SU, KUN. Ph.D. Implementation of a Diagnostic Classification Model for Middle-School Physics. (2022)
Directed by Dr. Robert Henson. 111 pp.

This dissertation provides a start-to-finish description of development, administration, and validation for an online middle-school physics test using a DCM framework with response-time. The first paper illustrated the process of implementing DCM with a careful selection of the content domain and a simulation approach for a Q-matrix construction. The results were promising despite some items that showed inadequate fit and quality. The second paper is a narration of a step-by-step validation process for the effects of the DCM-scored physics assessment on learning and teaching. While evidence was found to support multiple validity arguments and the usefulness of the diagnostic feedback, validity threats were also found because one of the students identified multiple strategies in solving some of the questions. The third paper investigates the potential benefits of incorporating response-time into the diagnostic model. Although the response-time variable did not improve the classification estimation in this case, different types of relations between the ability, time variable, or other ancillary variables should be investigated in the future. Although limitations were found in the dissertation, multiple actions could be taken to refine this process in future research, and the process could still be generalized into other domains and as guidelines for researchers and educators interested in DCM application.

IMPLEMENTATION OF A DIAGNOSTIC CLASSIFICATION MODEL FOR MIDDLE-SCHOOL PHYSICS

by

Kun Su

A Dissertation

Submitted to

the Faculty of The Graduate School at

The University of North Carolina at Greensboro

in Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

Greensboro

2022

Approved by

Dr. Robert Henson
Committee Chair

DEDICATION

To my unborn daughter, who motivated me to graduate;

To my husband, who held my hands throughout the journey;

To my parents, who supported my academic pursuits constantly.

APPROVAL PAGE

This dissertation written by Kun Su has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair

Dr. Robert Henson

Committee Members

Dr. Kyung Yong Kim

Dr. Richard Luecht

Dr. John Willse

March 21, 2022

Date of Acceptance by Committee

October 21, 2021

Date of Final Oral Examination

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor Dr. Robert Henson. To this day, I still remember our first skype meeting before I even entering this program. Since then, you have been a great mentor and provided me with constant support and encouragement, especially during the dissertation journey. I would not have been able to complete this dissertation without you.

I would also like to thank the rest of my dissertation committee of Dr. Kyung Yong Kim, Dr. Richard Luecht, and Dr. John Willse for their support and feedback during the dissertation process and their guidance throughout my years at UNCG.

Lastly, I would like to acknowledge and thank Liu Siying, Yang Jing for serving as Subject Matter Experts roles in my dissertation, and Miao Lixian, Qian lei, Jiang Minli for helping me collect real response data from students. Your support made this real-data application possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Diagnostic classification models (DCMs) aim to provide master status to examinees for a set of discretely defined skills, or attributes, as well as detailed information regarding specific strengths and weaknesses. Researchers have sought to implement DCMs into school education—for example, in math and science—to increase learning efficiency by targeting student weaknesses. Gierl et al. (2010) developed a diagnostic mathematics assessment in Grades 3 and 6 to provide teachers with diagnostic information about students' cognitive mathematical knowledge. This type of cognitive diagnostic assessment can help to identify examinees' strengths and weaknesses, and the intervention can reinforce strengths while overcoming weaknesses (Gierl et al., 2010). Zhang (2018) applied a diagnostic classification model with a learning model to a spatial rotation test with learning intervention based on what examinees had answered wrong on the test. This model of learning could help track students' learning progress, assess the effectiveness of learning interventions, and detect causes that may affect learning outcomes (Zhang, 2018).

Early DCMs development can be classified into three origins: latent class analyses, mastery models, and knowledge state and rule space approaches (von Davier, M., & Lee, Y. S., 2019). From early development to its current state, the field of DCMs have inspired substantial theoretical work. Models have evolved from simple structures with few parameters—for example, the Deterministic Input Noisy "and" Gate (DINA) model (Junker & Sijtsma, 2001)—to models composed of complex structures and parameters—for example, a family of learning models that integrates a cognitive diagnostic model and a higher-order, hidden Markov model in one framework ((Wang et al., 2018). Estimation methods range from an EM algorithm to a

Markov Chain Monte Carlo algorithm, and fit statistics have been developed for models, items, and people.

Despite the development in DCMs theoretical work, it is rare to see a diagnostic test designed and administered to verify DCMs theory in mathematics and science. Most of the studies with real data applications have adapted retrofitted approaches to demonstrate the findings. For example, Lee et al. (2011) applied the DINA model to the Trends in International Mathematics and Science Study (TIMSS) 2007 data set for extra fine-grained information that can be translated and used in classroom instruction. The data set was extracted from Booklets 4 and 5 of the TIMSS 2007 fourth-grade mathematics assessment and consisted of 25 items with 15 multiple-choice items and 10 constructed response items. Results from this study showed that the DINA model presented a better model fit by the lower AIC and BIC statistics compared to item response theory (IRT) models, which are used by the TIMSS to calibrate and scale student performance. Chen et al. (2010) investigated mathematics performances on the TIMSS 1999 using an existing attribute and compared the cognitive attributes between Taiwanese students of different achievement levels, as well as between male and female students, and 2874 of students' responses were used in the analysis. The results indicated that the highest and lowest achieving students differed most on mastery probabilities for three of the attributes. Even though this study took a retrofitting approach, using the rule space method to retrofit under a unidimensional framework proved useful. Both the studies of Lee et al. (2011) and Chen et al. (2010) demonstrated that retrofitting DCMs to a unidimensional assessment could potentially provide extra information with better model fit. Studies done by Kabiri et al. (2016), Groß et al. (2015), Fay et al. (2018), and Arican & Sen (2015) also implied that richer diagnostic information could be obtained by using DCMs for real mathematics and science test data.

However, according to Liu et al. (2018), retrofitting could be problematic in two aspects—assessment design and statistical quality. In terms of assessment design, three major problems have emerged. First, there is no principal assessment design for diagnostic measurement to theoretically support the use of multidimensionality extraction on the unidimensional test. Second, the multidimensional structure will be hard to extract from a unidimensional test, which is designed to have a high reliability by decreasing the dimensionality. Third, even if multiple constructs could be identified after a unidimensional test was created, the number of times that each attribute is measured on the test is not typically sufficient to produce reliable results. As for statistical quality, three problems have been identified by Liu et al. (2018): (1) a high correlation of attributes that originate from one trait; (2) a low or even negative correlation between attribute and total score; and (3) poor model fit statistics. As stated in Liu et al. (2018), one would undoubtedly obtain better diagnostic information through a diagnostic assessment than through retrofitting a non-diagnostic test.

Compared to the studies that have used retrofitting, authentic DCMs applications developed in the diagnostic test manner are rare. A few non-retrofitting examples can be found in Gierl et al. (2010), in which they developed a diagnostic mathematics assessment in grades 3 and 6; in Zhang (2018), the author used a modified Spatial Rotation Test with build-in cognitive features; and in Tjoe & de la Torre (2014) the authors identified and validated the attributes for proportional reasoning domain from the ground up. However, these non-retrofitting examples do not usually mention the process of domain selection, and no DCMs real application study of the discipline of physics has yet been performed.

Many researchers have criticized applied DCMs studies for not having adequate validity evidence and rarely informing practical decision-making (Sessoms & Henson, 2018). In

Sinharay & Haberman (2009), a lack of validity evidence for the DCMs results was identified as a major problem with DCMs applications. This conclusion was also supported by Sessoms & Henson (2018), in which they summarized that only 22% of the studies within their sample of DCMs applications presented validity evidence. In Bradshaw et al. (2014), the authors stated that rarely have DCMs applications provided valid feedback to examinees. Sessoms & Henson (2018) also documented that only 3 out of 36 DCMs studies used DCMs results to inform students or teachers. Part of the current study is to investigate the application of a middle-school physics diagnostic test and provide adequate validity evidence as well as practical diagnosis information.

Even before the formulation of a strong validity argument there are other constraints that must be considered prior to determining whether a DCM approach is appropriate. For example, given the hypothesized construct to be measured, is it reasonable with the amount of time and expected test length. According to a report about test taking behavior of children by the Council of the Great City Schools, students in 66 districts were required to take an average of 112.3 tests between pre-K and grade 12. Adding another test to the curriculum would be an extra burden on students. Zhu (2016) mentioned the same problem for students in China. In Zhu (2016), the author stated that the academic burden on Chinese middle-school students was heavier despite existing policies to ease the academic burden. Additionally, DCMs require more items for sufficient information because of its multidimensional nature. However, it may be possible to justify an additional test if sufficient amount of information could be extracted from the test. This information may save time in otherwise by provide tailored lesson plans or more effective interventions for some students. Thus, another goal of the current study is to keep the diagnostic

test as short as possible while still maintaining enough information for any diagnostic information to be useful for the students and teachers.

While researchers must be conscientious of test length, there may also be circumstances that additional information can be collected to supplement the information from the test. This additional information can sometimes help provide guidance about how best to help students. In fact, this information can even be incorporated in a modeling approach. Specifically, researchers have been trying to use other covariates to help increase estimation accuracy of a student's ability. For example, Ferrando & Lorenzo-Seva (2007) proposed an IRT model incorporating the time taken to respond to the item, which resulted in a modest increase in the accuracy of the estimated score. Note that this was based on the idea that how long a student takes to answer a question may be related to a person's ability in that same construct. Although most of the research on the incorporation of additional covariates has been done when the latent construct is continuous, some researchers have addressed this in the context of DCMs, when the construct is a set of dichotomous attributes. For example, Zhan et al. (2018) proposed a DCM model incorporating response-time (RT) called the joint responses and times DINA (JRT-DINA) model. In the current study, the joint response-time DCM model is investigated for the potential benefit of increasing the accuracy of the individual profile estimation by also modeling the amount of time used to answer each item.

The overarching purpose of the current study is to provide a guideline for implementing an authentic DCM application that also includes response-time for students in middle-school physics classes. In this study, the complete process is described from the identification of an appropriate domain to test development and diagnosis reports. Validity arguments of the process are also presented along with the test and the diagnosis results. Finally, a DCM incorporating

response-time is investigated and its ability to improve estimation precision with respect to mastery of attributes is assessed.

CHAPTER II: OVERVIEW OF THREE PAPERS

In this dissertation, I illustrate a detailed example of the application of a DCM. Thus, the process of identifying a domain, development of a test, administration of that test, and then the administration of an intervention for an online middle-school physics test using a DCM framework with response-time is provided. The main objective of this dissertation is to provide a set of guidelines for researchers and educators interested in DCMs and physics classroom education, while also exploring the potential of adding additional information such as time spent on each item.

In the first paper, I illustrated the process of implementing a DCM into a middle-school physics test. First, content domain was carefully selected using a set of defined criteria, which are purposely defined to improve the success rate of DCM implementation. Note that one of the criteria in this process was based on the ability to define and assess a set of dichotomous attributes of interest. Given a selected domain and its attributes, multiple subject matter experts (SMEs) identified and agreed upon what possible combination of attributes could possibly be measured by a single item. Given the constraints of what a single item could possibly measure, a simulation-based approach was conducted to determine the best combination of items. Best in this context was defined as a test that would do equally well at measuring all attributes as indicated by high correct classification rates (CCRs). This combination of items determined the target Q-matrix for item writers. Finally, a physics test on the final Q-matrix was developed, administered, and analyzed by the author and the SMEs.

The second paper provides a narration of a step-by-step validation process for the usefulness of the information provided from the DCM-scored physics assessment with respect to learning and teaching. In the second paper, I developed a validity argument based on multiple

sources of evidence. Various qualitative methods were then used, including a literature review, an analysis of skill profiles, interviews with the SMEs and the students, and a think-aloud protocol on students to determine whether the evidence supports the defined validity arguments. In addition, a follow-up intervention based on the assessment of attribute mastery for each student was completed. These results are analyzed to determine the usefulness of the DCM approach in this setting.

The third paper investigates the potential benefit of incorporating response-time into the modeling approach. Specifically, the third paper examines whether it seems feasible that the classification accuracy could be improved by including response time in the DCM. If response time is helpful, then it might be possible to develop a shorter test that maintains the same amount of precision with respect to final student classifications. In the third paper, the JRT-DINA is estimated, and model fit is explored utilizing the item responses and response-times from the test. The change of individual classification and the posterior probability of mastery for each attribute were then analyzed. Finally, a simulation study is used to compare the estimated correct classification rate (CCR) of the DINA model with the estimated CCR of the JRT-DINA model.

CHAPTER III: SIGNIFICANCE

In physics education, an experienced teacher can quickly provide a student with diagnostic information by observing and potentially analyzing the student's errors on a worksheet or test. Given this information, an experienced teacher can usually identify the steps that a student most likely knows. In addition, patterns of errors might be identified that indicate a conceptual misunderstanding. However, not all schools have experienced physics teachers, and new teachers might not accurately diagnose a student's capabilities. With the help of an assessment that is developed based on an established diagnostic modeling framework, a clear picture of each student's profile can be estimated, which would not require a teacher to be experienced. Teachers can use this information to adjust the curriculum and help students improve. The tool could also eliminate the disadvantages of small teacher-to-student ratios and help establish a positive teacher-student understanding without spending excessive in-class time on individual assessments. Thus, overall efficacy would be improved with the proper use of DCMs.

Many secondary school students are not interested in physics, as they perceive it as irrelevant and challenging, according to Williams et al. (2003). Regardless of academic success in the course, students still find it difficult (Ekici, 2016). If learning efficacy is improved by using DCMs in the classroom and individually tailored guidance can be provided, students might find physics easier to learn. Such efficacy may result in an increased level of interest in physics for individuals with low initial interest.

