.565	0.031	0.160	0.578	0.125	0.518	0.030	-0.095
MS6	Close	0.002	0.116	0.373	0.033	0.625	0.672	0.107	0.050	0.014	0.278
MS9	Near prox.	0.019	0.027	0.856	0.034	0.125	0.079	0.055	0.841	0.023	0.080
MS11	Near prox.	0.377	0.036	0.430	0.090	0.194	0.410	0.086	0.208	0.073	0.382
MS16	Near prox.	0.339	0.043	1.000	0.086	-0.339	0.185	0.045	0.552	0.056	0.263
MS18	Far prox.	0.286	0.103	0.343	0.030	0.371	0.442	0.128	0.204	0.025	0.354
MS20	Far prox.	0.461	0.075	0.996	0.057	-0.457	0.071	0.070	0.533	0.034	0.396

Table 6. *Item Parameters Related to Instructional Sensitivity*

Item ID	Pre-classified Proximity	Item Statistics based on CDM Analysis			Item Statistics based on CTT		
		Δ_{Guess} (G_{post} - G_{pre}))	Δ_{Slip} (S_{post} - S_{pre}))	Δ_{δ} (δ_{post} - δ_{pre})	Pretest *p*	Posttest *p*	PPDI (p_{post} - p_{pre})
LF Module							
LF2	Close	0.046	-0.183	0.137	0.41	0.66	0.25
LF3	Close	0.190	-0.059	-0.131	0.53	0.81	0.28
LF5	Close	-0.074	-0.074	0.148	0.56	0.75	0.19
LF10	Near prox.	0.243	-0.129	-0.114	0.46	0.77	0.31
LF13	Near prox.	-0.222	-0.155	0.377	0.47	0.63	0.16
LF14	Near prox.	-0.004	-0.040	0.044	0.48	0.74	0.26
LF17	Far prox.	-0.090	-0.015	0.105	0.39	0.50	0.11
LF20	Far prox.	0.044	0.001	-0.045	0.27	0.34	0.07
LF21	Far prox.	-0.077	-0.082	0.159	0.53	0.72	0.19
MS Module							
MS2	Close	-0.084	-0.357	0.441	0.42	0.84	0.42
MS4	Close	0.302	-0.047	-0.255	0.40	0.50	0.10
MS6	Close	0.670	-0.323	-0.347	0.53	0.94	0.41
MS9	Near prox.	0.060	-0.015	-0.046	0.08	0.15	0.07
MS11	Near prox.	0.033	-0.221	0.188	0.42	0.62	0.20
MS16	Near prox.	-0.154	-0.448	0.603	0.23	0.33	0.10
MS18	Far prox.	0.156	-0.139	-0.017	0.62	0.79	0.17
MS20	Far prox.	-0.389	-0.463	0.852	0.21	0.43	0.22

Conclusions

In this paper, we conducted a series of CDM analyses with the item level scores from two science modules at grade 5 to examine how the instructional sensitivity of the items is related to their cognitive diagnostic estimates. Informed by the curriculum mapping by researchers and teachers, we defined the measured attributes and generated the Q-matrix for each module. Using the DINA model, we produced three parameters for each item in both the pretest and the posttest: guessing, slip, and discrimination. The analysis attempted to test the primary hypothesis that the changes of item parameters from the pretest to the posttest should reflect the item proximity.

We examined the mastery level of students with respect to each attribute. The attributes with the higher gain have been heavily addressed in the intended curriculum (i.e., a greater amount of learning activities) compared to those with relatively smaller gain. Overall,

the increased mastery levels on all the attributes provide evidence that the items of close, near proximal, and far proximal are sensitive to the instruction.

The CDM analyses produced three types of item parameters. Although the item and test indices show that the data fit well the specified Q-matrix, the discrimination parameters of two MS items appeared to be extremely low. We need to closely analyze the two items to understand why they were unable to sufficiently discriminate mastery students from non-mastery students. The items will be reviewed by content experts and we will conduct qualitative analysis of the think aloud protocols of students we collected from a small sample of students. We can also perform exploratory analysis to revise the Q-matrix regarding these two items.

We also took a closer look at the patterns of item parameters between pretest and posttest in comparison to the descending patterns of the PPDI estimates by item proximity. Two findings were observed: (1) Item guessing parameters increased from the pretest to the posttest; the magnitude of the increase was positively correlated with the PPDI estimates, suggesting that the largest increase occurred with close items and the least with far proximal items. (2) As expected, the close items tended to have more desirable discrimination parameters for the pretest whereas the far proximal items behaved better at the posttest. The observed patterns tentatively indicate that item guessing parameter is a suitable measure of instructional sensitivity when considering the changes between the pretest and posttest which might be explained by other unspecified/hidden skill or understanding attributes involved in the assessment items.[4] Additional research is needed to understand why this parameter outperforms the other two parameters and evaluate how it is compared against other indicators of instructional sensitivity.

[4] The success rate of *educated guessing* increases at the end of the module even no evidence shows that students have grasped the specified attributes. The exposure to the instruction may improve the implicit learning processes or result in partial understanding, which cannot be statistically model in the CDM analysis.

Generally, the item quality are acceptable for the LF module but not for the MS module since the item estimates in the MS module seem to be inconsistent with the instructional sensitivity pattern. The item parameters changed a lot between pretest and posttest because examinees had gained knowledge from lesson. However, it is interesting to see guessing parameters increased from pre-test to post-test. We speculate two reasons: (1) items may involve several other attributes which are not initially included in the attribute definition and the Q-matrix; and (2) students may recall key words, if not concepts, from instruction. When reading the item prompts, we would suggest researchers to include more relevant attributes and more items to empirically test these explanations in the future.

As an exploratory analysis, the limitation of this study is it used only a subset of the data which is from the common items. By including items that not all students responded to, the alpha estimates can be more stable. In IRT and CDM, the person parameter estimates (i.e., theta and alpha) can be compared even if examinees did not take identical items, as long as the missingness is not systematic (e.g., students skipped items they find easy/difficult). Then, the tests students take can be entirely unique if the sample size is large and the test is long. Therefore, we will include all items in the model fitting in future analyses.

Model selection is another limitation. Theoretically it might be the case that an attribute is a prerequisite to another attribute. For example, Attribute "Processes of erosion and deposition" can be considered as a prerequisite to Attribute "Factors that influence erosion and deposition" in LF module. Therefore, the CDM model that matches this cognitive structure/model, e.g. higher-order DINA (HO-DINA) model (de la Torre & Douglas, 2004), should be used in the further data analysis.

In sum, this study demonstrated the use of CDM analysis to examine the instructional sensitivity inferences of item scores. Unlike the typical CTT or IRT approach with items or tests, the reported mastery levels by attributes allow teachers and students to focus on the

learning goals when planning subsequent instructional activities to address the attributes which fail to meet the curriculum standards.[5] This illustrates benefits of using the CDM analysis in tying closely the assessment interpretations to instruction and learning, a critical issue that instructional sensitivity aims to address. The findings also raise several important questions for future research: (1) why did the guessing parameter appear a better candidate compared to slip and discrimination to evaluate items' instructional sensitivity? (2) how can the master levels of attributes be used to evaluate the effects of curriculum designs and teachers' instruction? Or a CDM-based DIF analysis with two groups using different curricula or groups of taught and untaught condition to evaluate the sensitivity to curriculum or instruction, respectively? (3) do the number of attributes defined and specified influence the accuracy of the estimated item parameters? The ratio of attributes to items in the MS module was relatively higher than those in other studies. It is unclear how the number of attributes may influence the model fit and robustness of the item parameters. All of these questions call for studies with a larger sample of items and replications with different modules, subject domains, and grades.

[5] The decision of the cut-point value for meeting the standards requires the standards-setting work, verified by relevant statistical analyses.

SECTION 3

EXAMINING THE RELATIONSHIP OF CHARACTERISTICS OF WORD PROBLEMS AND ITEM PARAMETERS IN THE CONTEXT OF AN ONLINE MATH GAME

Phonraphee Thummaphan[1], Alec Kennedy[2], Min Li[1], Zoran Popović[3],

Roy Szeto[3], Robert Duisberg[3], & Gabriella Silva Gorsky[1]

[1]College of Education, University of Washington, Seattle, [2]Evans School of Public Policy and Governance, University of Washington, Seattle

[3]Department of Computer Science and Engineering, University of Washington, Seattle

Abstract

As emphasized by the Common Core State Standards (CSSS), interest in modeling in mathematics word problems has been on the rise. However, there is lack of empirical evidence on the relationship between word problem characteristics and modeling performance. This study aims to investigate this relationship using data from a math online game. The sample included 225 players in Grades 4-6 and their modeling performance on 22 items across two booklets. Correlation analysis was performed to investigate the relationship of item characteristics and Cognitive Diagnostic Modeling (CDM) item parameters and found that consistency and model type were significantly correlated with item difficulty. The sequence analysis with students' action log data provided visualization of their modeling strategies that further validated the results of CDM and correlation.

Introduction

Defining Modeling

Interest in the use of modeling in STEM assessment has seen a recent increase. The importance of modeling has been stressed by both the Common Core State Standards (CCSS) and the Next Generation Science Standards (NGSS). The Common Core State Standard Initiative (2010) states that "Modeling links classroom mathematics and statistics to everyday life, work, and decision-making. Modeling is the process of choosing and using appropriate mathematics and statistics to analyze empirical situations, to understand them better, and to improve decisions" (p. 72). Modeling can assist real world problem solving by enhancing understanding of situation and decision making.

Modeling is commonly used in math word problems. According to Török (2013; as cited in Debrenti, 2015), "Word problems are real-life, practical problems in which the correlation between the known and unknown quantities are provided in the form of text, and their solutions need some kind of mathematical model."(p. 22). Math word problems often require students to use models which are representations of problem situations to relate the underlying mathematical concepts to real-world situations (Llinares & Roig, 2006; Murata & Kattubadi, 2012). Specifically, during the word problem solving process, models need to be constructed by students to clearly understand the problem situation.

Despite the importance of modeling as a critical skill in the cognitive processes involved in word problem solving (Debrenti, 2015), little research specifically assesses students' modeling performance (Goldin & Kaput, 1996; Barbosa, 2006). In most studies, the outcome measure is the accuracy of solving word problems (which is largely contaminated by computational skills), rather than whether students are able to construct appropriate models (e.g., Koedinger & Nathan, 2004; Debrenti, 2015). Moreover, since students can create multiple models and modify them during the problem solving process, it would be

important to look at the different models to examine students' modeling strategies, particularly the initial one. Using the game platform, not paper-pencil test, allows us to track students' modeling process which demonstrate different models. Hence, measurement research needs to not only differentiate the modeling performance from other steps involved in the problem solving process, but also look at the initial approach students used in their modeling, so that we can obtain more accurate information about students' performance.

Most importantly, research remains unclear on word problem characteristics that influence student performance. For example, while some studies stress the importance of explicit keywords in word problem solving (see Reed, 1999), others reveal that students incorrectly solve problems if keywords are irrelevant to or inconsistent with the situation (Carpenter, Hiebert, & Moser, 1981; Clement & Bernhard, 2005). Empirical evidence(s) that substantially demonstrates the relationship of item characteristics and modeling performance is critical for interpreting assessment information and developing items. Otherwise, lack of understanding on this issue can bring about invalid interpretation of assessment results as well as ineffective classroom teaching. This study aims to examine the relationship of problem characteristics and item parameters by analyzing students' response patterns collected from an online math word problem game. The emphasis of the analysis is to reveal the relationship between item characteristics and student's use of a bar model construction performance so that findings can be used to guide assessment instrument development and instructional practices.

Item Characteristics that Impact Students' Performance

The literature of math word problems has studied four types of item characteristics related to student performance: problem type, model type, keyword explicitness and consistency, and reading complexity. However, we do not focus on the characteristic of problem type that indicates math operation involved in the problem such as addition,

multiplication, and multiplicative comparison, in this study because there is a constraint on the number of attributes that can be included in the analysis. Furthermore, we believe that problem type and model type characteristics are closely related meaning that the omission of problem type characteristic does not result in the loss of too much information. We provide details on the three item features included in this study.

Model type. Model type can be considered as the relationship of the entities in the model. The "Model Method," one of the visual methods introduced in 1983 by the Ministry of Education's Curriculum Development Institute of Singapore, divided model types using three groups: part-whole (the relationship between a whole and its parts), comparison (the comparison of two or more quantities), and change (the relationship between the new value of a quantity and its original value either in a before–after situation or after an increase or a decrease) (Kho, 1987). These model types involve conceptual understanding of multiple aspects, such as number sense, mathematical reasoning, abstraction of mathematical relationships, and knowledge of the four operations (Ng & Lee, 2009).

Different model types generally involved different modeling strategy types which are types of representation used for creating the model such as number only, number line. For example, experimenting with Grade 3 students, Murata and Kattubadi (2012) examined the modeling used in three different model types: separate (the action involves separation of quantities), part-whole, and comparison. Each model had two problems. Considering within each model type, the majority of students used number-only strategy types for both separate model problems, whereas more students used number-line strategy types in the first problem and switched to number-only strategy types in the latter problem for the part-whole and comparison model problems. Thus, the findings appear to indicate that different model types induced different modeling strategy types.

Keyword Explicitness and Consistency. Keyword explicitness is defined as the appearance of a relevant term in the context/question. In writing or generating the items, the relevant terms (e.g., "total" and "average") can be written either explicitly or implicitly in the context. When the relevant term or keyword is shown in the surface structure of the question, called explicit, students can more easily activate the relevant concepts that they have learned from their courses and use such concepts to solve the problem, whereas students may be more difficult recall the concepts to solve the implicit version with no or less keywords (Maloney & Siegler, 1993). Therefore, compared to items with implicit concepts/words, items with explicit concepts/words may considerably influence students to choose appropriate approaches for representing and solving problems, thus increasing the likelihood of producing correct models.