DCMs have been developed substantially in recent research and it has great potential for diagnosing a student's strengths and weaknesses in a specific domain. Gierl et al. (2010), Zhang (2018) and Jang (2009) applied DCMs in math, science, and reading and obtained significant

results. Additionally, it has the potential to help physics teachers make the best use of class time and focus on students' cognitive blind spots. Despite DCMs' advantages, most teachers and scholars are not familiar with the approach, and its level of development and application in China still needs work. Applying DCMs to physics instruction and providing validity evidence of the modeling could play a significant role in educational reform and take middle-school physics education to the next level. Furthermore, DCMs could remove experience-level barriers and promote education fairness in teaching instruction.

CHAPTER IV: PAPER 1: APPLICATION OF A DIAGNOSTIC CLASSIFICATION MODEL TO A MIDDLE-SCHOOL PHYSICS TEST

Introduction

In a school system, it may be challenging for newer teachers with less experience to accurately determine a student's capabilities. This deficiency can become more acute in physics education because of the large gap between scientific concepts and students' personal beliefs about the physical world (Wells, Hestenes, & Swackhamer, 1995). Using assessment tools to evaluate student's ability can help guide teachers to address the gap. It will be more effective to close the gap if we know precisely what students know and don't know. Research conducted by Wells, Hestenes, & Swackhamer (1995) showed that traditional methods of teaching physics are ineffective in closing the gaps, and an innovative teaching strategy with better assessment tools is needed.

Educational Assessment describes a field that focuses on measuring the ability of a student, usually based on test performance. Thus, Educational Assessment can help identify a student's ability. Typical frameworks in Educational Assessment are Classical Test Theory (CTT) and Item Response Theory (IRT). Both the CTT and IRT frameworks are used to rank individuals (from high to low) on hypothesized underlying traits. However, some researchers have argued that a unidimensional score indicating a student's proficiency level on a hypothesized underlying construct might not be sufficient information for a teacher who seeks to improve that student's competency (Izsak, Remillard & Templin, 2016). A unidimensional score can only describe how much the students are struggling with the construct, but not why do students are not doing well. Educators need more fine-grained information about students' skill

profiles in order to adjust the curriculum and assignments to address those areas of weakness for a student. If teachers knew specifically where these gaps were in the construct of physics, they might be able to provide more tailed lesson plans that more effectively minimize these gaps.

Diagnostic Classification Models (DCMs) is a very different approach to scoring tests compared to the CTT and IRT because it does not focus only on a unidimensional construct. In contrast, DCMs aim to diagnose examinees' mastery status of a multidimensional set of discretely defined skills or attributes, thereby providing examinees with detailed information regarding specific strengths and weaknesses. DCMs provide refined information on a test taker's skills profile by outputting statistical information describing skill mastery or non-mastery (Sessoms & Henson, 2018). Therefore, DCMs can be applied to meet educators' demands for fine-grained formative information of students' skill profiles rather than simply a unidimensional score.

There are many examples that DCMs were applied to generate fine-grained formative information about students. Chen et al. (2010) investigated Taiwanese mathematics performance on TIMSS 1999 by using an existing attribute and compared the cognitive attributes between students of different achievement levels as well as between male and female students. Lee et al. (2011) applied DCM to TIMSS 2007 fourth grade mathematics assessment and generated extra fine-grained information that can be translated and used in classroom instruction. Kabiri et al. (2016) utilized DCM to generate rich diagnostic information using TIMSS 2011 Iranian eighth-graders' science assessment data.

Despite the advantages of DCMs, it may not always be effective. For example, if a diagnostic test requires forty items for a reliable output, but the testing time only allows twenty items to be tested, DCMs might not be the optimal psychometric models. There may also be

structural challenges that prevent the use of DCMs even when theoretically in might be advantageous. For example, DCMs cannot be applied to a domain where no distinct traits can be identified.

This paper describes and illustrates the application of a Diagnostic Classification Model to a middle-school physics test by first, and most importantly emphasizing the real-life related characteristics that make DCMs application feasible. Of particular importance in this paper the real-life characteristics are related to what domains appear to be most appropriate and what constraints are placed on how the test can be written. Ideally, the diagnostic results will provide more fine-grained information about students' knowledge and help to inform the teacher in tailoring educational plans.

Literature Review

Researchers have been implementing Diagnostic Classification Models (DCMs) into school education, and it has been used to support and measure learning in math topics, such as proportional reasoning (Tjoe and de la Torre, 2013), as well in undergraduate mathematics (Mejía-Ramos, Lew, de la Torre, & Weber, 2017) and in grade 3 and grade 6 math (Gierl et al., 2010).

Although DCMs are being discussed more and some examples of its effectiveness have been demonstrated, it has remained underutilized. Compared with CTT and traditional IRT models, DCM is relatively novel and, in some cases, more complex. CTT has been around since the beginning of the 20th century and IRT as a theory occurred during the middle of the 20th century. While there are some DCM like approaches from the 70's and 80's, DCMs as a theory didn't start until the 2000's when DINA model has been defined by Junker & Sijtsma in 2001. Furthermore, unlike CTT or IRT models, DCM models take on many different forms. The reason

for the many different forms is because DCM models are describing the relationship between the expected response of examinee and things they know and do not know. This relationship is described by both the Q-matrix and the model that selected. Q-matrix depicts which attributes each item specifically measures. The complexity of the model characterizes the nature of what happens whenever somebody has mastered certain attributes not others with respect to how they response to items. One of the simpler DCM models is the DINA. In DINA, if you've mastered all the required attributes for an item, you should probably get that item right, and this probability is described by a parameter in the model. If you lack one or more required attributes, you should probably miss that item, and this probability is also described by a parameter in the DINA model. Examples of other DCM models include the Deterministic Input; Noisy "Or" Gate model (DINO; Templin & Henson, 2006), the Reparameterized Unified Model (RUM; Hartz, 2002), The Log-Linear Cognitive Diagnosis Model (LCDM; Henson et al., 2009).

Because DCMs are newer, there are not many examples where tests have been explicitly written to measure a set of dichotomous attributes. In contrast, there are many examples of unidimensional traits being measured. For example, the Trends in International Mathematics and Science Study (TIMSS) provides data on the mathematics and science achievement of students in different countries, as well as The Partnership for Assessment of Readiness for College and Careers (PARCC) which assesses students between Grade 3 and Grade 11 in Mathematics and English. As a result, many attempts to demonstrate the value of DCMs use unidimensional tests, and this method is called retrofitting. Retrofitting is a method that applies DCM to existing assessments that have CTT- or IRT- framework. Many studies have adopted this method (Chen et al., 2013; Chen & de la Torre, 2013; Cui et al., 2012; de la Torre, 2009; de la Torre & Douglas, 2004; Henson et al., 2009; Hou et al., 2014; Jang, 2009; Kim, 2015; Lee & Sawaki,

2009; Li et al., 2016; Li & Suen, 2013; Ravand, 2016; Templin & Bradshaw, 2013; von Davier, 2005, 2007, 2014; Warner, 2013), which might be due to the high input cost of creating new items. The Indiana Department of Education (IDOE) compensates item writers at a rate of $125 per completed state assessment item. By retrofitting, researchers can implement DCM theory without creating a new test.

While retrofitting may be a cost-effective approach to implementing DCM, retrofitting can also be problematic. Gierl & Cui (2008) argued that three characteristics of DCM lead to unsuccessful retrofitting: (1) a retrofitting data set rarely satisfies the high dimensionality demands of DCM, thereby resulting in a poor model fit; (2) the confirmatory nature of DCM requires a diagnostic test to specify the structure of the data. If a test does not contain such a psychometric structure, the relationship between the observation and the latent trait will be difficult to recover; and (3) the diagnostic nature of the interpretations requires that the test designer takes supporting instruction into account, yet existing educational assessment tests rarely incorporate this objective (Gierl & Cui, 2008). Because of these limitations, retrofitting can only be viewed as the addition of new technology to an older, unsuitable system and often results in unsatisfactory diagnostic classification outcomes (Gierl & Cui, 2008).

To overcome the limitations of retrofitting, people must describe frameworks that could be used to measure various constructs. For example, assessment engineering (AE) (Luecht, 2013) could be used as the framework during the test development process. AE outlines the evidentiary argument-based assessment design and aims to offer detailed, meaningful, cognitive difficulty descriptions regarding the scale (Luecht, 2013). Evidence-centered assessment design (ECD) also provides such a design framework, which can ensure a meaningful test score or subscores and provide guidelines for the test design system while maintaining a certain level of generality

that supports a broad range of assessment types—from widely used, large-scale standardized tests to a low-stake classroom quiz (Mislevy et al., 2004). Furthermore, the cognitive design system approach proposed by Embretson (1998) can as well as provide a framework that guides designing items to measure target aspects of the cognitive process. All three approaches emphasize the importance of take the construct into consideration and how the construct related to the task, and then how to write a test that is specific to the task.

Based on these three frameworks, we can notice that retrofitting could be problematic because it tries to measure skills that are not explicitly designed for the test. We can also learn that construct needs to be defined cautiously, and the task should be closely related to the construct. Each of the three frameworks has a slightly different approach to defining the construct, outlining the task, and designing the test. Many aspects from these three frameworks can be capitalized on to design the test for a multidimensional construct with DCMs so that it will contain integrated diagnostic features.

However, there are only limited examples of how to develop a test under a DCM framework. Culpepper (2015) applied a DINA model to a spatial visualization test. This test was developed based on an exam taker's four mental rotation skills: (1) 90° x-axis, (2) 90° y-axis, (3) 180° x-axis, and (4) 180° y-axis. Culpepper then analyzed the test data using the DINA model with the defined mental rotation skills as the attributes using a Q-matrix that was based on how the items were written. Because the skills had been previously defined, it was known which skills were intended to be measured. Although this example does not fall into the retrofitting category—because the test used in the study was a revision of an existing rotation visualization test—it provides minimal insight on developing a new diagnostic test from the start. Gierl et al. (2010) developed a diagnostic test for third- and sixth-grade math students using the Attribute

Hierarchy Method (AHM) for the Q matrix construction. The test was administered using a computer-based testing system, with 338 students from third grade and 184 students from sixth grade. After the test, an attribute-based scoring process was implemented, and the students were given a diagnostic report. This study developed the test based on a pre-determined Q-matrix and analyzed the test data using the same Q-matrix under the DCM framework. It provides a valuable reference for the current study and researchers who want to develop a new diagnostic test.

This paper discusses process that includes selecting a domain, defining the attributes, constructing a Q-matrix, test development and administration, data collection, and analyses. Because the current study is a real-life application, some operational limitations were considered from the beginning, for instance, allowable testing time, test length, and the feasibility of extracting separatable attributes from a specific domain. A critical characteristic of the current study is how to select a domain within the restrictions that will lead to the best success rate of a DCM application. Note that while DCM can be useful, it is believed that it cannot be applied in all domains given specific limitations and constraints, which has been discussed. No literature has been found that provides guidelines for specifically emphasizing real-life related characteristics that make DCM application feasible. In the current study, I discuss and examine four factors in selecting a domain that make DCM application realistic.

The first factor is about whether this domain can naturally be separated into different attributes. In order to measure attributes using a test, we need to identify the attributes first. Many people have identified the attributes during the DCMs related research. Most of them identified the attributes after the test is created and that can be a challenge (Jang, 2009; Kim, 2015; Li et al., 2016; Ravand, 2016; von Davier, 2005). There are also examples that identified the attributes before creating the test. For example, Tjoe & de la Torre (2014) identified six core

attributes in proportional reasoning through a series of meetings with subject matter expects without looking at the existing test. Bradshaw et al. (2014) identified set of attributes for a test of teachers' multiplicative reasoning before constructing the test. Identifying distinct but related traits could be a primary challenge while at attribute identification (Bradshaw et al., 2014).

The second factor concerns the number of the attributes. According to the evaluated sample of DCM applications in Sessoms & Henson (2018), the number of attributes in the DCMs related researches could range from 4 to 23, and the average number of attributes estimated was 8. A related topic for number of the attributes is the grain size of the attributes. If an attribute is too broad, it may be unreasonable to treat it as a dichotomous attribute and yet, having it too specific will result in too many attributes that might not be feasible to measure in a test.

The third factor is whether the attributes can be measured separately. If some attributes must be together all the time to form an item, there will be confounding issues during estimation. Bradshaw et al. (2014) also ran into the same issue in which the authors could not reliably discriminate between two related attributes and were compelled to combine them. Note that even if experts think the attributes are distinct, they may not be distinct in the population, which will also cause issues when we try to measure the target population using a test.

Lastly, even if items could be formed from different attribute combinations, such items should be familiar to students without introducing construct-irrelevant variation into the problem-solving process. Construct-irrelevant variation may give an unfair advantage or disadvantage to one or more subgroups of test-takers (Standards, AERA, APA, & NCME, 2014). With all these factors considered, I selected a target domain for the diagnostic test.

After selecting the domain, I formulated a list of attributes, constructed a Q-matrix using a simulation approach, and developed and administered a diagnostic test. Bradshaw et al. (2014)

and Gierl et al. (2010) provided a great example of developing and administering a diagnostic test. Bradshaw et al. (2014) presented a multidimensional test that analyzes middle- grades teachers' understanding of fraction arithmetic. Gierl et al. (2010) designed an operational diagnostic mathematics exam for grades 3 and 6 and evaluated the student response. Many aspects from these two studies can be capitalized on and guide the test development and administration of the current study.

When developing a new test, parallel test forms are occasionally required to decrease cheating effects. Gierl et al. (2012) demonstrated a method to create parallel forms by using automatic item generation to create multiple-choice test items with a cognitive model structure and item models. To decrease the cheating effect, rearranged questions and responses could be used in the test as standard procedures (Bresnock et al., 1989). However, a study showed that different item sequences of difficulty could impair test reliability (Hodson, 1984). It is important to maintain the same level of test reliability while decrease the cheating effect as much as possible. The current study employed a method to decrease the cheating effect while maintaining test reliability.

Besides the existing methods, a new simulation approach to identifying a Q-matrix to measure these attributes based on inferred Correct Classification Rate (CCR) was presented in this study. This new approach was used to create the Q-matrix and guide the writing of items.

Purpose

While the current literature on DCMs application includes many examples in mathematics and science, these examples do not do a sufficient job at describing the specific steps that were used to initially identify an appropriate domain from several domains. In addition, when constructing a test for a defined set of attributes, the possibility of constraints in

the item construction have not been explicitly considered when constructing the test. These constraints can impact the nature of the test and how all attributes can be reliably measured. Furthermore, no examples have been demonstrated in the physics domain, which could potentially be another example where DCMs could be beneficial.

The purpose of the current study is to describe and illustrate an application of a Diagnostic Classification Model to a middle-school physics test. This process includes (1) inspecting several domains and using four different factors to select a domain in which the DCM would be most useful and likely to succeed; (2) formulating a list of attributes that account for the theoretical and operational understanding of the selected domain relevant to eighth-grade physics that can serve as an effective tool to evaluate students' understanding; (3) introducing a new simulation approach to identifying an "optimal" Q-matrix on to measure these attributes given constraints of what items can measure; (4) constructing a diagnostic test based on the selected domain and the final Q-matrix; (5) administering the test to the target population and collecting response data; and (6) collecting and analyzing students' response data.

Research Questions

1. What are factors that should be used to select a domain such that a DCM approach is most applicable?

2. What are the steps to define attributes and Q-matrix for the selected domain in middle-school physics?

3. What are the steps to develop, administer and analyze a physics diagnostic test?

Methods

As with the research questions, this study was partitioned into three primary steps. The first step is selecting the domain, the second step is developing the test, and the finally the data

collection and model calibration. Prior to selecting the domain, SMEs needed to be familiarized with DCM. As will typically be that case, the SMEs are familiar with basic concepts of the content (physics) but are not familiar with thinking of the concepts and the context of DCMs. Thus, brief materials about DCMs were provided to the SMEs prior to the interview; then, during the interview, any questions and critical content about the basic concepts of attributes, Q-matrix, and DCMs was reviewed. At this point, SMEs had the knowledge of basic terminology such that we were able to work through the steps as a group. At each given step, I would ask them questions, and they would provide additional feedback, which resulted in the domain options being narrowed. Working as a group, they addressed each step, and final conclusions were made when consensus was reached

Once the SMEs were familiar with DCMs, four general factors are considered to identify a potential domain for the application of DCM. Precisely, in selecting a domain, the extractability, the number of attributes, uniqueness of the attributes in addition to the ability to write items in a commonly accepted format for the identified attributes.

As mentioned before, SMEs were asked to inspect the four factors on domain selection in a group interview. The first factor is whether the SMEs can extract attributes from the target domain. The SMEs were informed in a conversation during the interview that attributes are skills that a student can master or not; hence the attributes are dichotomous. To meet this requirement, the SMEs went through each domain for eighth-grade physics, then chose a few domains that could satisfy the first factor in the interview.

After identifying a few specific domains, the second factor—number of attributes—was addressed by considering the expected length of the test and the total number of attributes that could be feasibly measured in that time. This diagnostic test was administered in the form of

after-school homework or a class quiz. In the interview, I asked SMEs about how much time is available to administer the test and how many items can be included in the test. Both SMEs discussed and agreed that a thirty minutes twenty items multiple-choice test would be feasible. Based on the literature review, the appropriate number of attributes should be ranged from four to eight considering the test length and sample size. SMEs were asked to extract attributes from the few domains that satisfied the first factor. Any domain with a number of attributes more than eight and less than four would be eliminated.

To evaluate the third factor—whether the attributes can be measured separately—the SMEs were asked to create an item that only measures one attribute. Those attributes that could not be measured singly would be dropped or merged with others. Lastly, those items were examined by the SMEs on whether students are familiar with questions written in that way. Considering the abovementioned factors, one best domain was selected as the target domain for the diagnostic test. After deciding on the topic, the SMEs explicitly defined all the desired attributes based on the curriculum requirements. The detailed definition of all the attributes is described in Appendix A.

After the best domain is selected, some thought also went into which model would most like be appropriate. An appropriate DCM should be able to describe the relationship of the attributes mastery and getting the item correct in the test. It should also be well-studied in the literature. The relationship of the attributes mastery and getting the item correct can be described by conjunctive and disjunctive model. In conjunctive models, a lack of mastery in one attribute cannot compensate for a lack of mastery in other attributes (Henson, Templin & Willse, 2009). However, in disjunctive models, the mastery of a subset of skills, or sometimes one skill, can lead to a high probability of a correct response (Henson, Templin & Willse, 2009). In this study,

the SMEs were asked whether students needed to master all the required attributes to obtain a correct item, and then decide which type of model is the most appropriate.

In the second phase, the SMEs constructed a list of possible combinations of attributes that could be measured by a single item. Note that any given combination of attributes measured by a single item could be thought of as Q-matrix vector. Thus, this list contains all the possible Q-matrix vectors. Any test constructed for this domain can only be constructed using items that measure the attribute combinations in this list, which is a constraint for test construction. From this list, a subset of potential attribute combinations (allowing for replacement and so two items could theoretically measure the attribute combination) was determined such that if items were created, the attributes would be measured as well as possible. This section was completed taking into account the test length restrictions. To a simulation study was used predict how well each attribute would be measured given a specific test (i.e., the Q-matrix). Once a Q-matrix was identified with high correct classification rates (CCR) for all attributes, this Q-matrix will be used as the blueprint for item construction.

Note that ideally, one could optimize this process. However, because item parameters are unknown, the goal is to determine a process that would result in a good combination of items to measure the defined attributes. Besides not knowing the item parameters, computationally, it is impossible to try all the combinations in the form of a real test, and thus a simulation approach was defined to determine an effective test of fixed length to assess those attributes when not all combinations were possible.

To illustrate the process of the simulation study, x represents the number of items for the test based on real-life testing time limitation, and y represents the number of all the possible

attribute combinations that could generate an item. The simulation study was conducted as follows:

1. A total of *x* rows were randomly selected from the list of possible Q-matrix vectors (*y* rows) with replacement.

2. Five hundred students' profile data (mastery or nonmastery of the attributes) were randomly simulated assuming an equal probability for all the combinations of mastery patterns. Equal probabilities are used as the worst-case scenario. In the event that some attributes are more likely, this information could increase CCRs if a Bayesian approach was used.

3. The response data set was simulated using the generated Q-matrix and the profile data produced in #1 and #2 using the Reduced RUM Model. In the simulated data, the parameters were fixed such that the π_i^* has the value of 0.85, and the r_{ia}^* has the value of 0.2. These values are consistent with parameter values reported in the literature (Chiu & Köhn, 2016; Feng et al., 2014).

4. The simulated response data was then calibrated using the Reduced RUM Model again. This calibration resulted in item parameters and more importantly examinee posterior probability of mastery for each attribute. These posterior probabilities of mastery can then be used to estimate examinee mastery, α_{ik}, for the i^{th} simulated examinee on the k^{th} attribute such that if the posterior probability is great than 1 then the examinee is estimated to be a master (i.e., $\alpha_{ik} = 1$) and otherwise the examinee is estimated to be a nonmaster (i.e., $\alpha_{ik} = 0$)

5. The correct classification rate (CCR) was estimated by comparing the estimated profiles to the true originally simulated profiles in Step 2. Note that in this case the CCR is computed as the proportion of times that estimated mastery match true mastery across all attributes and simulated examines.

6. The process Step 1 through Step 5 was then repeated for 4,000 times. The Q-matrix that resulted in the highest CCR is used as the final Q-matrix and the test design.

Based on the final Q-matrix, which specified the blueprint for what items should measure, a multiple-choice item test was developed by two SMEs. This test was then reviewed and examined by the third SME. To decrease the cheating effect while maintaining test reliability, items were rearranged in three-item "difficulty blocks" to create parallel test forms. The procedure was as follows: all the SMEs were required to rate the items on whether they were easy, medium, or hard based on their professional judgment. If there was disagreement on the rating, SMEs were asked to discuss among themselves until they reached agreement. Items were arranged from easy to hard in terms of difficulty on the test. Items within the same difficulty rating were rearranged to create four different test forms.

After having the test (and corresponding forms) created the final step was to collect data and calibrate a DCM. Students were directed to different forms based on their student ID, and the number of students who took the forms was similar. The test was administrated through Qualtrics (Qualtrics, Provo, UT). Students could take the test anytime during the two days the Qualtrics link was open. While there was no time limit specified into the Qualtrics administration of the test, students were asked to finish the test within a certain time limit and informed that they could not go back and change answers once the test began. Only fully responded cases were recorded. The data was collected and downloaded through Qualtrics.

After the students' response data was collected, a DINA model was fitted and item parameters and students' profiles were generated using the R package CDM (Robitzsch et al., 2020). DINA mode was selected for two reasons. First, the SMEs determined that students need to master all the required attributes to get the item correct, and this quality is captured in the

conjunctive model such as DINA model. Secondly, DINA model is widely applied in the literature. After calibrating the data using DINA model, item parameters estimate, attribute probability and profile probability were generated, reported, and analyzed based on the selected model. Item discrimination index (IDI; Lee, de la Torre & Park, 2012) and the root mean square error of approximation (RMSEA) item fit index for each item j were computed and analyzed for each item.

$$IDI_j = 1 - s_j - g_j$$

$$RMSEA_j = \sqrt{\sum_k \sum_c \pi(\boldsymbol{\theta}_c) \left(P_j(\boldsymbol{\theta}_c) - \frac{n_{jkc}}{N_{jc}} \right)^2}$$

The IDI is a measure of how well an item is able to distinguish between masters and non-masters and the RMSEA is an absolute fit index. Both indices could be an indicator of the item quality. The IDI for each item was computed and summarized. Specific outliers to where they're highly discriminate or reasonably low discrimination compared to the rest the group was examined. For the item RMSEA values, any value greater than 0.1 is classified as poor fit, values less than .1 and greater than 0.05 is moderate fit and values less than 0.05 indicates good fit (Kunina-Habenicht et al., 2009). The guessing and slipping parameters for each item were computed and summarized. Specific outliers which have a guessing or slipping parameter greater than 0.5 were inspected. The *p-value* for each item was computed and summarized. Tetrachoric correlations between the attributes were computed and analyzed for all the attributes. I identified large value in tetrachoric correlation where they describe the association between in the event that associations are estimated higher than 0.7 and evaluated the relationship between the attributes.

The probability of different profile was analyzed. The ten profiles with the highest probability were evaluated. The SMEs were interviewed on their expectation on how many students should master all the attributes and how many students master none. This information was compared to the actual percentage of all-mastery and non-mastery. If there is any inconsistency, potential reason which might cause the inconsistency would be analyzed.

Lastly, the probability of mastery for all the attributes was computed and summarized. The SMEs were asked in the interview about their expectation for the most and the least mastered attributes, and then compared to the actual results.

Results

Domain Selection and Attributes Define Phase

There are 13 chapters in the eighth-grade physics curriculum. In this class, domains are taught in chapters, so each chapter represents one domain. The SMEs reviewed each chapter to determine whether dichotomous attributes could be extracted from the 13 domains. If there is a disagreement, they would discuss and come to a consensus in the end. After reviewing all the chapters, the SMEs concluded that all of the domains have attributes that can be extracted.

After having determined that all domains (all 13 chapters) could be broken down into dichotomous attributes, the next factor was to determine whether the number of attributes (and grain size) was reasonable for the testing window that was permitted. Because the diagnostic test was administered as after-school homework, the SMEs decided the total testing time should be less than thirty minutes, and as a result they felt that the test should be composed of no more than twenty multiple-choice items. According to the literature, the average number of attributes measured by a diagnostic test is eight (Sessoms & Henson, 2018). Considering that the test

length is short—twenty items—I only considered domains with fewer than eight attributes. Only four of the 13 domains were determined to have fewer than eight attributes.

The next factor to evaluate whether a domain would be appropriate to measure using a DCM was whether each attribute could be measured separately. In the four domains with fewer than eight attributes, SMEs were asked if separate items could be written to measure each of the attributes in the four domains. They determined that two out of the four domains had attributes that are typically measured with other attributes together as one question rather than being measured singly as one item, which would create a possible confound for a DCM. Therefore, two domains, buoyancy force and mechanical efficiency, where considered.

Recall that the goal is to evaluate the domain that could most reasonably work with diagnostic models. As such, in thinking about how to measure attributes in the given domain, I did not want to depart from the typical assessment (e.g., item type) students are usually exposed to. As an initial check, instead of evaluating all the possible ways of writing items, the SMEs were asked if items could be written to measure each of the attributes in the previous step, with follow-up questions exploring whether the items measured only a single attribute would also be familiar to students. Thus, in this step, SMEs evaluate whether students are familiar with items that only measure one attribute in these two domains.