Another variable that can affect student performance by interacting with keyword explicitness is the consistency of the keyword and operation involved. Sometimes the inconsistency can occur when the cue word does not match the math operation (see Lewis & Mayer, 1987; referred to as "marked" and "unmarked" in their study; Clement & Bernhard, 2005). For example, "Sara has 6 flowers. Mary has 3 flowers more than Sara. How many flowers does Mary have?" and "Sara has 6 flowers. She has 3 flowers more than Mary. How many flowers does Mary have?" Although, the same keyword "more than" appeared in both examples, it means add in example 1 but subtract in example 2. If students think "more than" generally implies addition, the keyword and math operation is consistent in example 1 but inconsistent in example 2. Lewis and Mayer (1987) found that students tended to miscomprehend the second example when the problem involved an inconsistency between the required arithmetic operation and the relational term.

Reading complexity. According to Edwards, Maloy, and Anderson (2009), "math word problems are intricate language constructions—they contain unfamiliar words, complex

combinations of text and numbers, and considerable amounts of information to decode and organize" (p. 1). Complexity of text, or reading complexity, is one of the item characteristics in word problems that can impact students' modeling performance. Barbu and Beal (2010) conducted a study to examine the effects of linguistic complexity and math difficulty on word problem solving performance of middle school English learners by defining text complexity as being made up of two components: vocabulary and grammatical structure of the problem. They found that students had lower performance on word problems with more complex text as compared to the same problems with easier text. Such complexity likely arises from the multiple requirements of text interpretation and reading comprehension for word problem modeling and solving.

Reading or text complexity can be defined and measured in different ways. Schuster and Erickson (2014) define text complexity as "the degree to which a text is easy or difficult to read and understand. Text complexity is influenced by factors at the word level, such as word frequency; the sentence level, such as the syntactic complexity of the sentence; the text level, such as number of connectives combining the sentences and paragraphs together; and factors outside the text, such as the characteristics of the task and the reader" (p. 4). So consideration of text complexity should include both vocabulary and sentence structure aspects which we believe are two main constraints for elementary school students.

Cognitive Diagnostic Modeling Approach

Cognitive diagnostic modeling (CDM) can help to parse out the underlying cognitive aspects of a word problem. CDM is a psychometric technique used for measuring latent constructs. It has various models with different characteristics. For example, DINA (Junker & Sijtsma, 2001; de la Torre, 2009) model is noncompensatory because it requires all measured attributes for producing correct answer. It is also parsimonious and interpretable since it estimates only two parameters for each item. So it is frequently used in the field.

CDM can be used as a method for looking into the diagnostic value of items or as a tool for validating the attributes of items. In this paper, it is a tool for validating the attributes in items. The study attributes have two main categories, the item characteristics and the underlying cognition needed to respond to the item. While the cognition is consistent with the typical CDM literature, exploration of item characteristics is still sparse. By considering item characteristics as attributes (knowledge or skills) in the modeling process, CDM can estimate students' attribute mastery level and item parameters which can be used to examine the relationship of word problem characteristics and item parameters.

In summary, there are four categories of variables that are associated with the modeling process. Two research gaps exist in the literature of measuring student modeling performance. First, many studies have not differentiated between two different performance outcomes: constructing an appropriate model versus generating a correct answer (see e.g., Edens & Potter, 2006; Murata & Kattubadi, 2012). Therefore, we do not yet know the impact item characteristics have on modeling. Second, there is insufficient empirical evidence regarding the effect of the item characteristics on modeling performance. Researchers should investigate the impacts of each characteristic of word problems as well as the interaction between them on modeling and solution performance. Therefore, in this study, we ask the following research question: *How are item characteristics of word problems associated with item parameters*?

This study uses several methods to examine the relationship between item characteristics and item parameters. Using CDM and correlation analysis, we examine the claim that the word problems from an online math game provide detailed and useful information on the relationship of item characteristics and student modeling performance. Furthermore, we assess the claim that the log data from the online game provides a detailed

description of the relationship of different item characteristics and student modeling processes that confirms the evidence provided by the CDM and correlation results. To be specific, we find that different item characteristics relate to student's modeling performance in different ways. Particularly, we hypothesize that items with keyword implicitness and inconsistency demonstrate distinct association with student modeling strategy and performance.

Methods

To overview the research methods, we began with defining the attributes for guiding the item development. Then we developed the assessment items and Q-matrix based on our attribute definition. After we collected the data, we estimated parameters in the CDM model fitting step. Finally, we validated the results of CDM using the log data. Further details of each step are provided below.

Attribute definition

We developed the attribute definitions based on the item characteristics literature as well as the attribute coding framework that we previously developed and validated specifically to analyze word problem solving skills (Thummaphan, Dong, & Li, 2015). Initially, nine attributes related to the item characteristics were developed. But only some critical attributes were chosen because of the literature support and redundancy which would be explained later. Finally, five attributes were employed to guide the item development as shown in Table 1. To make the consistent directions of the relationships between the item characteristics and modeling performance, we reversed the keyword explicitness and consistency to be keyword implicitness and inconsistency.

Table 1. *Attribute Definitions for the Items of the Two Booklets*

Label of Attribute	Explanation
Model type	Solving problems with model type 4a or similar (i.e., given the smaller quantity and the multiple, find the larger quantity.)
Complex procedures	Dealing with problems that involve two or more steps/sub-goals
Complex text	Reading sentences with longer than 31 words* and containing at least one difficult vocabulary word
Keyword implicitness	Dealing with problems that do not include any explicit keywords
Inconsistency	Dealing with problems that have inconsistency between underlying math concept(s) and keywords

Note: * The threshold for the item length was defined by calculating the average lengths of 50 NAEP released math word problems

Item Development

The instrument used in this study is an online game initially developed in 2014 for students at upper elementary and middle school grades, with a focus on problem solving and modeling. The bar model, which uses bar/rectangles to represent/model problem situations, was chosen as it aligns with the learning purpose of the online game and is consistent with the curriculum materials at the local school districts selected. So this study is an exploration of student's use of a particular model, rather than choose what model they were going to use. There were two major steps in playing the game. First players extract information and construct a model. Second, they set up the equation. Players can try multiple times/attempts to construct the model and set up the equation. They also can verify if their constructed models are correct or not. During playing the game, students can get the feedback or hints from the game. The entire set of a player's actions during the game playing is defined as action log data. In this study, we only analyzed the actions from the first step since they provided direct information on how students approached the modeling process.

The mathematics word problems were systematically constructed. First, relevant word problems were collected from released NAEP, TIMSS, PISA, and state test items and the profiles of these items were created as examples. Then, 47 word problem items were developed and verified by content experts based on several identified item characteristics (e.g., model type, keyword implicitness, and inconsistency) and mapped to the CCSS related

to math modeling. The items were put into the online game, piloted in multiple play tests, and then revised to improve the clarity of the language. To ensure a sufficient sample size, eleven items, representative of each of the item characteristics, were selected, and then randomly assigned into two booklets. The items were incorporated into the online game along with four filler items. Moreover, other items (e.g., automatically generated items and collected released items) were also put into the game to assure the educational benefit of the game to its users. Examples of word problem items with the respective model type are presented in Figure 1.

Figure 1. *Examples of Word Problem Items from Different Sources Included in the Game*
Note: The lower part is the modeling done by the player

Since we are interested in the students' ability to construct situation/problem models, rather than solution/computation models, the game was designed to allow students to create models and set up equations in steps. Then, the interface was created to differentiate modeling actions from initial set-up and later computation (See Figure 2.). At the beginning of the game, a tutorial is provided so that players can learn and become familiar with the interface to play the game.

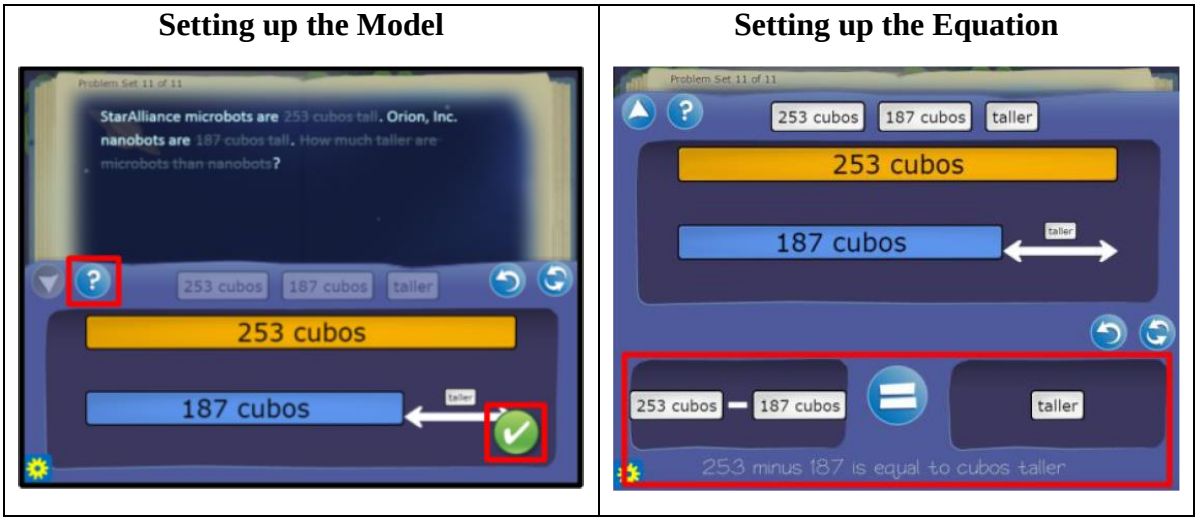


Figure 2. *Examples of Setting up the Model and Equation Steps in the Game*
Source: BrainPOP (n.d.)
Note: The below part is the modeling done by the player

Q-matrix

Three researchers created the Q-matrix for the math word problems (see Table 2 for the Q-matrix). After a thorough review of the items and Q-matrix by two content experts, we judged that there were no misspecifications of attributes in the Q-matrix. Each item measured one or more attributes. The list of nine initial attributes was narrowed, removing redundant ones and keeping only those that were more prevalent in the literature. Specifically, the attribute of "Understanding the concepts and operations involving fractions, decimals, and ratios," was dropped because it was an alternative version of the attribute "Whole number" in the coding; "Decoding the context that require retrieval of situational information," was dropped because it was highly correlated with "Complex text"; and "Creating models for problems where students can select different problem solving strategies," was omitted because the game does not really allow players to demonstrate the use of multiple strategies and was found to be highly correlated with the attribute "Complex text." The three attributes were dropped instead of being collapsed with other attributes because they were not conceptually compatible to combine. Due to the small number of items, we continued to drop one more: "Whole number" since most items (10 out of 11 items) measured this attribute

which then provide few variance in analysis. The five selected attributes were measured in all eleven paired items, with each item containing certain level of each attribute for each booklet,

Attributes measured in the two booklets are quite similar. Each item in both booklets assessed from one to five attributes. Most attributes were assessed by a moderate number of items, ranging from three to six. Moreover, the number of items that measured attributes is also comparable. Although the number of attributes seems too large for the number of items, these attributes represent critical item characteristics supported by the literature. So we decided to include them for the purpose of an exploratory analysis. The Q-matrix is shown in Table 2.

Model fitting

Data Collection

The data were collected during Winter 2016. K-12 teachers in Washington State were asked to register for the challenge to provide an opportunity for their students to participate in game learning. Then, teachers tasked the students with playing the game. When students started playing the game, they were automatically and randomly assigned to one of two booklets. At the end of data collection, a total of 2,860 students participated: 1,436 for Booklet 1 and 1,424 for Booklet 2. From this sample, we selected 1,354 players who were in grades 4-6 because the constructed word problems were appropriated to their grade level.

The log data was used to keep track of player actions throughout the game to understand modeling strategy. We focused our analysis of the log data on the first valid attempt which may be neither real “initial model” before hints or changes, nor their final model. Specifically, we define the first valid attempt in a way to make sure the log data truly reflected the actions of the game players when they engaged with the problem solving. First, we dropped undo, remove, reset, and adjacent actions, to include only actions that were

Table 2. *Q-matrix for Modeling*

Item ID	Model type	Complex Procedure	Complex text	Keyword Implicitness	Inconsistency	Total
	A1	**A2**	**A3**	**A4**	**A5**	
Booklet 1						
1	0	1	0	1	1	3
2	0	0	0	1	0	1
3	1	0	0	1	0	2
4	1	0	1	1	1	4
5	0	0	1	0	0	1
6	0	1	0	1	0	2
7	0	1	1	0	0	2
8	1	0	0	1	0	2
9	0	1	0	0	0	1
10	0	1	1	0	1	3
11	1	1	1	0	0	3
Total	4	6	5	6	3	
Booklet 2						
1	0	1	0	1	0	2
2	0	0	0	1	1	2
3	1	0	0	1	1	3
4	1	0	1	1	0	3
5	0	0	0	0	1	1
6	0	1	0	0	0	1
7	0	1	1	1	0	3
8	1	0	0	0	0	1
9	0	1	0	0	1	2
10	0	1	1	0	0	2
11	1	1	1	1	0	4
Total	4	6	4	6	4	

relevant to the verified model, not the trial and error style that players took. Second, we chose only those sequences of actions with at least two extracting information and two model construction actions since they are the minimum number of actions needed to demonstrate players made a serious attempt in solving the item and to maintain a sufficient sample size. Third, to be considered a *first* valid attempt, the sequence must appear as the first set of actions prior to receiving feedback from the game. Finally, to balance between focusing on a subset of serious players and achieving a sufficiently large sample size, we selected players

who played *at least 9* items (80% of all items) in each booklet, yielding a final sample size of 105 unique students for Booklet 1 and 120 for Booklet 2 out of about 1,354 students.

After identifying the first valid attempt, we had to determine what data to use for each part of the analysis. For CDM, we chose to model performance on the first valid attempt. Specifically, we considered the accuracy of the students' model solution as an indicator of their modeling performance; in other words, a binary outcome variable was used to indicate the correctness of a students' first valid attempt. Next, the item difficulty (1-slip) parameters from the CDM were used for correlation analysis. For the sequence analysis of the log data, we used all action logs included in the first valid attempt. We also used the indicator of model correctness from the first valid attempts to calculate the percentage of action sequences producing correct models.