Typical items in the mechanical efficiency domain are mostly computational questions. Students need to complete multiple steps to reach the final correct answer. There are a total of seven attributes in this domain and each one in the mechanical efficiency domain is equally important and easy to define. Each attribute is defined as being able to identify when is appropriate to use the formula, memorize the correct formula and know how to use the formula to obtain a specific value. SMEs can easily create items that measure one or multiple

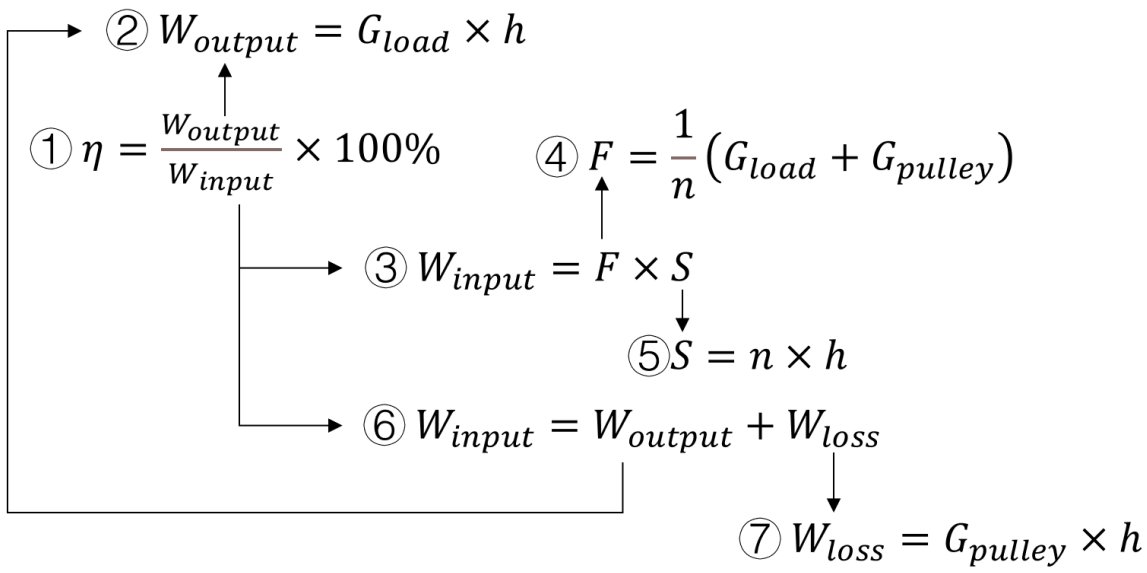
combinations of attributes by providing some information and asking students to compute the answer using the required attribute(s). Most importantly, these items will look very familiar to the items that students are usually given on the test. Although in buoyancy force, some attributes are computational skills, others concern analyzing, experimenting, and recognizing skills. Different types of skills make the test development process more difficult, and the items may need to differ from the typical questions students are usually exposed to. For example, a typical question in buoyancy force is usually a combination of computational, analytical, and recognition skills. Although SMEs could create an item that measures only one skill, it will look unfamiliar to students and have the risk of introducing irrelevant variance into the test design.

Based on these criteria and in addition to the number of attribute measures, mechanical efficiency was selected as the most suitable domain for constructing the DCM-scored test. Provided that this domain was most appropriate, some thought also went into which model would most like be appropriate. An appropriate DCM should be able to capture the relationship of the mastery of the attributes and getting the correct answer for the test and is well-studied in the literature. The most commonly used diagnostic models can be categorized into conjunctive and disjunctive models based on the attribute requirements for an examinee to obtain a correct item. In conjunctive models, a lack of mastery in one attribute cannot compensate for a lack of mastery in other attributes (Henson, Templin & Willse, 2009). However, in disjunctive models, the mastery of a subset of skills, or sometimes one skill, can lead to a high probability of a correct response (Henson, Templin & Willse, 2009). In this study, the SMEs were asked in the interview whether students needed to master all the required attributes to obtain a correct answer for the items. The SMEs discussed and agreed that students need all the required attributes to reach the correct answer, which means the test items follow a conjunctive model. Common

conjunctive models include the Deterministic Input; Noisy "And" gate (DINA; Junker & Sijtsma, 2001), Noisy Input; Deterministic "And" gate model (NIDA, Junker & Sijtsma, 2001), and the Reparameterized Unified Model (RUM, Hartz, 2002). DINA is commonly used in the literature, NIDA is rarely seen in the literature, and the RUM seeming over complex given the possibility of a small sample. Therefore, DINA was selected as the diagnostic classification model for this physics test.

Based on the SMEs' evaluation it is determined that mechanical efficiency would be a good domain for DCM and that a conjunctive model (i.e., the DINA model) would most likely describe the relationship between the attributes the probability of a correct response. Furthermore, SMEs were asked to define the attributes. In the mechanical efficiency domain, there are seven required formulas to solving questions in this domain. Each formula's mastery, including correctly memorizing and applying the formula, is defined as an attribute. The SMEs also helped describe the relationship of the seven formulas, which is shown in Figure 1 (a detailed description is presented in Appendix A). Although there are seven formulas, certain components in the formula could be computed from others. For example, W_{output} in formula 1 could be computed from formula 2. Thus, while there is a natural dependency among the seven formulas, such dependency could be avoided by providing particular values of the formula. For example, even though formula 1 relies on W_{output} and W_{input}, these values could be directly given in an item that asks students to calculate η, as opposed to asking for values of G_{load}, h, F , and S. As a result, while these formulas could build on each other in the way they are taught or assessed, they could be reasonably independent while measured in the diagnostic model.

Figure 4.1. The Relationship of the Seven Formulas



Q-Matrix Development Phase

After defining all of the attributes, the SMEs were asked to list all of the possible attribute combinations that could be realistically measured by a single item. Assuming no restrictions here are a total of 127 different combinations that could theoretically be measured when there are a total for seven attributes in a domain:

Numbers of combination=${}_7C_1 + {}_7C_2 + {}_7C_3 + {}_7C_4 + {}_7C_5 + {}_7C_6 + {}_7C_7$

$= 7 + 21 + 35 + 35 + 21 + 7 + 1 = 127.$

The SMEs considered all possible combinations and what a possible item would look like and eliminated those combinations that could not be measured by a single item. For example, attributes 3 and 6 could not be measured at the same time by a single item because both calculate W_{input} and could not be distinctively separated. The remaining combinations comprise the all-possible hypothetical items that could be written to create a test. Note that these combinations also represent all possible 0/1 rows of a Q-matrix for the constructed test. In this study, a total of

37 combinations of the original possible combinations were retained as possible items that could be included in the test. All possible combinations of attributes that could be measured by a single item are included in Appendix B.

Because the test length was set to 20 by the SMEs based on the allowable testing time, a subset of all possible items must be used to construct a test and its corresponding Q-matrix. Note that multiple items measuring the same attributes could be included in a test. However, when trying to measure seven attributes using only 20 items, not all combinations of items were considered the same. For example, if 20 items that only measured attribute 1 were included, then attribute 1 could be measured well, but no information about attributes 2–7 could be obtained. As a result, a "good" Q-matrix with only 20 items needed to be determined with the goal of measuring all attributes equally well while also addressing the fact that items cannot measure all possible attribute combinations.

A simulation approach was conducted to determine the Q-matrix structure that would do well at measuring all attributes given the constraints. Table 4.1 provides a summary of the estimated correct classification rates (CCR) for attributes on the 4,000 simulated 20 item tests including the average, minimum, maximum, and standard deviation of CCR. Recall that the Q-matrix of a simulated test was generated by randomly sampling, with replacement, from the list of 37 possible set of attributes measured by a single item.

Table 4.1 Summary of CCR for Attributes

Average	Min	Max	SD
0.746628	0.653143	0.832571	0.027048

The Q-matrix with the highest CCR for attributes was selected for the diagnostic test, and it is listed in Table 4.2. The final Q-matrix contains 20 items. Items 1–10 are simple structure

such that each item only measures one attribute. Items 11–14 measure two attributes, and items 15–20 measure three attributes. The Q-matrix measures attributes 1 and 6 the most (seven times) and attribute 4 the least (three times).

Table 4.2. Final Q-matrix with the Highest CCR

	attribute 1	attribute 2	attribute 3	attribute 4	attribute 5	attribute 6	attribute 7
item 1	1	0	0	0	0	0	0
item 2	1	0	0	0	0	0	0
item 3	0	1	0	0	0	0	0
item 4	0	1	0	0	0	0	0
item 5	0	0	0	1	0	0	0
item 6	0	0	0	1	0	0	0
item 7	0	0	0	0	1	0	0
item 8	0	0	0	0	0	1	0
item 9	0	0	0	0	0	1	0
item 10	0	0	0	0	0	0	1
item 11	1	1	0	0	0	0	0
item 12	0	0	1	0	0	1	0
item 13	0	0	0	1	0	0	1
item 14	0	0	0	0	0	1	1
item 15	1	1	1	0	0	0	0
item 16	1	0	1	0	1	0	0
item 17	1	0	1	0	1	0	0
item 18	1	0	0	0	0	1	1
item 19	0	0	1	0	1	1	0
item 20	0	0	1	0	0	1	1

Test Development and Administration Phase

After obtaining the target Q-matrix for the test, a multiple-choice item test was developed by two of the SMEs, and the test was reviewed and examined by the third SME. Specifically, for each row of the Q-matrix, the SMEs wrote an item that measures those specific attributes. Because of the previous review, it was known that items could be written to measure all combinations of attributes in this Q-matrix. Four test forms were created by rearranging the order of the questions. Students were directed to different test forms based on the first letter of their last name. The diagnostic test was administered to ninth-grade students in five middle schools in

Guangdong province through Qualtrics. Five-hundred and sixty responses were recorded in Qualtrics. After eliminating duplicate responses, as well as responses under 5 minutes and over 30 minutes, 397 were kept in the analysis.

Test Analysis Phase

After the data were exported from Qualtrics, the response data were analyzed in R using the "CDM" package. The item parameters and item statistics for the DCM-scored test are shown in Table 4.3. The item quality for items 1, 2, 4, and 8 were low with an item discrimination index (IDI) below 0.4. Items 1, 2, 4, 8, 9, and 10 have a guessing parameter estimate higher than 0.5. These items only measure one attribute. Items that measure two or three attributes have lower guessing parameters. Most of the items have a relatively small slipping parameter. The highest slip parameter is item 12 (0.238). The mean of the RMSEA item fit is 0.08. The RMSEA item-fit indices for the DINA model showed 3 items with good fit (RMSEA < .05), 13 items with moderate fit (RMSEA < .10), and 4 items with poor fit (RMSEA > .10) (Kunina-Habenicht et al., 2009). *p*-value is the proportion of times that item was answered correctly, and it is the indicator of item difficulty. The average *p*-value is 0.742, and 12 out of 20 items have a *p*-value greater than 0.7. Items that measure one attribute tend to have higher *p*-value than items that measure more than one attribute.

Table 4.3 Item Parameter and Item Statistics for the 20 Items

	Guess	SE (guess)	Slip	SE (Slip)	RMSEA	IDI	*p*-value
Item 1	0.803	0.052	0.013	0.005	0.054	0.184	0.929
Item 2	0.829	0.049	0.014	0.005	0.046	0.156	0.937
Item 3	0.404	0.045	0.058	0.014	0.046	0.538	0.791
Item 4	0.714	0.046	0.011	0.005	0.062	0.275	0.912
Item 5	0.395	0.048	0.026	0.006	0.075	0.579	0.761
Item 6	0.322	0.045	0.007	0.002	0.084	0.671	0.746
Item 7	0.022	0.004	0.040	0.009	0.016	0.939	0.660
Item 8	0.883	0.031	0.000	0.000	0.063	0.117	0.965

Item 9	0.565	0.046	0.017	0.007	0.079	0.418	0.856
Item 10	0.539	0.046	0.020	0.006	0.078	0.440	0.801
Item 11	0.415	0.039	0.010	0.004	0.076	0.575	0.763
Item 12	0.323	0.036	0.238	0.028	0.120	0.439	0.554
Item 13	0.265	0.035	0.046	0.012	0.099	0.689	0.615
Item 14	0.470	0.037	0.031	0.011	0.132	0.498	0.738
Item 15	0.495	0.040	0.011	0.003	0.082	0.494	0.766
Item 16	0.379	0.035	0.079	0.018	0.096	0.542	0.665
Item 17	0.249	0.030	0.167	0.026	0.076	0.584	0.557
Item 18	0.342	0.034	0.059	0.017	0.111	0.600	0.657
Item 19	0.198	0.029	0.156	0.025	0.094	0.647	0.524
Item 20	0.359	0.035	0.088	0.021	0.103	0.554	0.645

Table 4.4 provides the tetrachoric correlations between the seven attributes. The correlation between attributes 3 and 6 is the lowest with a value of 0.35, which could be due to the fact that both attributes are related to calculating W_{input}. Some students may tend to remember one of the formulas rather than both. The correlation between attribute 3 and the rest of the attributes are below 0.70, with the exception of attributes 2 and 7, which means attribute 3 has a low-to-moderate correlation with attributes 1, 4, 5, and 6. The remaining correlations are all higher than 0.7, which could be because the formula in attribute 3 is also taught and used in other four domains, whereas the rest of the formulas are specific to mechanical efficiency. Thus, attribute 3 has a relatively lower correlation with other attributes while correlations among the rest of the attributes are high. The correlation between attributes 4 and 5 has the highest value of 0.91, which might because both attributes are an element of attribute 3. This connection between attributes 4 and 5 makes it easier for students to remember both. The high correlation might be due to the fact that the two attributes are not distinct in nature. Mastery of one attribute would lead to the mastery of another in the population. It is necessary to have more discussions with SMEs on whether those attributes showed high correlation to each other could merge into one.

Table 4.4 Tetrachoric Correlations between the 7 Attributes

	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6	Attribute 7
Attribute 1	1.00	-	-	-	-	-	-
Attribute 2	0.76	1.00	-	-	-	-	-
Attribute 3	0.61	0.71	1.00	-	-	-	-
Attribute 4	0.89	0.84	0.70	1.00	-	-	-
Attribute 5	0.84	0.78	0.60	0.91	1.00	-	-
Attribute 6	0.89	0.72	0.35	0.80	0.77	1.00	-
Attribute 7	0.79	0.91	0.85	0.78	0.72	0.78	1.00

Table 4.5 provides the top 10 profiles with the highest probability, and they comprise almost 80% of the population. It also shows that 49.2% of the students have mastered all of the attributes, and 5.2% of the students have mastered none. I interviewed one SME on her expectation on how many students mastered all the attributes and how many students mastered none. She anticipated that approximately 40% percent of the students would master all the attributes, and approximately 5% would master none. This result is almost consistent with the SMEs' expectations. No particular hierarchical structure or learning pattern appears while examining the profiles. Table 4.6 provides the probability of mastery for all the attributes. The probability of mastery for each attribute is higher than 0.6, and the average probability of mastery is 0.7. Attributes 4 and 7 have the lowest probability of mastery, which is consistent with the SMEs' expectations. The SMEs expected attribute 4 to be the hardest formula to apply and attribute 7 as the most forgettable by the students. The rest of the attributes were expected to be equally easy.

Table 4.5 Partial Profile Probability

	Class probability	Cumulative probability
1111111	49.2%	49.2%
0000000	5.2%	54.4%
0010000	5.2%	59.6%
0000010	3.7%	63.3%
1000010	3.6%	66.9%
1101110	3.2%	70.1%

0101110	2.2%	72.3%
0100001	2.1%	74.4%
0110001	2.1%	76.5%
0000100	1.8%	78.3%

Table 4.6 Probability of Mastery for Each Attribute

	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6	Attribute 7
Probability of mastery	0.74	0.73	0.68	0.62	0.72	0.78	0.66

Discussion

Conclusion

The purpose of this study was to describe and illustrate a DCM application to a middle-school physics test while also considering how to select the most reasonable domain, and how to construct a Q-matrix with limitations on test length and the types of items that could be written. One important message of this study is that DCM cannot be applied to all domains without first considering the typical features of that domain. There are some characteristics of domains that make DCM a more favorable psychometric-scoring model than others. There are instances in which particular features prevented us from testing a domain with a DCM-scored test, and thus an examination needs to be undertaken prior to application.

The results demonstrate that it is possible to apply DCM to a middle-school physics test with careful selection of the content domain and a simulation approach for construction of a target Q-matrix. Only four out of 20 items showed poor fit and poor quality. The average correlation between the attributes shows moderate correlation, and some correlations are explainable based on the SMEs' professional judgement. Some of the attributes showed high correlation, and this might be because the two attributes are not distinct in nature. In the general population, mastery of one attribute would lead to the mastery of another. Further discussion with the SMEs would be necessary to determine whether the attributes that showed high

correlation to each other could be merged into a single attribute. The majority of the student profiles and the mastery probability of the attributes are both consistent with the SMEs' expectations. While examining the profiles, no learning or hierarchical structure pattern emerged, which is also consistent with the assumed independent relationship among the attributes.

While the results do have some issues, in a typical application, new items would be administered as experimental items first. After analyzing the item parameters based on responses, some would be dropped or refined in the next round if the item quality is insufficient. The current study only contains a single iteration. Even with some poor-quality items and high correlations among the attributes, the test is still considered an acceptable test with usable diagnostic information.

Although current literature on DCM applications includes many real data examples, most applications have used a retrofitting approach and did not design exams specifically as diagnostic models; however, the studies subsequently attempted to fit a diagnostic model. Even among few authentic DCM application examples, limited information exists on how to select a domain and construct a Q-matrix. Most of the studies have relied on SMEs for domain selection and Q-matrix construction. This study provides a new perspective in which the psychometrician can guide the domain selection process and the Q-matrix construction using practical methods.

Limitations and Future Directions

While this study showcases a domain selection and the Q-matrix construction procedure with a full-application process, it is not without limitations. Even though this was a low-stakes and easy assignment, many students were motivated, which explains the high proportion of all-mastered students (49.2%). However, no hierarchical structure and particular mastery profile were identified in cases of profile with very few mastered attributes, which could be due to some

students' low motivation, and their mastery of attributes might have been expected to be better. The mechanical efficiency domain is not the most challenging for ninth-grade students, as this domain is usually a small part of a test rather than the basis of a whole test. It is not uncommon if some of the attributes present high correlation because the coverage of this domain is rather small compared to the full test. Some of the attributes indeed showed high correlation, and this might be because the two attributes are not distinct in nature. Further investigation could be done on whether the attributes that showed high correlation are actually distinct, and whether they could be merged into one attribute. Typically, new items should be tested before actual administration. In the future, follow-up item analyses could be done to inspect the items, poor quality items could be dropped or refined, and the test could be improved. The current study provides some validity evidence; however, more evidence could be collected to show the validity of the test. In addition to the lack of validity evidence, no investigation was conducted on the usefulness of the diagnostic feedback.

In the future, the results of the current study could be used to improve items before administration, which might result in more reliable test outputs. Items with poor quality or fit could be appropriately examined for potential improvement or elimination before an actual test. Secondly, more work could be done to provide internal and external validity evidence to support student classifications. For example, to provide internal validity evidence, a think-aloud protocol with students could determine whether the SMEs' defined attributes are the skills students use during the problem-solving process. A correlation study could be done to compare this DCM-scored test with other physics tests to provide external evidence. Furthermore, future analysis should investigate how to present the feedback to students and teachers after verifying the

usefulness of the diagnostic feedback. Methods like surveys, interviews, or interventions could be used to evaluate the feedback.

CHAPTER V: PAPER 2: VALIDITY ARGUMENTS FOR APPLICATION OF THE DIAGNOSTIC CLASSIFICATION MODEL TO A MIDDLE-SCHOOL PHYSICS TEST

Introduction

Diagnostic classification models (DCMs) can provide student skill profiles that give statistical information of mastery or non-mastery of each skill. DCM is different from the Classical Test Theory (CTT) and Unidimensional Item Response Theory (UIRT) frameworks because it aims to provide examinees with detailed information regarding their specific strengths and weaknesses as opposed to only ranking individuals on hypothesized underlying traits. The hope is that this diagnostic feedback can help students target their weaknesses and increase study efficiency by having tailored learning material, while also providing teachers the necessary information to allow for classroom adjustments.

Although diagnostic feedback is the crucial feature of DCMs, its applications rarely provide adequate validity evidence and rarely inform practical decision-making (Sessoms & Henson, 2018). Within the sample of DCMs applications that Sessoms & Henson (2018) inspected, only 22% evaluated validity support for diagnostic results. Sinharay & Haberman (2009) identified a lack of validity evidence for the results provided by DCMs as one of the major issues with its application. Bradshaw et al. (2014) stated that rarely have DCMs applications given valid feedback to examinees.

The current study intends to develop and showcase the application of a validation process for the effects of the existing DCM-scored-physics assessment on learning and teaching. Validity arguments were established and analyzed for different sources of evidence. An intervention section was conducted to justify the utility of the diagnostic results.

Literature Review

Few researchers have systematically discussed validity in the context for diagnostic assessment. Sessoms & Henson (2018) provided validity arguments that included evidence for DCMs results in terms of construct representativeness, internal validity, external validity, and use of results. Yang & Embretson (2007) discussed the validity in terms of construct representativeness for cognitive diagnostic assessment from the perspectives of appropriateness, completeness, and granularity. The validity is most relevant to DCMs while in the context of classroom setting where DCMs are actually being used, because we would like to know whether DCMs are really useful to the students and teachers. However, only one example was found that examines diagnostic feedback in the classroom setting. Jang (2009) applied DCM to test data and provided diagnostic feedback to the students. The validity of the diagnostic information was evaluated using interviews, classroom observations, and surveys. Jang (2009) provides theoretical and practical guidance on how to collect validity evidence for a diagnostic test and could be capitalized for the current study.

The current study offers a step-by-step validation process for the effects of the existing DCM-scored-physics assessment on remediation based on current literature. The DCM-scored-physics assessment includes phases from test development, test administration, and test analysis that are similar to other existing studies (Bradshaw et al., 2014; Gierl et al., 2010). To validate the use of the diagnostic results, a validity argument was established and analyzed using corresponding evidence for each step. In addition, an intervention study was conducted to evaluate the usefulness of the diagnostic results. Five sources of validity evidence (evidence based on test content, response processes, internal structure, relations to other variables, and evidence for validity and consequences of testing) from Standards (AERA, APA, & NCME,

2014) were used as the guiding structure to formulate validity arguments on different aspects of the diagnostic test.

The first source of validity evidence is determined based on test content. Both construct representation and the construct-irrelevant variance should be evaluated for this source of evidence. Construct representation refers to the extent to which a test can capture important aspects of the target construct domain and it can be analyzed for appropriateness, completeness and granularity. Construct-irrelevant variance refers to the fact that particular features, such as demographics, social economic status, are impacting the meaning of the score in the way that is not related to the construct. Furthermore, in evaluating the general content (for construct relevant and construct irrelevant features), both the test developer's and the test taker's perspectives must be inspected because these viewpoints are not always parallel to each other. Construct is defined as the concept or characteristic that a test is designed to measure (Standards, AERA, APA, & NCME, 2014).

Particularly in DCMs, construct is represented by attributes which are multidimensional and fine-grained. Because of the additional complexity, to have valid construct representation across multiple attributes, Yang & Embretson (2007) introduced appropriateness and completeness to illustrate what is construct representativeness in DCMs. For DCM, appropriateness means whether the identified attributes are suitable to describe the cognitive process of interest (Yang & Embretson, 2007). For example, in basic math, addition, subtraction, multiplication and division are suitable to describe all the required skills. Whereas a reading skill should not be used to describe skills in basic math. Completeness means whether those attributes are adequate to represent the construct. (Yang & Embretson, 2007). For example, missing any of the skills in basic math will result in incomplete representation of the construct. Adequacy in the context of

completeness is referring to two different aspects. One aspect is adequate numbers of attributes, and the other aspect is the item/time adequacy used to measure each attribute. The former is relatively easier to decide through the literature and the experts while the second is related to the sampling items for CDA construct (Yang & Embretson, 2007). Unsolved issues remain in estimating the statistical distribution of the different attributes (Yang & Embretson, 2007).

Another aspect of construct representation is granularity. Using DCM, this is defined as the grain size for each skill. Grainsize of an attribute can vary because the attribute definition can be at different levels of complexity. For example, one can define addition, subtraction, multiplication, and division as four different math attributes. However, in a more advanced test these "four" attributes may instead be defined as one attribute representing basic math skills. An attribute with too-large of a grain size (i.e., being too general or containing too many combined "skills") can be viewed as a continuous ability. In cases where the grain size is too large, it can be difficult to determine the definition of "mastery" or trying to make a dichotomous decision. An attribute with too small of a grain size, however, can result in a dichotomous decision, which might limit the scope of the content coverage (de la Torre & Minchen, 2014) and jeopardize the generalizability of the diagnosis inference (Yang & Embretson, 2007). Because grain size is directly related to the number of attributes measured for a given test, the choice of grainsize is also limited by characteristics such as the capacity of measurement models, the sample size, and the computational power (Yang & Embretson, 2007).

Besides construct representation, construct-irrelevant variance should also be considered while collecting validity evidence based on test content. The existence of construct-irrelevant variance that is not part of the construct might systematically influence test scores (Standards, AERA, APA, & NCME, 2014). Spurgeon (2017) identified testing factors as one source of the

construct-irrelevant variances. Testing factors such as item creation, item wording, setting, and context could contribute to construct-irrelevant variation (Spurgeon, 2017).

The second source of evidence is based on examinees' response processes. Standards (AERA, APA, & NCME, 2014) defined this as the evidence concerning the fit between the construct and the detailed nature of the performance or response actually undertaken by test takers. Methods used to obtain validity evidence related to response processes include using think-aloud procedures and verbal probes; interviewers ask students to describe their thinking either simultaneously as they answer each question or retrospectively after they complete the test (Peterson et al., 2017). Note that these methods are similar and have been used in the DCMs studies. For example, Wang & Gierl (2011) validated the cognitive models by having a sample of students think aloud as they solved each item.

The third source of evidence is based on internal structure. According to Standards (AERA, APA, & NCME, 2014), "analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (p. 16). To evaluate internal structure, confirmatory factor analysis could be used to inspect if the numbers of factors we expected could explain the outcome, and if the correlations between those attributes is low enough to justify the factors are distinct. In DCMs, we can use similar process to inspect the internal structure. Sessoms & Henson (2018) suggested that evaluating the alignment for the diagnostic results and theory-based expectations could serve as internal structure validity. For example, the diagnostic results should show that the correlation between different attributes is low, and fewer students master a difficult attribute.