To make sure that the prior performance of students included across two booklets was comparable, we employed a t-test of the average scores of the four filler items that were presented to all players who played the game. Student performance of the two groups was comparable because average modeling performance of those items did not differ significantly ($t = .163$, $p = 0.817$). Hence, any difference in players' modeling performance on the other booklet items is due to item manipulations.

Data Analysis Methods

CDM helps us understand students' attribute mastery levels based on item characteristics. Given the small sample size, we analyzed the data using the DINA (Junker & Sijtsma, 2001; de la Torre, 2009) model with the GDINA package in R (Ma & de la Torre, 2016). This package provides a framework for cognitive diagnosis analyses for dichotomous data within the generalized DINA model framework. The results of this analysis included the students' attribute mastery level and item parameters: guessing and slip.

Validation

Two strategies were used to validate CDM results including correlation and sequence analyses. Correlation helps us examine the association of item characteristics and CDM estimated item parameter. After we identified the difficulty parameter, which is 1-slip (implying that players who mastered all measured attributes can produce the correct model), we ran the correlation to understand how item characteristics are associated with item parameters using the item as a unit of analysis. We included five important variables: model type, keyword implicitness, inconsistency, reading time, and item difficulty. We anticipate that the correlations between item characteristics and item parameters should exhibit the same direction as the CDM results.

Sequence analysis was selected as a method because it helps us recognize the relationship of item characteristics and students' modeling patterns/strategies. By capturing students' modeling actions, we were able to analyze students' modeling strategies and tie them back to item characteristics. The sequence analysis provides the modeling action patterns that display the actions relevant to significant item characteristics based on the CDM results.

To avoid overwhelming information, we collapsed some of the actions from the action logs together. This made the descriptive plots from the analysis more interpretable and readable. For example, we chose to combine actions "Add horizontal bracket to bar model," "Add vertical bracket to bar model," and "Change the name on a box, this is a bar model step" into a single category related to labeling. Actions related to the mathematical operations, including addition, subtraction, multiplication, and division, made up their own categories since they could be evidence of players' strategies due to the item features. Eventually, we created 12 separate groupings of actions that took place between the start of

the item and the validation of the model or quitting/skipping the item, as shown in Appendix A.

Besides collapsing actions, items with a similar modeling interface in the game were grouped together to make more interpretable findings, resulting in five groups of items, with 1-4 items in each group, as presented in Table 3. The sequence analysis was part of the validity research to investigate how item characteristics are related to modeling actions.

Table 3. *Groups of Items*

Group	# of Items	Model Interface	Operation Type of Items in the Group	*Implicitness & Inconsistency	
				Booklet 1	Booklet 2
1	1	1a	Addition	IC	EC
2	3	2a	Addition, subtraction, multiplication	II, IC, EC	IC,II,IC
3	1	3b	Division	EC	EI
4	4	4a	Multiplication, addition	IC,II,IC,EC	II,IC,EC,IC
5	2	4c	Multiplication, subtraction	EC,EI	EI,EC

Note: II = Implicit and Inconsistent, IC = Implicit and Consistent,
EI = Explicit and Inconsistent, EC = Explicit and Consistent

To illustrate, we used the TraMineR R package (Gabadinho et al., 2011) that allows us to describe and compare data in which there is an explicit ordering of categorical events/states. Specifically, we used TraMineR to calculate similarities between different action sequences in order to group students who exhibited similar modeling strategies using cluster analysis. Since a majority of students (82.09%) completed problems prior to the 18th action, we chose to cut the plots to only show actions up to this point. Also, in the figures presented, we only show actions 2-18 (after removing the first step, which is always "start of the problem"). Moreover, in each item group, we provided three clusters at most for the purpose of interpretability.

After clustering the sequence patterns based on a similarity matrix computed using the R package TraMineR, we visually compared the clusters across different types of item

features. We then considered the percentage of students correctly answering the model for each cluster. This could demonstrate the relationship of item features and modeling strategies for different types of students (i.e., high vs. low performers). This descriptive analysis was used to explore how items with certain features correlated with student modeling strategies, and functioned as a way of understanding why some items appeared more difficult than others. The set of findings using the action log data was also cross-validated with the mastery patterns from CDM.

Results

In this section, we present the results from our analysis. First, the CDM results display mastery level and estimated item parameters that infer the relationship between item characteristics and item parameters. The correlation results reveal the relationship between item characteristics and CDM item parameters. Finally, sequence analysis results validate the correlation results by showing the modeling strategies associated with item characteristics.

CDM results

The CDM results presented here included model fit indices, attribute mastery and item parameters as follow.

Model Fit Indices. The absolute fit statistics evaluate the relation between word problems in the game and data as shown in Table 4. The three fit indices for both booklets were close to zero. For instance, Mean Absolute Deviation (Mean Abs. Dev.) of the proportion correct (Prop) was 0.0044 for Booklet 1 and 0.0047 for Booklet 2. This indicated that the model, specified by the Q-matrix, was a good-fit with the data.

Table 4. *Item-level Fit Statistics*

	Booklet 1			Booklet 2		
	Prop	Z(Corr)	Log(OR)	Prop	Z(Corr)	Log(OR)
Mean Abs. Dev.	0.0044	0.0662	0.2897	0.0047	0.0724	0.3269
Max. Abs Dev.	0.0156	0.1956	0.8672	0.0105	0.2301	1.0871
SE(Max. Abs Dev.)	0.3265	1.9757	2.0830	0.2360	2.4893	2.6638

Note. *Mean Abs. Dev.* is Mean Absolute Deviation; *Max. Abs Dev.* is Maximum Absolute Deviation; *SE(Max. Abs Dev.)* is Standard Error of Maximum Absolute Deviation; *Prop* is residuals between the observed and predicted proportion of correct individual items; *Z(Corr)* is residuals between the observed and predicted Fisher-transformed correlation of item pairs (referred to as transformed correlation); *Log(OR)* is residuals between the observed and predicted log-odds ratios of item pairs.

Attribute Mastery. Mastery levels are presented in Figure 3. Mastery level of students in Booklet 2 were higher than those of students in Booklet 1 despite that students' performance was comparable across the two booklets on the filler items; Mastery levels ranged from 55% to 61% in Booklet 1 and 61% to 91% in Booklet 2. While "Keyword implicitness" was the highest mastery level, students had the lowest possession of the attribute "Inconsistency" for Booklet 1. In Booklet 2, students possessed "Complex procedure" as the highest attribute with 91%, and the lowest mastery level was "Model type". These results suggested that students in both groups lacked skill in solving problem with a model of type 4a or similar (Given the smaller quantity and the multiple, find the larger quantity.). It is interesting to see that students in group 2 had higher skills in dealing with problems that have those item characteristics, particularly a complex procedure. Since the students had the same performance level on filler items and the items between the two booklets were carefully paired up based on similar characteristics, we speculate that the students in Booklet 2 showed higher mastery levels because of the hints they requested from the game that helped their modeling. The numbers/types of hints will be considered in further analysis.

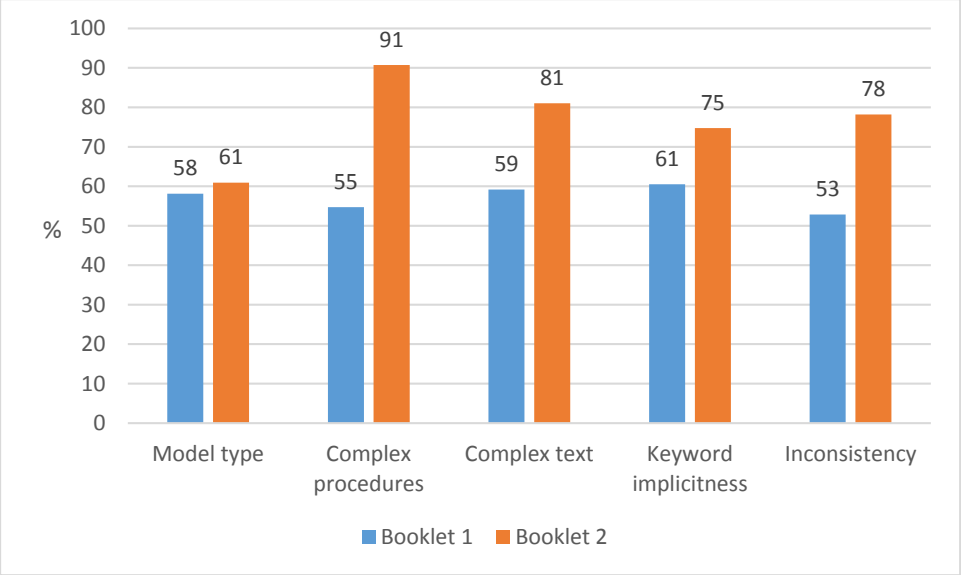


Figure 3. *Attribute Mastery Level by Attribute (in percentage)*

Item Parameters. The guessing and slip parameters for the two booklets varied, as presented in Table 5. While guessing parameters in Booklet 1 ranged from 0.000 to 0.761 with one item of 0.000 guessing values, those in Booklet 2 ranged from 0.000 to 0.744 with three items of 0.000 guessing values. This suggested that items in Booklet 1 had higher probabilities of providing correct answers by guessing than those in Booklet 2. In contrast, while the range of slip parameters were comparable with 0.000 to 0.613 for Booklet 1 and 0.000 to 0.631 for Booklet 2, number of items with 0.000 slip value was three items in Booklet 1 and one item in Booklet 2. This indicated that items in Booklet 1 had higher probabilities of providing incorrect answers although students possessed all measured attributes than those in Booklet 2.

Regarding the item difficulty (1-slip), most item difficulties were high for both booklets, except Item 11 in Booklet 1 (item difficulty of .387) and Item 6 in Booklet 2 (item difficulty of .369). These two items are explicit, but one of them is model type 4a which involves the integration of part-whole and comparison models, and the other is inconsistent. Considering item difficulty in relation to the item features of interest (i.e., model type, keyword explicitness, and consistency), the results revealed a high value of item difficulty for

most items with model type 4a. We also found that all items with no explicit keyword had at least a 0.5 item difficulty value in both booklets. It is interesting that while all items with inconsistency in Booklet 1 had over a 0.9 difficulty value, these kinds of items in Booklet 2 had a wide range of item difficulty ranging from .369 to .808. While most items had high item difficulties, the findings showed that model type and inconsistency may correspond with low item difficulty values.

Correlation analysis results

The correlation coefficients between item characteristics and difficulty are presented in Table 6. There was only one variable in each booklet that was statistically significant with item difficulty at .05 level: inconsistency and model type (r=.715 and .663, respectively). These findings suggest that the inconsistency between the mathematical operation and key word, and model type 4a or similar (Given the smaller quantity and the multiple, find the larger quantity.) are positively associated with item difficulty. This suggests that greater inconsistency and model type 4a are associated with higher item difficulty.

The results strongly supported the importance of model type and inconsistency. As mentioned by Murata and Kattubadi (2012), students used different modeling strategies with different model types. To illustrate, Model type 4a involves multiplicative comparison that requires the integration of part-whole and comparison models, then require complex thinking than other simpler or single model types. Also, results suggested that inconsistency decreases item difficulty value, which aligns with the findings from Lewis and Mayer (1987) and Clement and Bernhard (2005). These findings suggest that inconsistency between mathematical concept and wordings make the item more difficult because it causes confusion or misunderstanding of the problem requirements.

Table 5. *Item Parameters*

Item	Model type (4a)	Explicitness & Consistency	Guess	SE(Guess)	Slip	SE(Slip)	Difficulty
Booklet 1							
1	N	II	0.295	0.052	0.000	0.219	1.000
2	N	IC	0.278	0.098	0.430	0.075	0.570
3	Y	IC	0.761	0.052	0.171	0.064	0.829
4	Y	II	0.554	0.054	0.000	0.271	1.000
5	N	EC	0.000	0.092	0.464	0.088	0.536
6	N	IC	0.213	0.058	0.397	0.099	0.603
7	N	EC	0.260	0.080	0.463	0.087	0.537
8	Y	IC	0.000	0.132	0.000	0.210	1.000
9	N	EC	0.017	0.057	0.349	0.083	0.651
10	N	EI	0.210	0.067	0.027	0.142	0.973
11	Y	EC	0.683	0.065	0.613	0.125	0.387
Booklet 2							
1	N	IC	0.000	0.105	0.497	0.068	0.503
2	N	II	0.115	0.078	0.377	0.075	0.623
3	Y	II	0.595	0.062	0.192	0.075	0.808
4	Y	IC	0.423	0.080	0.000	0.054	1.000
5	N	EI	0.001	0.121	0.504	0.063	0.496
6	N	EC	0.107	0.210	0.603	0.051	0.397
7	N	IC	0.101	0.063	0.413	0.074	0.587
8	Y	EC	0.000	0.121	0.358	0.084	0.642
9	N	EI	0.000	0.139	0.631	0.062	0.369
10	N	EC	0.034	0.153	0.271	0.063	0.729
11	Y	IC	0.744	0.066	0.371	0.084	0.629

Note: Y = Model type 4a, N = other model types, II = Implicit and Inconsistent, IC = Implicit and Consistent, EI = Explicit and Inconsistent, EC = Explicit and Consistent

Table 6. *Correlation Coefficients between Item Characteristics and Item Difficulty*

	Operation step	Model type	Keyword implicitness	Inconsistency	Reading complexity
Item difficulty of Booklet 1	-.253	.237	.493	.715*	.293
Item difficulty of Booklet 2	-.572	.663*	.472	-.185	.448

Note. Item difficulty = 1-slip, N=11 items, * $p < .05$.