According to the Standards (2014), the fourth source of evidence is “evidence based on relations to other variables”. In a unidimensional test, the test score is expected to correlate with some criteria or other tests hypothesized to measure the same constructs (Standards, AERA, APA, & NCME, 2014). In a context where the test is multidimensional, we would expect each dimension or a combination of dimensions to be correlated to a construct that is related to. In the current study, the diagnostic test is multidimensional and the estimated attribute mastery from a test could be considered related to other physics skills or physics tests. Sessoms & Henson (2018) also mentioned that DCM mastery classifications of a math test should be moderately correlated with other math skills.

The last source of evidence for a validity argument refers to the consequences of testing. In DCMs, we expect the diagnostic results to be useful to the students and teacher. Positive post-intervention consequences based on diagnostic feedback could serve as strong support of the usefulness of the DCM results (Sessoms & Henson, 2018). Evaluating the usefulness of the diagnostic feedback using interviews, classroom observations, and surveys could provide validity evidence for the uses of test score (Jang, 2009). Evaluating the extent to which various groups of users interpret assessment results appropriately is also suitable for a validation study (Lane, 1999).

Purpose

Although current literature on DCM application has included real data examples, most of these focus on model fit and estimation of the attribute mastery profile. However, these studies have not provided sufficient validity evidence of the estimate of the attribute profile and test use. In fact, few theoretical guidelines have been found in the literature, and not many researchers discuss validity in DCMs (Sessoms & Henson, 2018). The purpose of this study is to examine

the validity of the DCM application to middle-school physics tests critically using multiple pieces of evidence. This paper narrates a step-by-step validation process for the DCM effects of the scored physics-assessment classification results based on different sources of validity evidence. This process includes: (1) constructing several validity arguments on different sources of such evidence; (2) using various qualitative methods including a literature review, information gathering, interviews with the SMEs and the students, and think-aloud protocol on students to analyze the validity arguments for a specific physics test; and (3) verifying the usefulness of diagnostic feedback with a follow-up intervention.

Research Questions

1. How can a validity argument be constructed for DCMs based on the Standards?
2. What evidence could be used to support the validity argument for a diagnostic test?
3. How does one verify the usefulness of feedback when using a diagnostic test?

Methods

Participants

DCM-scored physics test takers: a total of 524 test-takers took the online physics test on the mechanical efficiency. They are ninth-grade students from four different school in Guangdong province, China.

Subject matter experts: three physics teachers from three different middle school in Guangdong province, China.

Instruments

The pre-test is an existing middle-school physics DCM-scored test including twenty items measuring the domain of mechanical efficiency. The pre-test was constructed based on seven attributes with four parallel forms. Pre-test items were administrated through Qualtrics

(Qualtrics, Provo, UT). Students can take the test anytime during the two days the Qualtrics link is open. While no time limit is built into the Qualtrics program, students were asked to complete the test in 30 minutes. Students were informed that they cannot go back and change answers once they have moved to the next item. Only fully completed tests were recorded. The data was collected and downloaded through Qualtrics.

The intervention material was created by one of the SMEs in an electronic version and verified by another SME. The material is an instructional handout for each attribute that assists students who do not understand (have not mastered) some of the attributes. More specifically, the attributes represent mastery of a specific set of equations in mechanical efficiency, so the handout includes a detailed explanation about the formulas that student has not mastered and examples of how to use each formula. The appropriate tailored handouts were given to each student in the intervention group based on students' mastery profiles, skills, and misconceptions. The more skills not mastered by a student, the longer the handout. Students in the control group were not given any intervention material.

The post-test contained the same item stems as the pre-test but had some changes to the numbers and options to reduce any memory effects from the pre-test. The main item stems remained the same. Gierl et al. (2012) used the item model to generate different items with similar item difficulty. The same idea was used to create a post-test by keeping the same item stem and changing elements such as numbers and options in the pre-test. The pre-test and post-test were parallel but not identical to a memorization effect, thus the effect of the intervention could be evaluated.

In the post-test, students in the intervention group were asked if they studied the handout. If they had not, they were directed to the end of the test and asked to study the handout first. The

post-test was open for two days. Students in both the intervention and control groups were given the same instructions about the test, which includes the testing time is thirty minutes, and do not look up for the test answer from textbook or online.

Validity Arguments and Evidence

In this section, validity arguments and related evidence are presented for five sources of evidence separately. Each source of evidence has one or two validity arguments that are related to the test use. Evidence from different perspective are used to support the claim of each argument.

From the Source of Evidence Based on Test Content:

Validity Argument 1.1. The test appropriately measures mechanical efficiency using seven defined attributes

Evidence. I conducted extensive literature and database reviews to understand the target domain and document information about the skills required for it. I interviewed a subject matter expert (a middle-school physics teacher) on the target domain and asked her to verify the appropriateness and completeness of the defined attributes in the domain by previous subject matter experts (SMEs). In the interview, I first showed the SME the seven attributes that are defined by the other two SMEs previously, and then asked her if she think these seven attributes can represent the skills a student should know about mechanical efficiency domain according to the curriculum. Then I asked her if these attributes cover everything a student should know about mechanical efficiency. I randomly selected and interviewed three students who finished the pre-test about their understanding of the test. Specifically, I asked each student individually in a separate interview which chapter they think this test is design to measure.

Validity Argument 1.2. Students' performances on the test are only affected by their competence in the target domain.

Evidence. I interviewed the same three students who finished the pre-test individually about their testing experience. In the interview, I asked them if they were familiar with the test format, and if there were any questions or instructions they did not understand during the test.

From the Source of Evidence Based on Response Processes:

Validity argument 2. Students use the designed required attributes while engaging in problem-solving in the test.

Evidence. I proceeded with a think-aloud protocol with the same three students and documented their various ways of problem-solving. Each participant was asked to verbalize the problem-solving processes and strategies after completing the test. This process is recorded while participant looked at the questions one by one. I identified the problem-solving skills from the think-aloud data (voice recording) and matched the skills from students with the SMEs' defined attributes for the test.

From the Source of Evidence Based on Internal Structure:

Validity Argument 3. The association between the attributes should be low to reflect the multidimensional structure of the construct and fewer students would be expected to master a difficult attribute.

Evidence. I examined the correlation between different attributes and the estimated probability of mastery for all attributes. I interviewed the SMEs about whether the correlations between attributes and the probability of mastery for attributes are consistent with what to expect theoretically based on their knowledge of the attributes being measured. In the interview, I showed the SMEs correlations between attributes which greater than 0.9. I explain to them that

high correlation between two attributes means if a student master one of the attributes, he or she has a very high probability to also master another attribute. Then I asked the SMEs, according to their professional judgment, do they expect these pairs of attributes to be highly correlated to each other and what are the potential reason to explain the high correlations. After we reviewed the correlations, I showed the SMEs the probability of mastery for each attribute. I explained to them that the probability means after the model calibration, among all the students who took the test, that many percentages of students have mastered each attribute. Then I asked them if these probability for each attribute match with their expectation based on their knowledge of the attributes. I then summarized their responses in word format.

From the Source of Evidence Based on Relations to Other Variables:

Validity Argument 4. Estimated skill masteries can predict the end-of-year physics scores.

Evidence. A multiple regression analysis was performed using the seven attributes to predict the end-of-year physics score. From the multiple regression analysis, the R^2 is examined to describe the proportion of variability of the end-of-year physics score that can be explained by attribute mastery. The effect-size index, Cohen's f^2, was used to evaluate the magnitude of a treatment effect for each coefficient in the multiple regression. Cohen's $f^2 \geqslant 0.15$ indicates moderate effect and Cohen's $f^2 \geqslant 0.35$ indicates large effect-sizes (Cohen, 1988). In addition, the point-biserial correlations are computed and analyzed between the end-of-year physics score with each attribute.

From the Source of Evidence for Validity and Consequences of Testing:

Validity Argument 5. Diagnostic test feedback is helpful to students and teachers.

Evidence. Evidence for this argument seeks to demonstrated that the information obtained from the diagnostic test can be used to improve students' performance on mechanical efficiency and help teachers to adjust the curriculum based on students' current performance. To verify the usefulness of the feedback, I conducted an intervention study. First, I randomly selected two classes of students from the target population. One is randomly identified as the control group and the other as the experimental group. The class size in the selected school was approximately 40 for each group. Based on a student's profile from the pre-test, intervention handout material was generated for students in the intervention group that only included non-mastered skills. The full handout material which includes all the attributes is included in the Appendix D. Handouts were sent to students individually after the pre-test. Each student in the intervention group was then asked to study the handout without looking at any other materials for that day and to prepare for the post-test the next day. No individualized handouts were sent to the control group. Students in both groups were given a post-test the next day.

Based on the item parameters generated from the pre-test, student profiles for the post-test were estimated and compared to the profiles in the pre-test. Attributes mastery is compared between the pre-test and post-test on possible profile changes for both student. Both aggregated and individual attribute mastery changes were analyzed. In addition, for this comparison, it is important to determine the "equivalence" of the two groups. Although specific demographic information was not available, to compare the groups initially, the pre-test is used to establish baseline equivalence. Specifically, Cox's index is used to evaluate the baseline equivalence for two groups. Cox's index is an effect-size indication for baseline equivalence (Sánchez-Meca et al., 2003). In the current study, p_i is the probability of master on each attribute in the

experimental group and p_c is the probability of master on each attribute in the control group. ω is the small sample size correction, which is equal to $1 - \frac{3}{4N-9}$, where N is the total sample size.

$$d_{cox} = \omega \left[\ln\left(\frac{p_i}{1-p_i}\right) - \ln\left(\frac{p_c}{1-p_c}\right) \right] / 1.65$$

According to Ferguson (2009), the recommended minimum effect-size for the odds ratio index in social science is 2.0. Cox's index is the natural logarithm of the odds ratio divided by 1.65, so the minimum effect-size for the absolute value of Cox's index is 0.42 accordingly. When analyzing the specific attribute mastery changes in both groups, students were categorized into three groups. The first group is labeled as "same", indicating that mastery status stayed the same for the post-test when compared to the pre-test (i.e., was a master in both tests or classified as a non-mastered in both tests). The remaining two groups are "non-master to master" and "master to non-master", which means that many of students' mastery status change from non-master or master in the pre-test to master or non-master in post-test. The frequency of students in each of the three groups was computed for the control group and for the experimental group. Fisher's exact test is used as the index to test whether both groups showed a significant difference in the distribution of different mastery status change categories under the null that there are no differences. In addition, students in the intervention group were asked in a survey whether they found the handouts useful in helping them with non-mastered skills.

Results

To investigate whether the test was designed to measure ninth-grade students' mastery of seven attributes in the mechanical efficiency domain, I conducted literature reviews. In this case, the domain is a typical introduction to concepts of physics. The memorization and use of these seven equations is reasonably standard and as a result, the same required formulas for mechanical efficiency were found in the textbook and the curriculum requirements. Furthermore,

the third SME was showed the seven attributes that are defined by the other two SMEs previously, and then was asked if she think these seven attributes can represent the skills a student should know about mechanical efficiency domain according to the curriculum (appropriateness). Then I asked her if these attributes cover everything a student should know about mechanical efficiency (completeness). This SME confirmed that these seven attributes are the required skills and agreed on the appropriateness and completeness of the defined attributes.

To collect validity evidence based on test content and response processes for validity arguments 1.1, 1.2, and 2, three interviews were conducted with each of the three students. Students were first asked what skills or domain they think this test is measuring. They all identified that the target domain was mechanical efficiency. Students were then asked if they had any difficulties while taking the test and if the test format looked familiar to them. None of students experienced trouble while completing the test. One of them reported that the test was easy while another reported that it was similar to what they were usually given. After that, each student was asked to describe how they solved each question, including the formulas and the process of plugging in which numbers. Their answers were recorded and analyzed to compare with the SMEs' defined attributes.

Based on the results, all three students were able to identify the seven attributes. However, for specific items, not all the students could identify the same required attributes as the per the SMEs' item definition. Table 5.1. shows a summary of the think-aloud protocol results. If the student answered the question incorrectly, that question is marked as "incorrect" in the table. Student 1 identified 14 questions with the exact required attributes defined by the SMEs. For item 12, this student identified a different set of attributes than the SMEs' and applied the formulas incorrectly, but still got the correct answer. For items 14, 16, 17, 19, and 20, this

student used a different set of required attributes correctly and was also able to calculate the correct answer. The student used an extra attribute 4 to solve items 16, 17, and 19, and a completely different set of attributes to solve items 14 and 20, which implies potential multiple strategies for items 14, 16, 17, 19, and 20. However, the other two students identified the same required attributes with the SMEs except item 17, suggesting it is possible that only some students use a different strategy. For student 2, this student matched the required attributes for 17 out of 20 items. For items 15 and 17, this student incorrectly applied attribute 7, while the correct attribute was attribute 3. Although both formulas look very similar except for the subscript, neither item require distinguishing the subscript for the solution, and thus this student accidentally got item 15 correct and item 17 incorrect. The third student matched 18 items out of 20 with SMEs' defined attributes. For item 13, this student did not identify attribute 7 explicitly in the recording; however, based on the answer, attribute 7 was used in the problem-solving process to reach the final answer. For item 17, this student missed attribute 5 and got the item incorrect.