Other item characteristics, including keyword implicitness, and reading time were not statistically significant, which contradicts findings from previous studies (Maloney & Siegler, 1993; Koedinger & Nathan, 2004; Ng & Lee, 2009; Kattubadi, 2012). These differences could potentially be explained by differences in the levels of those variables for different items. For examples, keyword implicitness in this study is operationalized by the level of the

keyword appearance to make it possible to manipulate different versions of inconsistency; implicit items still have some keywords. Moreover, students perceive the appearance or importance of keywords depending on how they are taught in the classroom; if they are familiar with solving the problem by not using the keyword as a strategy, they may not pay attention to the keyword. In addition, because most of the word problems were of similar length in the present study, reading time did not vary much. Nevertheless, these findings were further validated by the sequence analysis results presented in the next section that examine how students' strategy uses may be linked to specific item features.

Sequence analysis results

The findings from the sequence analysis focus on five groups of items that share a similar modeling interface, as described earlier in Table 3. In each item group, we produced 2-3 clusters of sequence patterns as shown in the Appendix B. However, we highlight only patterns that demonstrated significant relationships of item features and item parameters in the correlation analysis. These features are presented in two parts: (1) inconsistency, and (2) model type.

***(1) Inconsistency*.** Since the items in Item Groups 3 only differed in terms of consistency, we determined the relationship of this item feature and modeling strategy by comparing patterns across booklets. There were three clusters of action patterns in both booklets: (1) used some extracting information with adding row and bar and minimal adding unit bar actions, (2) used many extracting information with adding unit bar actions, and (3) used few extracting information and jumped into adding unit row and bar, but minimal adding unit bar actions. Although those three clusters were similar across booklets, we found that students used more extracting information steps in Cluster 2 of Booklet 2 than the same cluster in Booklet 1 that could identify the relationship of inconsistency and modeling strategy. We also found that students in Cluster 2 of Booklet 2 used extracting information

without other actions in the first eight actions. This comparison is illustrated in Figure 4. This indicates that students used more extracting information actions to make sure that they correctly understood the problem situation when dealing with inconsistency in the item which was associated to having better modeling performance. Also, students may be confused or misunderstand the problem requirements as shown by fewer students used of the "divide" (brown) action in Booklet 2, although division is a required operation for this item pair.

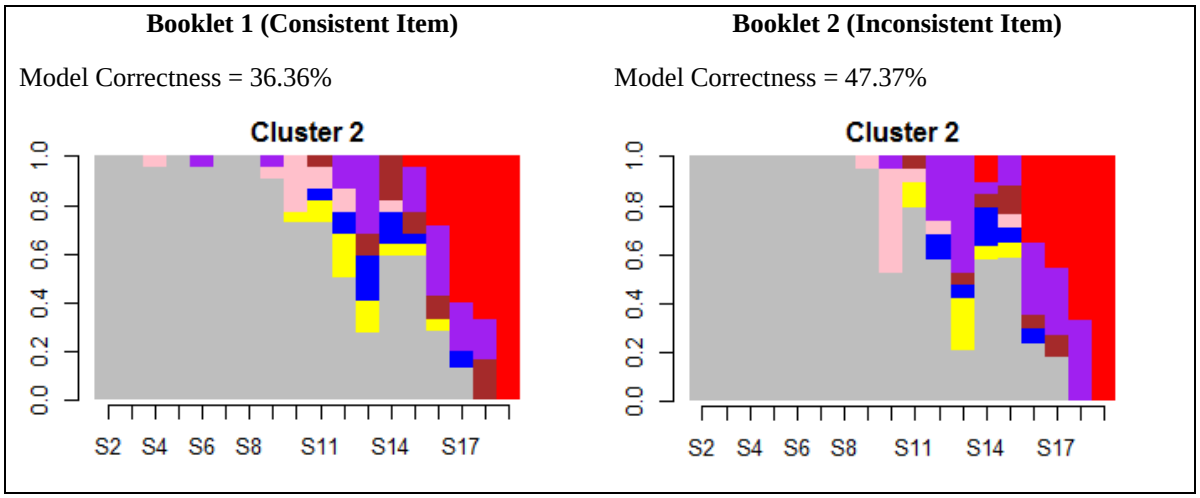


Figure 4. *Sequence Analysis Results of the Cluster 2 of Item Group 3*
Note: 'gray' = 'Extract info', 'yellow' = 'Add row', 'blue' = 'Add bar', 'lime green' = 'Subtract', 'pink' = 'Add unit bar', 'brown' = 'Divide', 'purple' = 'Add label', 'light blue' = 'Resize label', 'red' = 'Validate', 'coral' = 'Quit/Skip'

(2) Model type. To represent the relationship of model type and modeling strategy, we compared item groups which are different in model types but similar in other item characteristics. We compared Item Group 4, which included all items with model type 4a, with Item Group 2 which included items with model type 2a (Given two quantities, find the difference) in Booklet 1 only since the two item groups in this booklet included both versions of keyword implicitness and inconsistency, whereas Booklet 2 had incomparable versions of implicitness. Specifically, Item Group 4 involved the integration of part-whole and comparison model, whereas Item Group 2 involved only comparison model. The sequence analysis produced three clusters of action patterns for Item Group 4, but two clusters for Item

Group 2. To illustrate, the three clusters in Item Group 4 include (1) heavily used many extracting information with adding unit bar actions, (2) used many extracting information with adding row and bar actions, and (3) used few extracting information and jumped into adding unit row and bar, but minimal adding unit bar actions, whereas no cluster in Item Group 2 involved adding unit bar actions. It seems if students missed the actions of adding a unit bar (an important action in the integration of part-whole and comparison model), they will have a lower percentage of model correctness in Item Group 4. To explain, one cluster from Booklet 1 with the least model correctness for each item group was shown in Figure 5. A cluster of Item Group 4 with observed less adding unit bar (pink) actions resulted in 45.47% model correctness which was much smaller than the highest model correctness of this booklet's Item Group 4 with 64.47% (See Appendix B.). Whereas adding unit bar action was not needed for items in Item Group 2, the cluster involved a few subtraction, which is the right strategy in modeling and solving the problem, after adding row and bar actions yielded 38.14% model correctness which was not much differed from the highest performance level of 43.75% (See Appendix B.). Therefore, adding unit bar is the critical action that makes a big difference in modeling problem with the integration of part-whole and comparison model.

Comparing between the item groups, it was interesting to see that the percentage of students getting the correct model in Item Group 4 were higher than those in Item Group 2. The items with only comparison model were harder than items with the integration of part-whole and comparison model. We speculate that the part-whole model component in the integration of part-whole and comparison model can be a bridge or scaffolding for students' modeling strategy for the items with the integration of part-whole and comparison model.

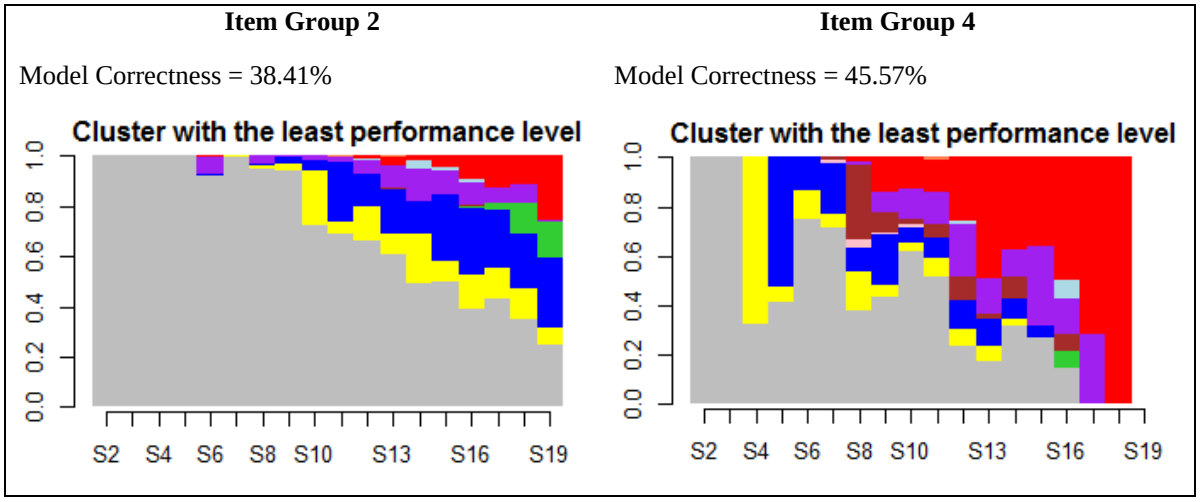


Figure 5. *Sequence Analysis Results of the Cluster 1 of Item Group 2 and 3 of Item Group 4 for Booklet 1*
Note: 'gray' = 'Extract info', 'yellow' = 'Add row', 'blue' = 'Add bar', 'lime green' = 'Subtract', 'pink' = 'Add unit bar', 'brown' = 'Divide', 'purple' = 'Add label', 'light blue' = 'Resize label', 'red' = 'Validate'

Conclusions

In this paper, we conducted multiple analyses of psychometrics and learning analytics to provide empirical evidence on the relationship of item characteristics and students' modeling performance. Using the DINA CDM model, we produced two item parameters, guessing and slip, as well as students' attribute mastery level.

To validate the CDM results, we used correlation and log data analysis. Using the CDM estimated item parameter as a variable, we ran a correlation to explore the relationship of item features. The results indicate that model type and consistency significantly correlated with item difficulty. These findings suggest these item characteristics are critical in students' modeling processes.

Given the information from the correlation, we used learning analytics exploring modeling strategy patterns of student log data to incorporation with the percentage of correct items to explain more about the relationship of item features and modeling, particularly model type and consistency. With the grouping of items, the visualizations of sequence and cluster analysis results show the modeling actions that are relevant to those item characteristics, as we found that students may be confused or misunderstand the problem

requirements and may need to spend longer time in extracting information if they faced inconsistency in the item. Moreover, we found that unit representation or adding unit bar is a critical action that has to be included in modeling to produce the correct model.

The present study has several limitations. First, the sample size in this analysis is too small. There were only about 100 students who played at least 9 out of 11 items in each booklet. We chose to limit our sample to these students to ensure that the data were from serious players and thus valid for analysis. Another constraint that limited our sample is the criteria defining the first valid attempt for sequence analysis. This was done to reflect student modeling strategy for those who made a valid attempt at solving the problem. In the further study, data analysis with a larger sample size would be of interest in order to reach more reliable results.

There were pros and cons associated with our decision to group items for the sequence analysis. Because we grouped the items, they encompassed different versions of item features in one group, which made it hard to conclude the relationship of specific item features and modeling strategy. However, not grouping the items, while it may have made it easier to determine the relationship of individual item features and the modeling strategy to validate CDM results, would have resulted in too much detailed information to yield understandable results. Nonetheless, future research could employ another method, which is manipulating items in the same group but in different booklets in order to have contrasting versions of the same item characteristics. This would allow us to clearly compare the two booklets by using the item group.

Another limitation is related to the data that we analyzed. The results of this study were based on the first valid attempt. The definition of the first valid attempt impacts the model correctness and number of actions included in the analysis. In the present study, the data from the first valid attempts were not truly the first time students validated the model.

However, we needed data from students who know how to play the game to avoid students having problems with the game interface. Therefore, we deleted all irrelevant data such as undo and remove actions, to clean the data as much as possible. Although, this method yielded the most reliable results, it will be interesting to analyze all the attempts in one analysis to gain a fuller picture of students' thinking. This will provide more understanding of the relationship of word problem characteristics and item parameters.

In addition, determination of the similarity of the modeling strategies or action patterns in sequence analysis is based on visual consideration in the present study. Such analysis is sufficient for the purpose of this study. Nonetheless, it would be useful to use statistical analysis to test the similarity of the clusters. Then, the results would inform the level of similarity with more objective and detailed information.

To summarize, the findings of the relationships between item features and students' modeling performance can inform educators and researchers about assessment design. Specifically, the item characteristics found to be highly related with modeling skills should be taken into account when developing assessment instruments. The resulting assessments can potentially provide more information about students' strengths and weaknesses, which can be used for planning and modifying instructional activities that address individual students' learning needs.

Thus, the current study yields two measurement contributions. First, these results contribute to pinpointing which characteristics are relevant for item development in measuring students' modeling performance and strategy. Within the game platform that allows us to see students modeling actions, rather than the final written models from the paper-pencil test, CDM and log data analyses confirmed that model type and consistency are critical item features. These findings guide word problem development in several ways. If we want to make items easy, particular for young or novice students, items should contain

consistent information of mathematical concepts and keywords and include a simple model type. Keywords can be either explicit or implicit. If we want to make the item more complicated in some situations, such as for higher grade students, we can include more complex model type. This will help us examine students' thinking in mathematical modeling processes, particularly in extracting information.

Second, the use of CDM can provide useful, targeted information about students' performance, which enables teachers and students to effectively plan and focus their teaching and learning. For example, in the present study, it appears from CDM results that students were deficit in dealing with the attribute "Model type". Teachers and students can know that this skill is the priority for their learning. With the aid of such diagnostic information, instruction can be more targeted and personalized.

CONCLUSION

The three papers represent the "*Innovative Assessments that Support Students' STEM Learning*" by operating the integrative framework for CDM based on the components of cognition, observation, and interpretation. Specifically, this dissertation demonstrates how this framework combines the sciences of psychometrics and cognitive psychology, and utilizes the integrative nature of those three components in science and math assessments. Given this, the dissertation provides useful knowledge for the measurement contributions and assessment implications.

Given the theory and application of CDM in science and math education, the dissertation provides a concrete example of an innovative assessment that support students' learning in STEM. Back to the first research question, *What is the integrative framework for CDM?*, the conceptual framework based was developed. The integrative framework for CDM consists of two elements: CDM as theory-driven model and statistical procedures. It intends

to improve the capacity of CDM to increase the formative values of assessment and validate cognitive/learning theory.

This framework not only stress the two elements, but also highlight the importance of integrating cognitive models and stat in the entire procedure of the CDM analysis. As demonstrated in this dissertation that both cognition model and statistics are integrated in the whole procedure of the work. Then this can fulfill the innovation definition based on the assessment triangle. The interpretation of observation is limited if the assessment tasks are not well-designed for measuring cognition or cognition model is not well-developed. Hence, this integrative framework essentially shows how those three elements are interconnected, especially the path between cognition and interpretation.

To answer the second research question, *How can this CDM integrative framework be applied to develop and validate STEM assessments?*, two research studies in science and math were demonstrated. Basically, the framework can help construct the items or assessment, provide targeted information, and facilitate assessment data interpretation. It can be used for many situations such as examining students' mastery level and learning gain as useful information for diagnostic and formative purposes and estimating item parameters for investigating item features related to students' performance that can guide assessment practices. Specifically, it can be used to provide the indices of the instructional sensitivity of assessment items which can be used as indicators for further item development and interpretation. Moreover, it can be used to examine the relationship of item characteristics and item parameters that can guide the assessment development and interpretation.

Who is the user of this framework? Since this framework aims to support the development and validation of STEM assessments, measurement research scientist, assessment specialist, and learning scientist should be the ones who mainly use this framework. As the framework is integrative by nature, it requires collaboration of the team

with at least measurement specialist and cognitive scientist or teachers. Then, the work can reach the full potential of CDM and fulfill the needs of innovative assessments that can support students' learning.

FUTURE DIRECTIONS OF RESEARCH

This dissertation raises several future research directions. First, how do we employ more complex CDM models with the data? In this current research, there are several issues such as sample size, number of items, number of attributes, and number of missing values, which are concerned in the selection of the model used in the analysis. For the purpose of exploratory analysis, the DINA model is sufficient. In the future, a larger sample size should be used to allow using more complex models, e.g., HO-DINA (de la Torre, & Douglas, 2004) which can reflect the sequential structures of attributes or G-DINA (de la Torre, 2011) which can provide the probability of attribute mastery patterns. So the analysis can provide more valid and robust measurement results.

Second, how do we use the capacity of this framework to explore common constructs that can be applied in cross-disciplines to understand the similarities and differences of those constructs across disciplines? Specifically, I would like to continue my current work by studying diagnostic assessments and constructs that are stressed in the CCSS and NGSS, yet have not been adequately captured in commonly used assessments, for example, modeling, problem-solving strategies, and understanding and application of cross-cutting concepts. Generating the attributes and Q-matrix for these complex constructs can involve potential challenges. For example, what should be the fine grained size of the attributes, what should be the moderate numbers of the attributes that can yield the valid assessment results, and how to make sure that the Q-matrix for the complex constructs are correct. Moreover, we need to

create innovative tasks for assessing constructs and eventually produce validated assessments that can be used for further research and in practical settings.

Third, how to incorporate this integrative framework with the learning technology for both formal and informal learning environments, for example, tablet- and game-based assessment to provide diagnostic and formative information to support student learning? Learning analytics can help validate the CDM results by providing more detailed information of the attributes. Based on the validated assessment information, these methods can shape the assessment design and thus guide the technology use. Such methods can help provide individualized assessment and instruction, which can in turn allow us to provide adaptive support for individual students, especially those with additional needs (e.g., English language learners).

Last but not least, how do we combine machine learning and psychometric methods to mine game data (or online instruction data) to model the growth of student learning, particularly based on learning progression? These methods will allow us to better track student learning progress and use the assessment information for both formative (to adjust instruction to individual student learning plans) and summative (to compare student performance with benchmarks) purposes.

REFERENCES

Airasian, P. W., & Madaus, G. F. (1983). Linking testing and instruction. *Journal of Educational Measuremen*t, *20*, 103-118.

Alvers, C. (2012). *Making diagnostic inferences about student performance on the Alberta Education Diagnostic Mathematics Project: An application of the attribute hierarchy method*. Ph.D. dissertation, University of Alberta (Canada), Canada. Retrieved April 22, 2012, from Dissertations & Theses: Full Text.(Publication No. AAT NR81451).

Ananda, S., & Rabinowitz, S. (2001). *High-stakes and assessment innovation: A negative correlation?* San Francisco, CA: WestEd. (ERIC Document Reproduction Service No. ED462446).

Ayers, E., Rabe-Hesketh, S., & Nugent, R. (2013) Incorporating student covariates in cognitive diagnosis models. *Journal of Classification, 30,* 195-224.

Baker, E. L. (1994). Making performance assessment work: The road ahead. *Educational Leadership*, 51(6), 58-62.

Barbosa, J. C. (2006). Mathematical modelling in classroom: A socio-critical and discursive perspective. *Zentralblatt für Didaktik der Mathematik, 38*(3), 293–301.

Barbu, O. & Beal, C. R. (2010). Effects of linguistic complexity and math difficulty on word problem solving by English Learners. *International Journal of Education, 2*, 1-19.

Black, P. (2001). Dreams, strategies and systems: Portraits of assessment past, present and future", *Assessment in Education: Principles, Policy and Practice, 8*(1), 65-85.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). The nature and value of formative assessment for learning. *Improving Schools, 6*, 7-22.

BrainPOP. (n.d.). *Riddle Books step guide*. Retrieved December 20, 2016 from https://cdn-educators.brainpop.com/wp-content/.../01/RiddleBooksStepGuide-2.pdf

Briddle, B.J. (1997). Foolishness, dangerous nonsense, and real correlates of state differences in achievement. Phi Delta Kappan, 79(1). Retrieve January 4, 2007 from http://0-web.ebscohost.com.skyline.cudenver.edu/ehost/detail?vid=2&hid=117&sid=9fa94981-b195-42fe-bbc2-08fd5b9b01ba%40sessionmgr106.

Broaddus, A., & Shaftel, J. (2012). *Cognitive diagnostic assessment - informing responses and interventions.* Retrieved April 10, 2015 from https://cete.ku.edu/sites/cete.drupal.ku.edu/files/docs/Presentations/2012_05_Broaddus_Shaftel%20CDA-RtI.pdf.

Burstein, L., Aschbacher, P., Chen, Z., & Lin, L. (1990). Establishing the content validity of tests designed to serve multiple purposes: Bridging secondary-postsecondary mathematics. CSE Technical Report 313. Center for Research on Evaluation, Standards, and Student Testing. Los Angeles, CA: University of California, Los Angeles.

Carpenter, T. P., Hiebert, J., & Moser, J. M. (1981). Problem structure and first grade children's initial solution processes for simple addition and subtraction problems. *Journal for Research in Mathematics Education, 12*, 27-39.

Chappuis, S., &Chappuis, J. (Dec 2007-Jan 2008). The best value in formative assessment. *Educational Leadership*, 65(4), 14-19.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*, 123-140.

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

Clement, L. L., & Bernhard, J. Z. (2005). A problem-solving alternative to using key words. *Mathematics Teaching in the Middle School, 10*(7), 360-365.

Common Core State Standards Initiative. (2010). *Common Core State Standard for Mathematics*. Retrieved April 5, 2015 from http://www.corestandards.org/wp-content/uploads/Math_Standards.pdf

Common Core State Standards Initiative. (2010). *Common Core State Standard for Mathematics*. Retrieved April 5, 2015 from http://www.corestandards.org/wp-content/uploads/Math_Standards.pdf

Commission on Instructionally Supportive Assessment. (2001). *Building tests to support instruction and accountability: A guide for policymakers* (James Popham, Commission Chair). Washington, DC: Author. Retrieved December 23, 2007, from http://www.testaccountability.org/.

Cox, R. C., & Vargas, J. S. (1966). *A comparison of item selection techniques for norm referenced and criterion referenced tests*. Internal Manuscript: University of Pittsburg.

Cronbach, L. J. (1988). *Five perspectives on validity argument.* In H. Wainer & H. Braun (Eds.). Test validity (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Darling-Hammond, L. (1999). *Teacher quality and student achievement: A review of state policy evidence* (Document R-99-1). Retrieved January 4, 2008, from the Center for the Study of teaching and Policy Web site: http://www.ctpweb.org.

Darling-Hammond, L. (2007). Race, inequality, and educational accountability: The irony of 'No child Left Behind'. *Race Ethnicity and Education*, *10*(3), 245-260.

Davey, T. (2011). *A guide to computer adaptive testing systems*. Retrieved August 1, 2016 from http://www.ccsso.org/Documents/2011/Guide_to_Computer_Adaptive_2011.pdf

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*, 8-26.

Debrenti, E. (2015). Visual representations in Mathematics teaching: An experiment with students. *Acta Didactica Napocensia, 8*(1), 19-25.

de la Torre, J. (2009a). A cognitive diagnosis model for cognitively-based multiple-choice options. *Applied Psychological Measurement, 33*, 163-183.

de la Torre, J. (2009b). DINA model and parameter estimation: a didactic. *Journal of Educational and Behavioral Statistics, 34*, 115–130.

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

de la Torre, J. (2014). Cognitive diagnosis modeling: A general framework approach. Retrieved September 10, 2014 from http://rci.rutgers.edu/~jdelator/Download/handout.pdf.

de la Torre, J., & Chiu, C. Y. (2015). General empirical method of Q-Matrix validation. *Psychometrika*. Advance online publication. doi:10.1007/s11336-015-9467-8.

de la Torre, J., & Douglas, J. (2004). A higher-order latent trait model for cognitive diagnosis. *Psychometrika*, *69*, 333–353.

de la Torre, J., & Douglas, J. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*, 595-624.

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*, 227-249.

de la Torre, J., & Lee, Y.-S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*(4), 355-373.

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa, 20*, 89-97.

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

DiBello, L.S., & Stout, W. (2003). Student profile scoring for formative assessment. In H.Yanai, A. Okada, K. Shigemasu, Y. Kano, &J. J. Meulman (Eds.), *Proceedings of the International Meeting of the Psychometric Society IMPS2001* (pp. 81-92). Osaka, Japan, July 15–19, 2001.

DiBello, L.S., Stout, W., & 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). Mahwah, 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.

Doornik, J. A. (2002), Object-oriented matrix programming using Ox, 3rd ed. London: Timberlake Consultants Press and Oxford: www.nuff.ox.ac.uk/Users/Doornik.

Edens, K. & Potter, E. (2008). How students "unpack" the structure of a word problem: Graphic representations and problem solving. *Science and School Mathematics Journal, 108*(5), 184-193.

Edwards, S. A., Maloy, R. W., & Anderson, G. (2009). *Reading coaching for math word problems*. Urbana, IL: Literacy Coaching Clearinghouse.

Embretson S. E., & Yang X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika*, *78*, 14-36.

Falchikov, N. & Thompson, K. (2008). Assessment: what drives innovation? *Journal of University Teaching & Learning Practice, 5*(1), 49–60.

Feasel, K., Henson, R., & Jones, L. (2004). *Analysis of the Gambling Research Instrument (GRI)*. Unpublished manuscript.

Feng, Y. (2013). *Estimation and Q-matrix validation for diagnostic classification models*. (Doctoral dissertation). University of South Carolina.

Ferguson, R. (2012). *The state of learning analytics in 2012: A review and future challenges.* Technical Report KMI-12-01, Knowledge Media Institute, The Open University. kmi.open.ac.uk/publications.

Frejd, P. (2011*). Mathematical modelling in upper secondary school in Sweden an exploratory study.* Licentiate thesis. Linköping: Linköpings universitet.

Gabadinho, A., Müller, N. S., Ritschard, G., & Studer, M. (2011). *Mining sequence data in R with the TraMineR package: A user's guide (for version 1.4-2)*. Retrieved May 15, 2016 from http://mephisto.unige.ch/pub/TraMineR/doc/1.4/TraMineR-1.4-Users-Guide.pdf

Giamellaro, M., Lan, M-C., Ruiz-Primo, M. A., Li, M., & Tasker, T. (2011, April). *Mapping science curricula: A method for supporting teachers in the articulation of learning goals*. Paper presented at the American Educational Research Association. New Orleans, LA.

Giamellaro, M., Ruiz-Primo, M. A., Li, M. (2012, March). *Quality teaching as reflected in productive failure*. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Indianapolis, IN.

Goldin, G. A., Kaput, J. (1996). A joint perspective on the idea of representation in learning and doing mathematics. In: Steffe, L., Nesher, P., Cobb, P., Goldin, G. & Greer, B. (Eds.), *Theories of mathematical learning, Hillsdale Erlbaum*. PP. 397-430.

Gu, Z. (2011). *Maximizing the Potential of Multiple-Choice Items for Cognitive Diagnostic Assessment*. (Doctoral Dissertation). University of Toronto, Ontario, Canada.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333-352.
Haertel, E. H. (2013), *'Reliability and Validity of Inferences about Teachers Based on Student Test Scores'.* Report based on the 14th William H. Angoff Memorial Lecture at the National Press Club, Educational Testing Service, Princeton, NJ.
Hartz S. A. (2002). *Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. (Doctoral Dissertation). Urbana-Champaign: University of Illinois.
Henson, R, & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*, 262–277.
Heritage, M. (2007). Formative assessment: What do teachers need to know and do? *Phi Delta Kappa, 89*(02), 140-145.
Heritage, M. (2010). *Formative assessment and next-generation assessment systems: Are we losing an opportunity?* A project of Margaret Heritage and the Council of Chief State School Officers (Paper prepared for the Council of Chief State School Officers). Washington, DC: Council of Chief State School Officers.
Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnosis modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement, 51*(1), 98-125.
Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 117-131). Ames, IA: Iowa State University.
Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for applying Fusion Model to LanguEdge assessment. *Language Testing, 26*(1), 31–73.
Jeong, A. (2005). A guide to analyzing message-response sequences and group interaction patterns in computer-mediated communication. *Distance Education, 26*(3), 367–383.
Junker, B. W. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment.* Unpublished manuscript. Prepared for the Committee on the Foundations of Assessment, National Research Council, November 30, 1999. Accessed June 9, 2014 from http://www.stat.cmu.edu/~brian/nrc/cfa/documents/final.pdf.
Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.
Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*, 527–535.
Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 21 17–64). Westport, CT: American Council on Education and Praeger.
Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.
Kho, T. H. (1987). Mathematical models for solving arithmetic problems. *In Proceedings of the fourth Southeast Asian conference on mathematical education (ICMI-SEAMS).Mathematical education in the 1990's Vol. 4*, (pp. 345–351). Singapore: Institute of Education.
Koedinger, K. R. & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *The Journal of the Learning Sciences, 13*(2), 129-164.