Table 5.1 SMEs' and Students' Identified Attributes for All Items

	Identified attributes			
	SMEs	Student 1	Student 2	Student 3
Item 1	1	1	1	1
Item 2	1	1	1	1
Item 3	2	2	2	2
Item 4	2	2	2	2
Item 5	4	4	4	4
Item 6	4	4	4	4
Item 7	5	5	5	5
Item 8	6	6	6	6
Item 9	6	6	6	6
Item 10	7	7	7	7
Item 11	1, 2	1, 2	1, 2	1, 2
Item 12	3, 6	3, 4, 5, 6	3, 6	3, 6
Item 13	4, 7	4, 7	4, 7	4
Item 14	6, 7	3, 4	4, 7	6, 7

Item 15	1, 2, 3	1, 2, 3	1, 2, 7	1, 2, 3
Item 16	1, 3, 5	1, 3, 4, 5	1, 3, 5	1, 3, 5
Item 17	1, 3, 5	1, 3, 4, 5	1, 7 (incorrect)	1, 3 (incorrect)
Item 18	1, 6, 7	1, 6, 7	1, 6, 7	1, 6, 7
Item 19	3, 5, 6	3, 4, 5, 6	3, 6	3, 5, 6
Item 20	3, 6, 7	2, 4, 5	3, 6, 7	3, 6, 7

To provide evidence based on the internal structure for validity argument 3, the association between the attributes was examined. In addition, these results were discussed with the SMEs to determine if they could be theoretically supported based on their knowledge of the construct. Table 5.2 provides the tetrachoric correlations between the seven attributes. Even though some attributes are highly correlated, reasonable explanations could be found after interviewing the SEMs. For example, the correlation between attributes 4 and 5 has the highest value of 0.91, which might be because both attributes are an element of attribute 3. This connection between attributes 4 and 5 makes it easier for students to remember both. The correlation between attributes 3 and 6 is the lowest, with a value of 0.35, which may be because both attributes are related to calculating W_{input}. Students may tend to remember one of the formulas rather than both. The correlations between attribute 3 and the rest of the attributes below 0.70, with the exception of attributes 2 and 7, which means attribute 3 has a low-to-moderate correlation with attributes 1, 4, 5, and 6. The rest of the correlations are all higher than 0.7, which may be due to the rest of the formulas being specific to mechanical efficiency. Also, these formulas were taught together as a bundle, whereas the formula in attribute 3 is also taught and used in four other domains. Thus, attribute 3 has a relatively lower correlation with other attributes while correlations among the rest of the attributes are high.

Table 5.2 Tetrachoric Correlations between the 7 Attributes

	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6	Attribute 7
Attribute 1	1.00	-	-	-	-	-	-
Attribute 2	0.76	1.00	-	-	-	-	-
Attribute 3	0.61	0.71	1.00	-	-	-	-
Attribute 4	0.89	0.84	0.70	1.00	-	-	-
Attribute 5	0.84	0.78	0.60	0.91	1.00	-	-
Attribute 6	0.89	0.72	0.35	0.80	0.77	1.00	-
Attribute 7	0.79	0.91	0.85	0.78	0.72	0.78	1.00

More evidence based on the internal structure for validity argument 3 is based on the proportion of students who have mastered each of the attributes. Specifically, it should be true that those fewer students have mastered the "difficult" attributes relative to the proportion of students who have mastered the "easier" attributes. Table 5.3 provides the probability of mastery for all the attributes. The probability of mastery for each attribute is higher than 0.6, and the average probability of mastery is 0.7. Attributes 4 and 7 have the lowest probability of mastery, which is consistent with the SME's expectations. She had expected attribute 4 to be the most challenging formula to apply, and attribute 7 as the one most often forgotten. The rest of the attributes were expected to be equally easy to master.

Table 5.3 Probability of Mastery for Each Attribute

	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6	Attribute 7
Probability of mastery	0.74	0.73	0.68	0.62	0.72	0.78	0.66

To inspect validity argument 4 as to whether estimated skill masteries can predict the end-of-year (EOY) physics scores, a multiple regression analysis was performed. This analysis allows for the investigation as to whether or not there is a relationship between the end-of-year physics score and the estimated attribute masteries. Table 5.4 provides a summary of the coefficients, significance tests, and effect-sizes for attribute effects. Because schools used

different end-of-year (EOY) physics tests, only 103 responses from the same school were used to perform the multiple regression. Results showed that the R^2statistics is 0.52, which means 52% of the variability in the EOY physics scores can be explained by the attributes. Recall that this test is based on one domain (i.e., chapter) in a class that covers a total of 13 domain. As a result, the end-of-year physics test is intended to assess knowledge across all 13 domains (i.e., chapter). A total of 52% is relatively high given that mastery on the 7 attributes describes mastery of only one domain. However, considering that 49% of the students were classified as all-mastered for mechanical efficiency, a relationship could exist in which mastery of one domain tends to lead to mastery of others.

Table 5.4. also reports the significance of the test results and the effect-size index for each coefficient. Cohen's $f^2 \geqslant 0.15$ indicates a moderate effect and Cohen's $f^2 \geqslant 0.35$ indicates large effect-sizes (Cohen, 1988). Only attribute 3 and the intercept showed a significant coefficient, and only attribute 3 showed a large effect-size. After interviewing the SMEs about this result, they indicated that this is because the other attributes are taught and used only in mechanical efficiency, whereas attribute 3 is also used in the other four domains. This unique feature of attribute 3 leads to a significant coefficient and relatively high correlation with the EOY score. In Table 5.5, point-biserial correlations are reported for the EOY physics score with each estimated attribute mastery. The average correlation is 0.46, the lowest correlation is 0.28, and the highest correlation is from attribute 3 with 0.70. The highest correlation (0.7) is between attribute 3 and the EOY physics score, which could also be explained by the fact that the other attributes are taught and used only in mechanical efficiency, whereas attribute 3 is also used in the other four domains. Hence the mastery of attribute 3 is highly related to the EOY physics score.

Table 5.4 Summary of Multiple Regression Analysis

	Estimate	Std. Error	t value	*p*-value	Cohen's f^2
(Intercept)	51.62	4.55	11.34	< 2e-16 ***	-0.00
Attribute 1	5.61	3.94	1.42	0.16	0.14
Attribute 2	0.58	4.90	0.12	0.91	0.01
Attribute 3	18.76	4.41	4.25	5.44e-05 ***	0.52
Attribute 4	1.34	5.07	0.27	0.79	0.03
Attribute 5	3.58	4.61	0.78	0.44	0.09
Attribute 6	5.02	4.37	1.15	0.25	0.11
Attribute 7	0.20	4.20	0.05	0.96	0.01

Note. Significance codes: 0 '***' 0.01 '**' 0.1 '*'.

Table 5.5 Point-biserial Correlations for Attributes and End-of-Year Physics Score

	Attr 1	Attr 2	Attr 3	Attr 4	Attr 5	Attr 6	Attr 7
EOY Physics Score	0.43	0.47	0.70	0.56	0.49	0.28	0.33

Finally, to evaluate validity argument 5, which is about whether or not diagnostic test feedback is helpful to students and teachers, a control group is compared to an experimental group for "usefulness" of the diagnostic feedback. Here usefulness is assessed by whether or not an additional study guide can be constructed based on the diagnostic feedback to improve mastery.

In this study, a total of 67 students were given a pre-test and a post-test. After deleting the duplicated responses and eliminating responses less than five minutes or greater than 30 minutes, 22 responses were retained in the experimental group and 15 responses in the control group. Each member of the experimental group was given an individually tailored handout after the pre-test based on his or her mastery profile and allowed time to study it. The control group received no additional material based on performance of the pre-test. A post-test was then administered to both groups, and both pre-test and post-test data were collected. Pre-test profiles were calibrated with the rest of the students, and the post-test profiles were calibrated using the same item

parameters in the pre-test, as the pre- and post-tests were considered parallel. After comparing the pre- and post-test profiles, results showed that a higher percentage of students had mastered more attributes after the intervention. Table 5.6 shows the summary of change for the seven attributes together. As shown, a higher percentage of students mastered more attributes, while a fewer percentage mastered less attributes in the post-test for the experimental group. The average-mastered attributes only increased by 0.3 in the control group, whereas the experimental group increased by 1.3. When asked, 66% of students in the experiment group found the handout helpful in the survey attached with the post-test.

Table 5.6 Attribute Mastery Comparison for Control Group and Experimental Group

	Master more attributes	Master same attributes	Master less attributes	Average mastered attributes in pre-test	Average mastered attributes in post-test
Control group	40.0%	33.3%	26.7%	3.6	3.9
Experimental group	68.2%	22.7%	9.1%	3.9	5.2

While it does appear that there were some differences between the control and experimental groups, Table 5.6 only shows aggregated changes between the control and experimental groups. As an approach to better explore the effects of the tailor handouts, Tables 5.7 and 5.8 indicate the change per attribute. Specifically, Table 5.7 contains information on the proportion of mastery for both groups in the pre- and post-test. Prior to testing for the post-test difference between the control and experimental groups, baseline equivalence needed to be assessed between the groups. The Cox's index is an effect-size indication for baseline equivalence (Sánchez-Meca et al., 2003), and this index was used for testing the proportion of attribute-mastery baseline equivalence in the pre-test for both groups. According to Ferguson (2009), the minimum effect-size for the absolute value of the Cox's index is 0.42. With the exception of attribute 6, all the other attributes showed baseline equivalence in the pre-test.

Given that evidence of baseline equivalence was achieved for all but attribute 6, a significance test was used to test for differences in the post-test. In the post-test, two sample z tests were computed for each attribute on the proportion of mastery in both groups. The *p*-value and significance level are listed in the last column of Table 5.7. The proportion of mastery for attributes 1 and 6 in both groups are significantly different; however, baseline equivalence was achieved for all but attribute 6 in the pre-test, so the difference in attribute 6 is meaningless. It maybe indicating that the experimental group showed considerable improvement on the mastery of attribute 1 after the intervention.

Table 5.7 Proportion of Mastery Comparison for Both Groups

	Pre-test			Post-test		
	Control Group	Experimental Group	Cox's Index	Control Group	Experimental Group	*p* value
attribute 1	60%	68%	-0.21	60%	86%	0.07*
attribute 2	47%	59%	-0.29	60%	64%	0.83
attribute 3	53%	55%	-0.05	60%	68%	0.61
attribute 4	53%	50%	0.07	40%	64%	0.16
attribute 5	47%	32%	0.38	60%	77%	0.26
attribute 6	47%	82%	-0.97	53%	91%	0.01***
attribute 7	53%	41%	0.29	53%	68%	0.36

Note. Significance codes: 0 '***' 0.01 '**' 0.1 '*'.

Table 5.8 summarizes the mastery status change in the pre- and post-test for both groups on each attribute. Then column labeled as "Same" provides the count of the number of students whose mastery status stayed the same when comparing the pre-test to the post-test (i.e., the student was a master in both tests or classified as a non-mastered in both tests), and the column labeled as "non-master-to-master" indicates the number of students whose mastery status changed from non-master in the pre-test to master in post-test. The column "master to non-master" indicates the opposite of the previous column, and thus indicates the number of students who were initially estimated to be a master in the pre-test and then where estimate to be a non-

master in the post-test. The percentage in the parenthesis is the row percentage for the student count in each category.

Fisher's exact test is the index for whether both groups showed a significant difference in the distribution of different mastery status-change categories. According to the results, attributes 4, 5, and 7 showed a significant difference. For example, in attribute 7, there is a big difference in the "same" category. The significance for attribute 7 might be contributed from the "same" category rather than "non-master to master" column, hence there is no evidence to support that the intervention helped on the mastery of attribute 7. In conclusion, no validity evidence was found to support the usefulness of the intervention material and the diagnostic feedback.

Table 5.8 Summary of Mastery Status Change in Pre- and Post-test for Both Groups

		Same	Non-master to master	Master to non-master	Fisher's Exact Test
Attribute1	Control Group	11 (73%)	2 (13%)	2 (13%)	*p*=.59
	Experimental Group	14 (64%)	6 (27%)	2 (9%)	
Attribute2	Control Group	7 (47%)	5 (33%)	3 (20%)	*p*=.91
	Experimental Group	9 (41%)	7 (32%)	6 (27%)	
Attribute3	Control Group	6 (40%)	5 (33%)	4 (27%)	*p*=.50
	Experimental Group	13 (59%)	6 (27%)	3 (14%)	
Attribute4	Control Group	11 (73%)	1 (7%)	3 (20%)	*p*=.09*
	Experimental Group	9 (41%)	8 (36%)	5 (23%)	
Attribute5	Control Group	13 (87%)	2 (13%)	0	*p*=.02*
	Experimental Group	10 (45%)	11 (50%)	1 (5%)	
Attribute6	Control Group	8 (53%)	4 (27%)	3 (20%)	*p*=.43
	Experimental Group	16 (73%)	4 (18%)	2 (9%)	

Attribute7	Control Group	7 (47%)	4 (27%)	4 (27%)	*p*=.03*
	Experimental Group	16 (73%)	6 (27%)	0	

Note. Significance codes: 0 '***' 0.01 '**' 0.1 '*'.

Discussion

Main Findings

This study investigates the validity of a DCM-scored physics test. After the literature review and interviews with SMEs and students, evidence based on test content was found to support validity arguments 1.1 and 1.2. Thus, evidence supported that statement that the test was designed to test mechanical efficiency, and there was no evidence to suggest that there were any sources of irrelevant variance in the test.