Koshy, S. (2013), *'I have a dream … for innovative assessments', World Academy of Science, Engineering and Technology,* Retrieved March 1, 2016 from http://ro.uow.edu.au/cgi/viewcontent.cgi?article=1608&context=dubaipapers

Küchemann, D., Hodgen, J., & Brown, M. (2011). Models and representations for the learning of multiplicative reasoning: Making sense using the Double Number Line. *Proceedings of the British Society for Research into Learning Mathematics* 31(1) March 2011.

Lan, M-C., Li, M., Ruiz-Primo, M. A., Wang, T., Giamellaro, M., & Mason, H. (2012). Linking quality of instruction to instructionally sensitive assessments. Paper presented at the American Educational Research Association. Vancouver, BC, Canada.

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing*, *11*, 144-177.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *EducationalMeasurement: Issues and Practice, 26,* 3-16.

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41,* 205-237.

Leinhardt, G. (1983). Novice and expert knowledge of individual student's achievement. *Educational Psychologist*, *18*, 165-179.

Lesieur, H. R., & Blume, S. B. (1987). The South Oaks Gambling Screen (SOGS): A new instrument for the identification of pathological gamblers. *American Journal of Psychiatry, 144*(9), 1184-1188.

Lewis, A. B. & Mayer, R. E. (1987). Students miscomprehension of relational statements in arithmetic word problems. *Journal of Educational Psychology, 79*, 363-371.

Li, M., Lan, M-C., Ruiz-Primo, M. A., Giamellaro, M., & Wang, T. (2012, March). *Supporting students to make conceptual connections*. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Indianapolis, IN.

Li, H., & Suen, H. K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment, 18*(1), 1–25.

Linacre, L. M. (n.d.). *Computer-adaptive testing: A methodology whose time has come*. Retrieved August 1, 2016 from http://www.rasch.org/memo69.pdf.

Lingefjäerd, T. (2002). Teaching and Assessing Mathematical Modelling. *Teaching Mathematics and Its Applications, 21*(2), 75-83.

Linn, R. L., & Harnisch, D. L. (1981) Interactions between item content and group membership. *Journal of Educational Measurement, 18,* 109-118.

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement. 33,* 579-598

Llinares, S. & Roig, A. I. (2008). Secondary school students' construction and use of mathematical models in solving word problems. *International Journal of Science and Mathematics Education, 6*, 505-532.

Llinares, S. & Roig, A. I. (2008). Secondary school students' construction and use of mathematical models in solving word problems. *International Journal of Science and Mathematics Education, 6*, 505-532.

Looney, J. (2009). *Assessment and Innovation in Education.* Retrieved March 1, 2015 from http://www.oecd.org/edu/43338180.pdf

Ma, L. (2014). *Validation of the item-attribute matrix in TIMSS-Mathematics using multiple regression and the LSDM*. (Doctoral dissertation). University of Denver.

Ma, W., & de la Torre, J. (2016). *GDINA R package*. Unpublished R package.
Madaus, G. F., Airasian, P. W., & Kellaghan, T. (1980). *School effectiveness: A reassessment of theevidence.* New York: McGraw-Hill.
Madaus, G. F., Airasian, P. W., & Kellaghan, T. (1980). *School effectiveness: a reassessment of the evidence.* New York: McGraw-Hill.
Maloney, D. P., & Siegler, R. S. (1993). Conceptual competition in physics learning. International *Journal of Science Education, 15*, 283-295.
Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 197-212.
Martin, S. & Bassok, M. (2005). Effects of semantic cues on mathematical modeling: Evidence from word problem solving and equation construction. *Memory & Cognition, 33*, 471-478.
Mason, H., Ruiz-Primo, M. A., Giamellaro, M., & Li, M. (2012, March). *What do students' science notebooks reflect about the quality of teaching students receive?* Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Indianapolis, IN.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: The American Council on Education & the National Council on Measurement in Education.
Mislevy, R. J. *(*1996*).* Test theory reconceived. *Journal of Educational Measurement, 33*(4), 379-416.
Mislevy, R.J. et al. (1998). *A cognitive task analysis, with implications for designing a simulation based performance assessment*. University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.
Mislevy, R.J. et al. (2001). *Making sense of data from complex assessments.* University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.
Mowl, G. (2006). Red guides, paper 17: Innovative student assessment: what's the point? Northumbria University. Retrieved March 15, 2016 from https://www.northumbria.ac.uk/static/5007/arpdf/academy/redguide17.pdf
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: IEA.
Murata, A., & Kattubadi, S. (2012). Grade 3 students' mathematization models and solution models with mutli-digit subtraction problem solving. *Journal of Mathematical Behavior, 31*, 15–28.
Nathan, M. J., Kintsch, W., and Young, E. (1992). A theory of algebra-word-problem comprehension and its implications for the design of learning environments. *Cognition and Instruction, 9*(4), 329–389.
National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.
National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, D.C.: National Academy Press.
Ng, K. H. R., Hartman, K., Liu, K., & Khong, A. W. H. (2016). *Modelling the way: Using action sequence archetypes to differentiate learning pathways from learning outcomes*. In Proceedings of the 9th International Conference on Educational Data Mining. Retrieved August 5, 2016 from http://www.educationaldatamining.org/EDM2016/proceedings/paper_154.pdf.
Ng, S. W., & Lee, K. (2005). How primary five pupils use the model method to solve word problems. *The Mathematics Educator, 9*(1), 60-83.

Ng, S. W., & Lee, K. (2009). The model method: Singapore children's tool for representing and solving algebraic word problems. *Journal for Research in Mathematics Education, 40*(3), 282-313.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999), "Addressing the 'Two Disciplines' problem: Linking theories of cognition with assessment and instructional practice. *Review of Research in Education, 24,* 307-353.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments, *Educational Measurement: Issues and Practices*, *29*(4), 3-14.

Popham, W. J. (2006). Diagnostic assessment a measurement mirage? *Educational Leadership*, *64*(2), 90-91.

Popham, W. J. (2007a). *Conducting instructional-sensitivity reviews of educational accountability tests*. Unpublished paper. Los Angeles, CA: University of California, Los Angeles.

Popham, W. J. (2007b). Instructional sensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan*, *89*(2), 146-150, 155.

Popham, W. J., Keller, T. Moulding, B., Pellegrino, J., & Sandifer, P. (2005). Instructionally supportive accountability tests in science: A viable assessment option? Measurement, 3(3), 121-179.

Rabinowitz, S. N., & Brandt, T. (2001). *Computer-based assessment: Can it deliver on its promise?* San Francisco, CA: WestEd.

Reed, S. K. (1999). *Word problems. Research and curriculum reform*. Mahwah, NJ: Lawrence Erlbaum

Rojas, G., de la Torre, J., & Olea, J. (2012, April). *Choosing between general and specific cognitive diagnosis models when the sample size is small*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

Roussos, L., DiBello, L., Henson, R., Jang, E. E., & Templin, J. (2010). Skills diagnosis for education and psychology with IRT-based parametric latent class models. In S. E. Embretson & J. Roberts (Eds.), *New directions in psychological measurement with model-based approaches* (pp. 35-69). Washington, DC: American Psychological Association.

Roussos, L., Templin, J., & Henson, R. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement, 44*(4), 293-311.

Ruiz-Primo, M. A., & Li, M. (2012). *Assessing transfer of learning: Instructionally sensitive assessments, curriculum, and instruction*. Paper presented at the AERA annual meeting, Vancouver, Canada.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic education reform: Searching for instructional sensitivity. *Journal of Research in ScienceTeaching*, *39*(5), 369-393.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*, 219.

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

Sajadi1, M., Amiripour, P., & Rostamy-Malkhalifeh, M. (2013). The examining mathematical word problems solving ability under efficient representation aspect. *Mathematics Education Trends and Research, 2013*, 1-11.

Schuster, J. & Erickson, K. (2014). *Text Complexity in the Dynamic Learning Maps™ Alternate Assessment System (White Paper No. 14-01)*. Lawrence, KS: University of Kansas Center for Educational Testing and Evaluation.

Shum, S.B., & Crick, R.D. (2012). *Learning dispositions and transferable competencies: Pedagogy, modelling and learning analytics.* Proc. 2nd International Conference on Learning Analytics & Knowledge, (29 Apr-2 May, Vancouver, BC). ACM Press: New York. Retrieved March 1, 2016 from http://projects.kmi.open.ac.uk/hyperdiscourse/docs/LAK2012-SBS-RDC.pdf

Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. *Educational Researcher, 37*(4), 189–199

Songer, N. B., & Ruiz-Primo, M. A. (2012). Assessment and science education: Our essential new priority? *Journal of Research in Science Teaching, 49*(6), 683–690.

Stiggins, R. J. (2008). *Assessment manifesto: A call for the development of balanced assessment systems.* Portland, OR: ETS Assessment Training Institute.

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

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale, NJ: Erlbaum.

Templin, J., & Henson, R. A. (2005). *The random effects reparametrized unified model: A model for joint estimation of discrete skills and continuous ability*. Unpublished manuscript.

Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11,* 287–305.

Templin, J., Henson, R., Rupp, A., Jang, E., & Ahmed, M. (2008, March).*Cognitive diagnosis models for nominal response data.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York City, NY.

Templin, J., Henson, R., Templin, S., & Roussos, L. (2008). Robustness of unidimensional hierarchical modeling of discrete attribute association in cognitive diagnosis models. *Applied Psychological Measurement, 32,* 559-574.

Thummaphan, P., Dong, D., & Li, M. (2015, August). *The use of cognitive diagnosis modeling analysis of word problem solving*. Paper presented in the 12th International Postgraduate Research Colloquium (IPRC), Bangkok, Thailand.

Toutkoushian, R. K., & Curtis, T. (2005). Effects of socioeconomic factors on public high school outcomes and rankings. *The Journal of Educational Research, 98*(5), 259-270.

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

Verschaffel, L., De Corte, E., & Lasure, S. (1994). Realistic considerations in mathematical modelling of school arithmetic word problems. *Learning and Instruction, 4*, 273-294.

Vos, P. (2013). Assessment of modelling in mathematics examination papers: Ready-made models and reproductive mathematising. In G.A. Stillman, G., Kaiser, W., Blum, J.P.Brown, (eds.), *Teaching Mathematical Modelling: Connecting to Research and Practice, International Perspectives on the Teaching and Learning of Mathematical Modelling*, DOI 10.1007/978-94-007-6540-5_41

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2014), *The Log-Linear Cognitive Diagnostic Model (LCDM) as a Special Case of the General Diagnostic Model (GDM)*. ETS Research Report Series. 2014(2), 1–13, Retrieved April 20, 2015 from http://onlinelibrary.wiley.com/doi/10.1002/ets2.12043/abstract

von Davier, M., & Yamamoto, K. (2004, October). *A class of models for cognitive diagnosis*. Paper presented at the 4th Spearman Conference, Philadelphia, PA.

Wang, T., Lan, M-C., Giamellaro, M., Zhao, D. Y., Birkby, E., Ruiz-Primo, M. A., & Li, M. (2012, March). Knowledge of learning goals as a navigation tool in curriculum implementation. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Indianapolis, IN.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*(3), 461-481.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance.* San Francisco, CA: Jossey-Bass.

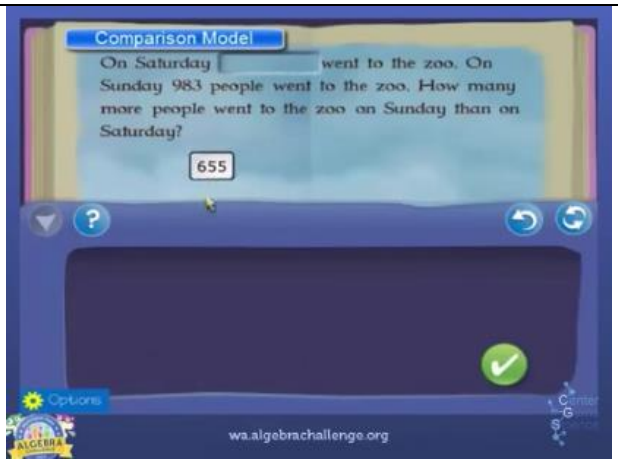
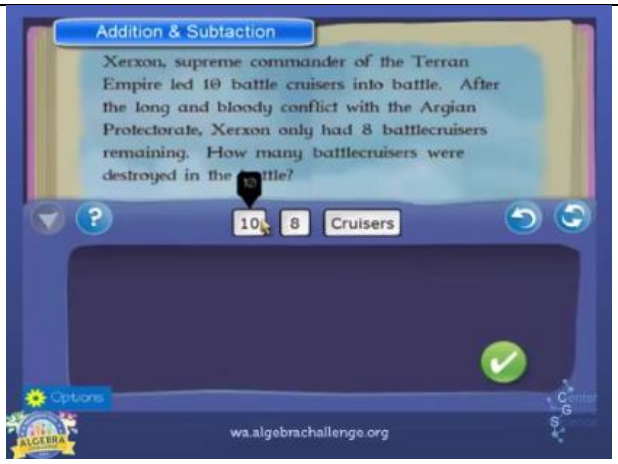
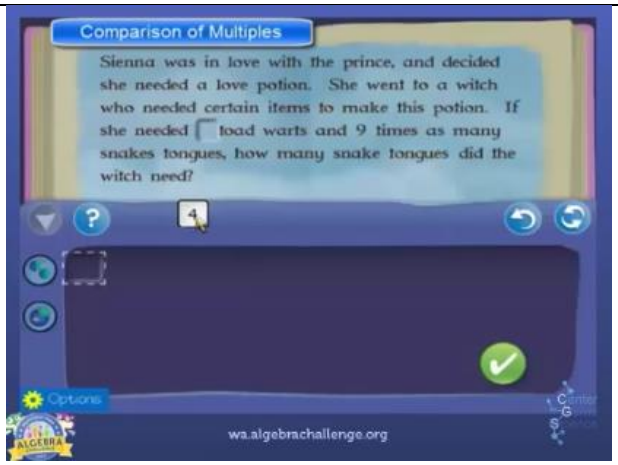
Wiliam D. (2007, September). *Sensitivity to instruction: The missing ingredient in large-scale assessment systems?* Paper presented at the Annual Meeting of the International Association for Educational Assessment. Baku, Azerbaijan.

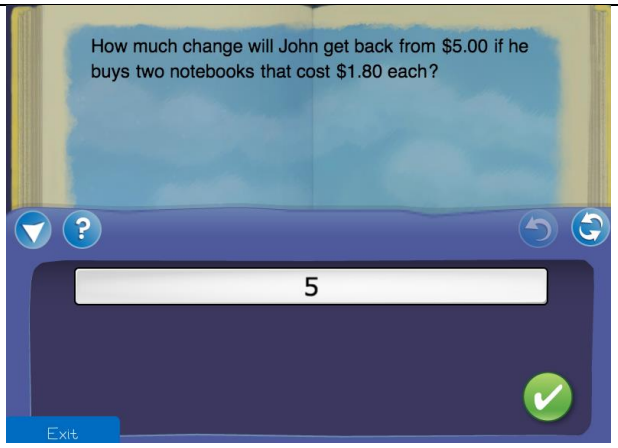
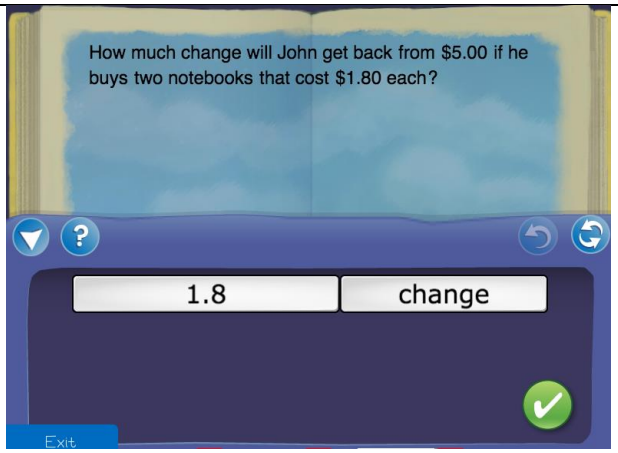
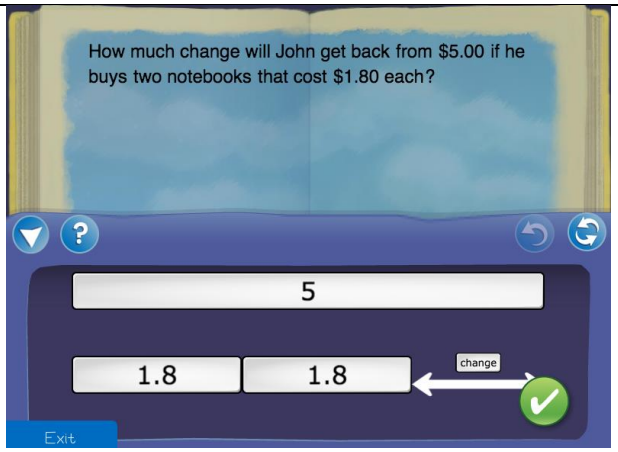
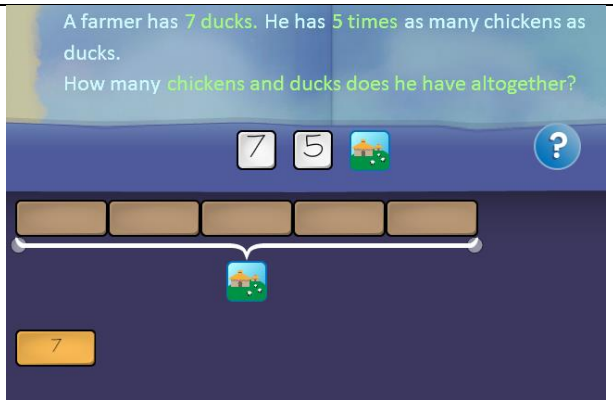
Xu, X., & von Davier, M. (2006). *Cognitive diagnosis for NAEP proficiency data (Research Report No. RR-06–08).* Princeton, NJ: Educational Testing Service.

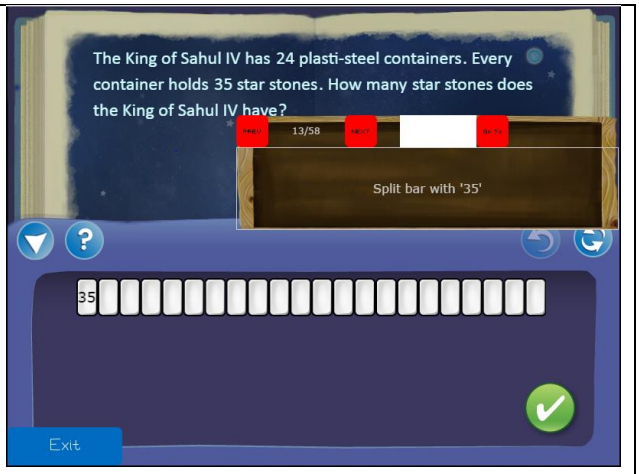
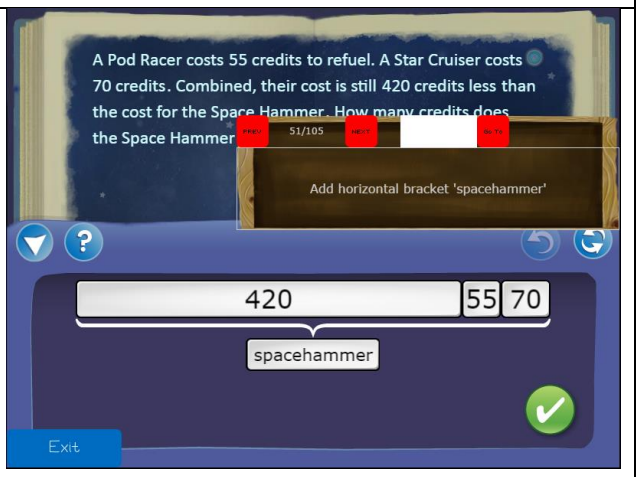
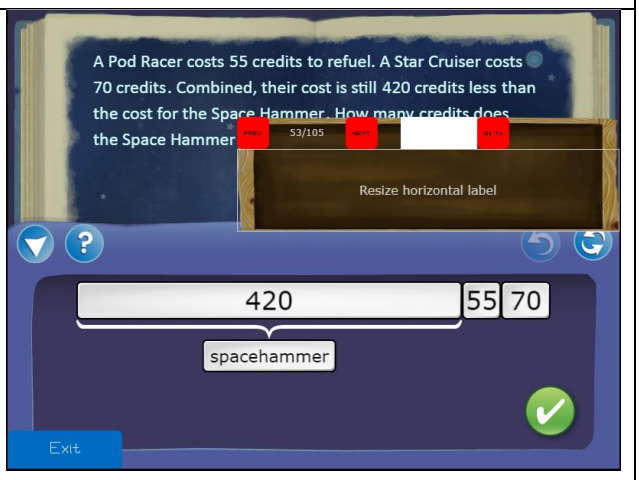
Yan, D., Mislevy, R. J., & Almond, R. G. (2003). *Design and analysis in a cognitive assessment (Research Report No. RR-03–32).* Princeton, NJ: Educational Testing Service.

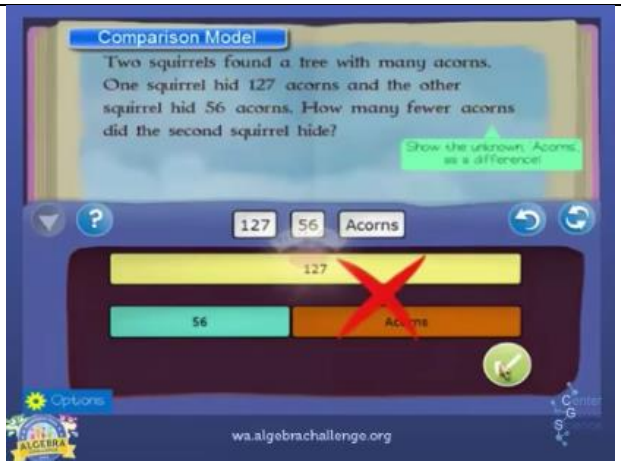
APPENDIX

Appendix A. Descriptions and examples of actions for the first level of the modeling cycle

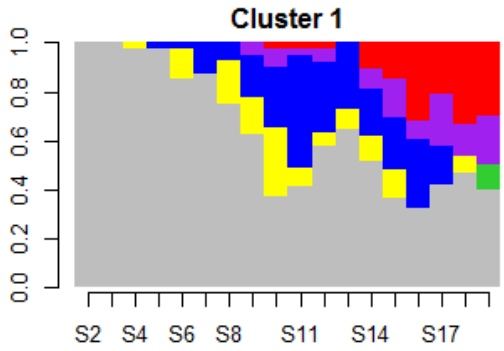
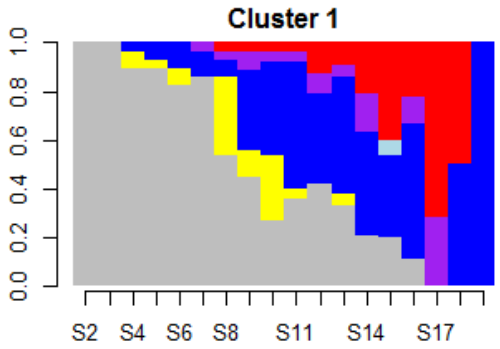
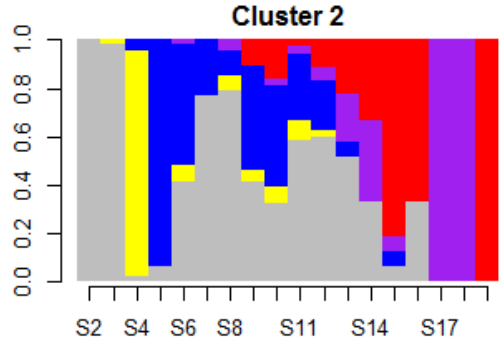
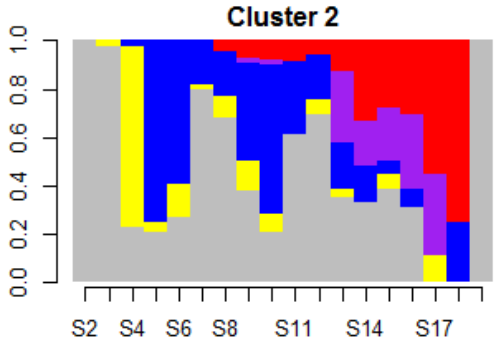
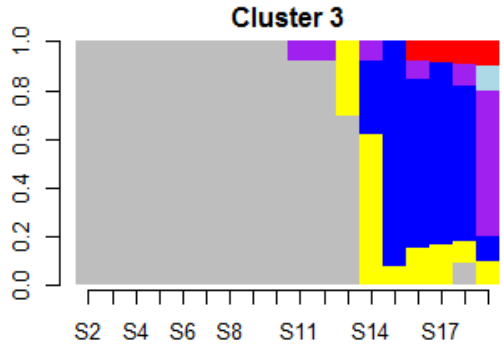
Actions	Descriptions	Example
1	**Start:** Students click a bottom to start with the item.	
2	**Extract information:** This step involves multiple actions related to extracting information ***Pick info from text:*** The mathematical information in the text is clickable. Students can pick up a number and variables from the text or deck by dragging the box of that number.	
	Pick info from the deck: Deck is the area where you see the arrow points to in the figure on the right. It is used to store values that students put there so that when students construct models, they can easily fetch the numbers from the deck instead of going back to the text.	
	Drop info Elsewhere: Students pick up a number and drop it somewhere else other than the modeling area. Usually it indicates students either put the number back to the text or drop it on the deck.	

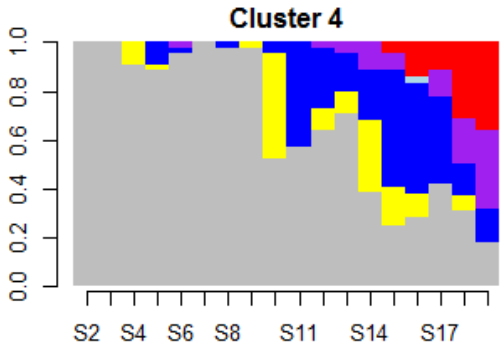
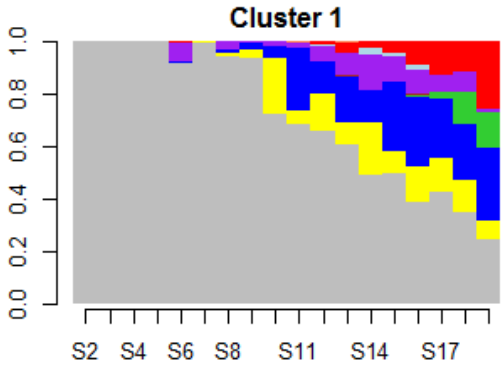
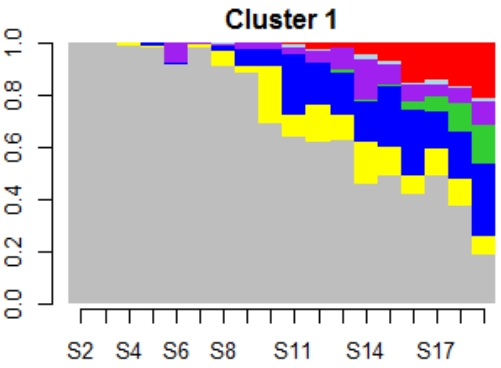
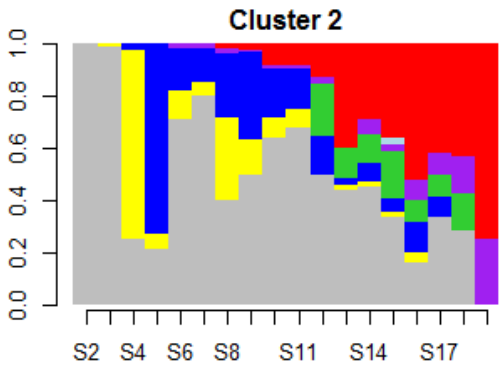
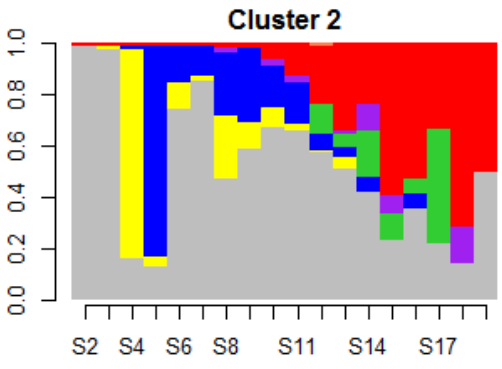
3	**Add row:** Students build a first bar in a row to create the model which can be either the first bar in the modeling area or the first bar in another row.	
4	**Add bar:** Add new bar segment to a row. In this case a student build a new bar with unknown (the unknown in this item is named "change") in the existing bar.	
5	**Subtract:** Add subtraction part to the bar model.	
6	**Add unit bar:** Add many new boxes at once to a row; in the picture it was five grey boxes on the first row in the modeling area.	

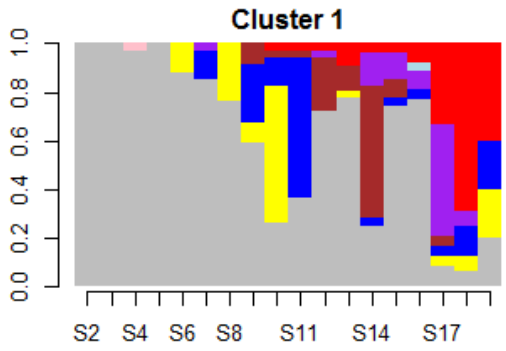
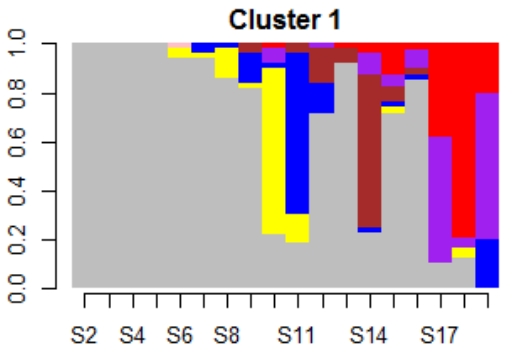
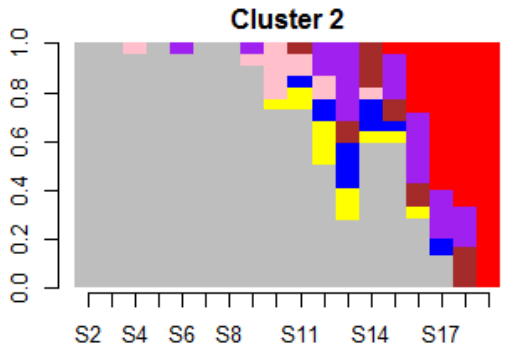
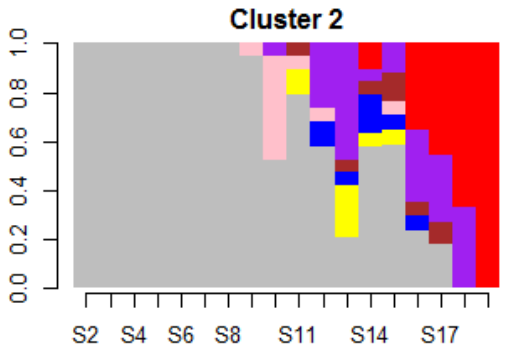
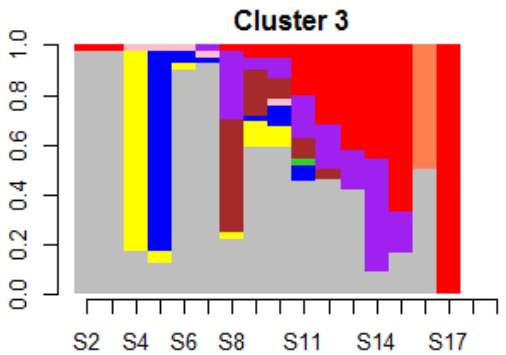
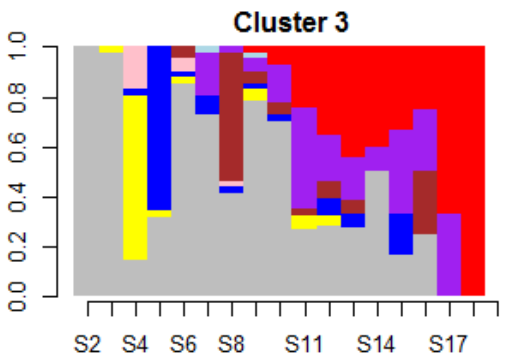
7	**Divide:** Divide box into smaller equal parts. In the picture it is split each bar with 35.	
8	**Add label:** Add the bracket to bar model. In the picture it is the label of "spacehammer" for three bars.	
9	**Resize label:** Change size of the bracket. In the picture it is the decrease of the "spacehammer" size to cover only the bar of 420.	

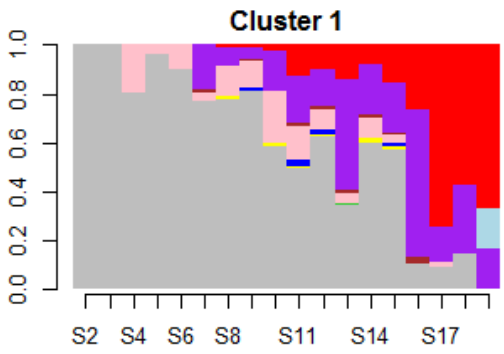
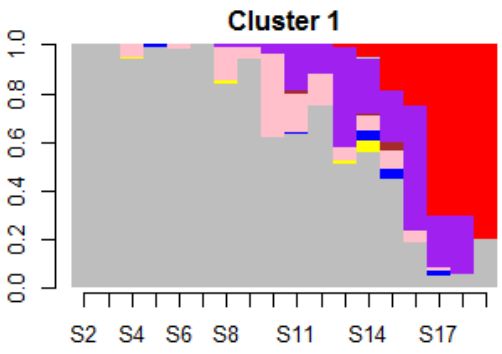
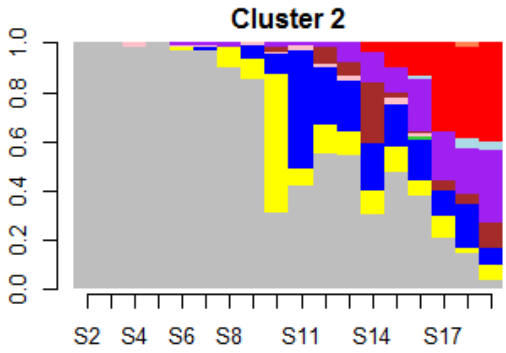
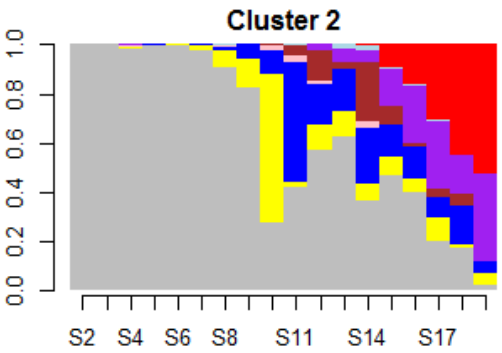
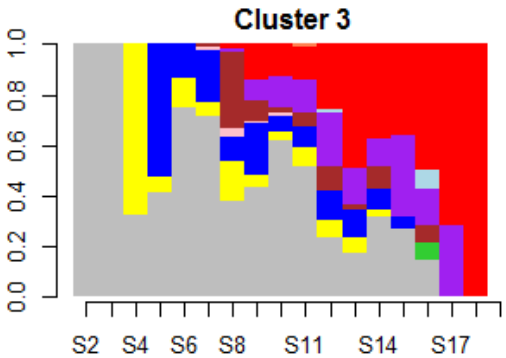
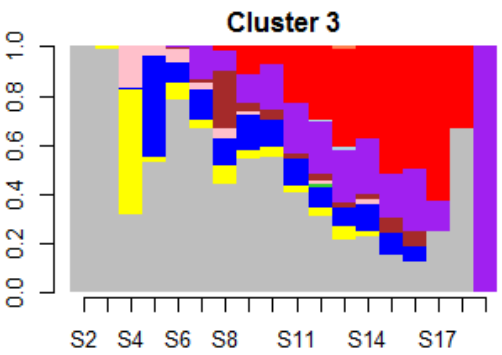
10	**Validate model:** Students checking if bar model is correct. Students can do this many times. In the picture the model is incorrect.	
11	**Quit/Skip:** Any quit or skip level actions, did not play later Skip level, played again later	

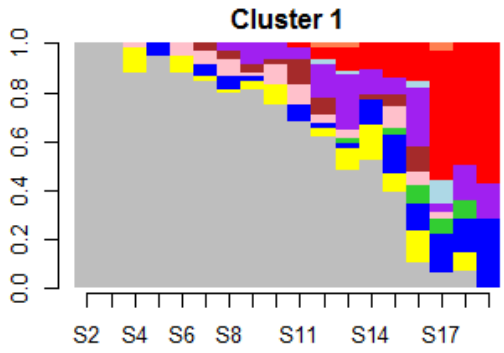
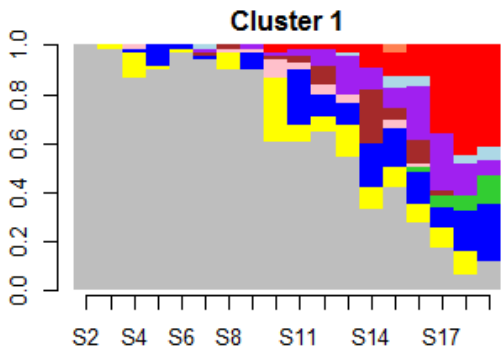
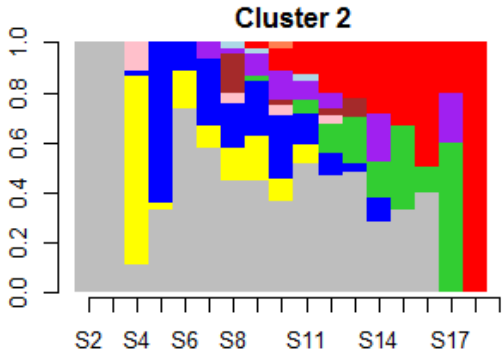
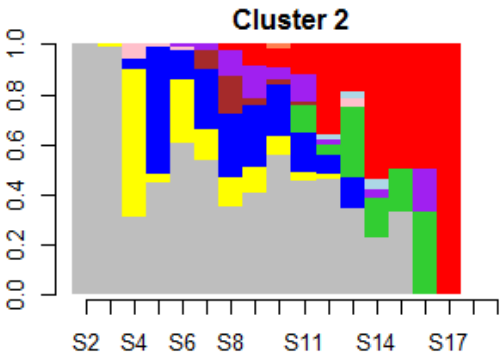
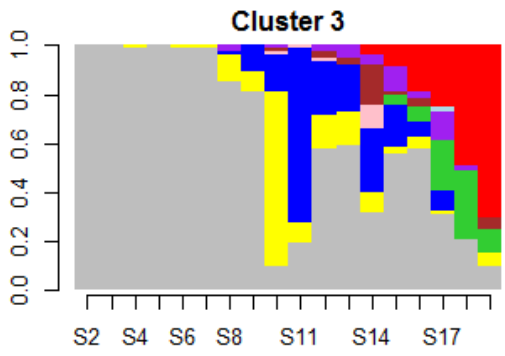
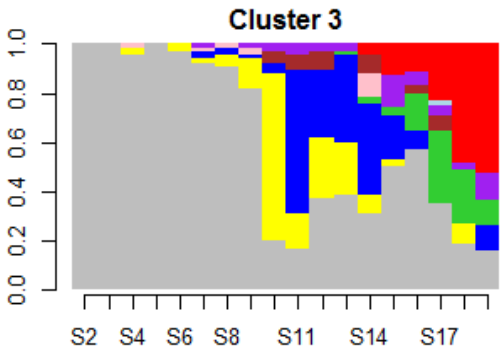
Appendix B. Cluster Analysis Results Ordered with Percentage of Correct by the Similarity

Cluster	Booklet 1	Booklet 2
Group 1: Model Type 1a.		
1	27.50% 	21.43% 
2	30.43% 	54.54% 
3	76.92% 	

4		29.54% 
Group 2: Model Type 2a.		
1	38.41% 	37.98% 
2	43.75% 	37.60% 

Group 3: Model Type 3a.		
1	35.29% 	36.00% 
2	36.36% 	47.37% 
3	25.00% 	36.59% 

Group 4: Model Type 4a.		
1	64.47% 	64.00% 
2	64.07% 	55.82% 
3	45.57% 	65.28% 

Group 5: Model Type 4c.		
1	18.64% 	17.65% 
2	48.89% 	33.80% 
3	46.57% 	69.23% 

Note: 'gray' = 'Extract info', 'yellow' = 'Add row', 'blue' = 'Add bar', 'lime green' = 'Subtract', 'pink' = 'Add unit bar', 'brown' = 'Divide', 'purple' = 'Add label', 'light blue' = 'Resize label', 'red' = 'Validate', 'coral' = 'Quit/Skip'