However, for validity argument 2, evidence based on the think-aloud protocol suggested that some students did not use the intended attributes while answering specific items of the test. One student applied an alternative strategy for solving questions 14, 16, 17, 19, and 20, while the other two identified the same attributes that had been identified by the SMEs. The student employing an alternative strategy used an additional attribute, attribute 4, to solve items 16, 17, and 19. In addition, this student used a completely different set of attributes to solve items 14 and 20, which implies a potential for multiple strategies with respect to items 14, 16, 17, 19, and 20. Whereas the other two students identified the same attributes as the SMEs with the exception of item 17. Based on these results, the Q-matrix might not specify the only problem-solving process that is used to answer the questions, which is a violation of one basic assumption of most DCM. While this violation is expected to threaten the validity of the test, it does appear that it would be less of an impact on items 16, 17, and 19 than on items 14 and 20 because that student identified an extra attribute rather than a totally different set of attributes. For example, for item

16, the SMEs identified required attributes to be 1, 3, 5, whereas student 2 identified attributes 1, 3, 4, 5 as the required attributes. The only difference between SEMs' judgement and student 2's judgement is attribute 4. However, for item 14, the SMEs identified required attributes to be 6, 7, whereas student 2 identified attributes 3, 4 as the required attributes. These two sets of attributes are totally different without any overlap. Suggesting that it would be less of an impact on items 16, than on items 14.

For validity argument 3, evidence based on the internal structure demonstrated a range of associations between the 7 attributes. Even though some attributes were highly correlated, reasonable explanations could be found after interviewing the SEMs. For example, he correlation of attributes 4 and 5 has the highest value of 0.91, suggesting that in this particular sample these two attributes are not well distinguished, whereas correlations of attributes 3 and 6 is the lowest, with a value of 0.35, suggesting that the attributes are in fact distinct. Upon further review, the high association can be explained based on the fact that both attributes 4 and 5 are an element of attribute 3. This connection between attributes 4 and 5 makes it easier for students to remember both. In addition to the general associations, evidence of the internal structure was evaluated based on the proportion of student expected to have master each attribute. In this sample, it is predicted that difficult attributes did have a smaller proportion students who had mastered the attribute compared to attributes that were perceived to be easier by the subject matter experts. This result provides some validity evidence for the internal structure of the test and the construct being assessed.

With respect to providing evidence for relations to other variables—i.e., validity argument 4—the result showed that more than 50% of the variance in end-of-year physics scores could be explained by the examinees estimated mastery profiles. In addition, primarily attribute 3

is highly correlated with the end-of-year physics score. Although this result was not originally expected, upon further reflection, this result can be explain based on the differences of the two constructs. The end of grade test represents the overall knowledge whereas mechanical efficiency is just one of 13 domains for eighth-grade students. It seems unlikely that knowledge of a single domain could completely predict the final score describing knowledge on all 13 domains. That said, one attribute is relevant in other domains of physics, attribute 3. In the follow-up interview with an SME, she identified that attribute 3 is also used in the other four domains, which explains the high correlation for attribute 3 and the EOY score. The fact that attribute 3 predicts the end-of-year score and not the other attributes does provide validity evidence with respect to its predicted relationship.

For the last validity argument—whether diagnostic test feedback is helpful to students and teachers—the result provides little evidence that a student's mastery profile could be used to provide a tailored study plan in the way that was provided in this design, which in turn could improve mastery status. Even though both attributes 1 and 6 are significantly different between the pre and post-test, only attribute 1 satisfied baseline equivalence in the pre-test. In fact, attribute 1 did show considerable improvement for the experimental group with respect to mastery of attribute 1 after the intervention when compared to the control group. In addition, when comparing the mastery status change in the pre- and post-test for both groups, attributes 4, 5, and 7 did demonstrate a significant difference. However, this significant result may not have been specifically to an increase in the "non-master to master" group in the experimental group. Thus, no validity evidence was found to support the last validity argument.

Collectively, all the evidence suggested that this diagnostic test is lacks sufficient validity for the purpose of providing tailored remediation to students. However, some evidence was

found to support validity arguments with respect to the test content, internal structure, and relations to other variables. In general, the DCM feedback did not appear to be useful in diagnosing student weaknesses and providing the teacher with extra fine-grain information. In addition, a validity threat was detected on the source of response processes in which one student revealed alternative strategy in solving the problems for five items. Upon further inspection, the student identified only one extra attribute for three items. It is believed that this does lessen the impact on those three items. In conclusion, even though the study results showed validity threats to the diagnostic test, this study potentially provides practical guidance for researchers interested in validity evidence in DCM applications and encourages future studies to include validity analyses.

Limitations and Future Directions

The major limitation of the study is the small sample size in the control and the intervention groups. In the future, additional students could be included in the intervention study to produce a more reliable comparison result. This study was also limited to examining the item characteristics. For example, a distractor analysis of the items could reveal why students used a different set of attributes and still obtained a correct item. Furthermore, more research could be performed on the study's generalizability and help future researchers analyze the validity of their DCM studies in a more systematic way.

CHAPTER VI: PAPER 3: INCORPORATING THE RESPONSE-TIME FOR A PHYSICS DIAGNOSTIC ASSESSMENT

Introduction

Educators are demanding diagnostic information from assessments that go above and beyond a traditional score, only places students along a unidimensional scale (Huff & Goodman, 2007). This type of diagnostic information can only be provided by using multidimensional psychometric models in the educational setting. Using a multidimensional model would provide information about more specific dimensions that a student should focus on improving while also indicating the students' strengths. Diagnostic Classification Modeling (DCM) is a type of multidimensional model that focuses on diagnosing a student's skill profile by outputting statistical information on skill mastery or non-mastery (Sessoms & Henson, 2018). DCM could potentially meet the demands of educators and produce useful diagnostic information from assessments.

However, multidimensional models tend to require more items because of the needs to estimate a score or status for each dimension, which in turn requires more time from students and teachers. According to a test report by the Council of the Great City Schools, the nation's urban public schools administer many tests. Statistically, the average student takes roughly 112 tests between pre-K and grade 12. Consequently, students spend a fair amount of time taking tests. The report showed that the time needed on average to take mandatory tests is approximately 25 hours or between four and five days of school per school year. A similar issue was discussed by Zhu (2016) who indicated that the academic burden on Chinese middle-school students was heavier despite the series of policies designed to ease the academic burden.

While the burden is higher for multidimensional assessment, not all modeling approaches are the same. With the same number of items, DCM can typically extract more dimensions than other continuous latent-variable multidimensional models such as MIRT, because as opposed to locating a student on multiple continuous scales, information is only used to obtain mastery or nonmastery on a set of dichotomous dimensions (Templin & Bradshaw, 2013). By making the dimensions dichotomous, fewer items are needed for a reliable result. Furthermore, besides reducing each dimension to a dichotomous master decision, other methods should continue to be explored to help increase accuracy and reduce test length. For example, studies have revealed that it is possible to use variables other than student responses to increase the accuracy of the ability estimation (de la Torre, 2009; Ferrando & Lorenzo-Seva, 2007).

The current study investigates the joint response-time DCM model to determine the potential benefit of increasing the accuracy of the individual profile estimation. Noting that if additional information such as response-time can help with estimation, it may then be possible to create shorter tests that still provide useful formative information about each student.

Literature Review

Many studies have investigated the incorporation of addition information other than item responses into psychometric models as a way to increase estimation accuracy. Mislevy (1986) displayed the utilization of ancillary information in increasing item-parameter estimation accuracy for item response models. De la Torre (2009) introduced a general model that incorporates both ancillary variables and the correlational structure of latent abilities to improve estimation accuracy. The results showed that using ancillary variables provides better ability estimates, which in turn would suggest that a shorter test with ancillary information could provide the same accuracy as a longer test without the ancillary information.

One of the commonly discussed ancillary variables that may be related to ability and, thus, useful to model is response-times. Van der Linden (2006) proposed a lognormal model for Response-times (RT) on test items, and it showed an excellent fit to the response times for the adaptive version of a test from the Armed Services Vocational Aptitude Battery. The author stated that this model could be used as a "plug-in" model that would allow for the simultaneous modeling of item responses and item response-times. With the addition of response times to the model, they could be used to help improving accuracy of the updates of the ability estimates in adaptive testing (van der Linden, 2006). Ferrando & Lorenzo-Seva (2007) presented an item response theory model for incorporating response-time data when modeling binary personality items and concluded that a gain exists in the precision of estimating individual-trait levels when response-times are used.

Similar to unidimensional continuous models, response-time has also been investigated for diagnostic tests. (e.g. De Boeck & Jeon, 2019, Zhan et al. , 2018). Of particular importance to this study is the model proposed by Zhan el al. (2018) refer to as the joint response-time deterministic input, noisy "and" gate (JRT-DINA) model. They showed that assuming there is a relationship between the response time and the mastery, there were improvements on attribute and profile correct classification rates.

In the current study, I plan to calibrate a specific model using physics data with intent of improving estimation by also incorporating RT. An investigation was conducted on whether incorporating response-time can result in an improvement in accuracy for diagnosis. The following paragraphs first introduce the Deterministic Input; Noisy "And" gate (DINA; Junker & Sijtsma, 2001), and then define the JRT-DINA.

The DINA model is selected as the diagnostic model for all the items. The DINA model is defined as follows:

$$P(X_{ij} = 1|\xi_{ij}) = (1 - s_j)^{\xi_{ij}} g_j{}^{(1-\xi_{ij})}. \quad (1)$$

As defined in equation, $\xi_{ij}$is an indicator for whether students have mastered all the required attributes. Note that based on how ξ_{ij} is defined (mastered all required attributes versus lack mastery of at least one attribute), the DINA model is a non-compensatory conjunctive model that requires examinees to master all the needed attributes to have a high probability of providing the correct response. If an examinee masters all the required attributes, ξ_{ij}equals one. However, if one or more attributes have not been mastered, then, $\xi_{ij} = 0$. Specifically, for examinee i for item j,

$$\xi_{ij} = \prod_{k=1}^{K} \alpha_{ik}{}^{q_{jk}}. \quad (2)$$

Thus, examinees lacking mastery of all required attributes are assumed to be similar to those lacking mastery of only one required attribute.

Given the value of ξ_{ij}, the probability of a correct response is defined using two parameters, s_j and g_j. Where s_j, the slip parameter, represents the probability of obtaining an incorrect answer while mastering all of the needed attributes. In another words, s_j is the estimated probability of "slipping" up. The parameter g_j, the guessing parameter, represents the probability of answering a question correctly without mastering at least one of the required attributes, which represents the chances of "guessing" a correct response. Equations (3) and (4).provide the formal definition of the slip and guess parameters

$$s_j = P(X_{ij} = 1|\xi_{ij} = 1), \quad (3)$$

$$g_j = P(X_{ij} = 1|\xi_{ij} = 0). \quad (4)$$

For these two parameters, the DINA model has one constraint, (1-s_j) > g_j, so the probability of not slipping (i.e., getting the item correct for masters) will always exceed the probability of guessing (i.e., getting the item right for nonmasters). This constraint ensures that masters are always predicted to perform as well or better than nonmasters.

According to Zhan et al. (2018), the JRT-DINA model jointly models the probability of a correct response using a DINA model and the predicted distribution of the response-time. Equation (5) gives the model for response time such that T_{ni} is the observed item response time of person n to item i; ζ_i is the time-intensity parameter that represents the populations' average time needed to complete item i; τ_n is the person-speed parameter that represents the average speed of person n on a test; and ε_{ni} is the normally distributed error term, indicating that the RT model belongs to a lognormal family.

$$\log(T_{ni}) = \zeta_i - \tau_n + \varepsilon_{ni}, \quad \varepsilon_{ni} \sim N\left(0, \sigma_{\varepsilon_i}^2\right), \tag{5}$$

In addition to response time the model simultaneously models the probability of a correct response using an alternative parameterization of the DINA model. Specifically, when modeling the probability of a correct response, $P(Y_{ni} = 1)$, for person n answering item i correctly the model still uses the slip and guess parameters, s_i and g_i. These two parameters are still defined in the same way using equation (3) and (4), and as a result, they represent the probability of an incorrect answer for masters and the probability of a correct response for nonmasters, respective. The model also specifies item masters and item nonmasters, using $\prod_{k=1}^{K} \alpha_{nk}^{q_{ik}}$ measures whether the examinee master all the required attributes. Using the slip, guess, profile and Q-matrix, the probability is described in equation

$$P(Y_{ni} = 1) = g_i + (1 - s_i - g_i) \prod_{k=1}^{K} \alpha_{nk}^{q_{ik}}, \tag{6}$$

To add in a constraint that the probability cannot exceed 1 and must be greater than 0, the JRT DINA, instead models the log-odds the probability of a correct response. As a result, this model is transformed such that s_i and g_i are re-parametrized as in equation (7) and (8).

$$\beta_i = logit(g_i), \tag{7}$$

$$\delta_i = logit(1 - s_i) - logit(g_i), \tag{8}$$

Given the transformation $logit(x) = logit(\frac{x}{1-x})$, the log-odds of the probability of a correct response can then be defined as seen in equation (9).

$$logit\big(P(Y_{ni} = 1)\big) = \beta_i + \delta_i \prod_{k=1}^{K} \alpha_{nk}^{q_{ik}}. \tag{9}$$

As these two models have been defined, they separately model the probability of a correct response and the expected distribution of time. So that time spent on an item also provides some information of mastery. Zhan et al. (2018) used a higher-order structure to link the correlated attributes to the expected time latent variable.

$$logit\big(P(\alpha_{nk} = 1)\big) = \gamma_k \theta_n - \lambda_k. \tag{10}$$

Because the joint model uses the hierarchical modeling framework, item parameters of the JRT-DINA model are assumed to follow a trivariate normal distribution. The distribution with the respective mean vectors and variance and covariance matrix is as in equation (11).

$$\boldsymbol{\Psi}_i = \begin{pmatrix} \beta_i \\ \delta_i \\ \zeta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\beta \\ \mu_\delta \\ \mu_\zeta \end{pmatrix}, \Sigma_{item} \right). \tag{11}$$

Similarly, the hierarchical person parameters of the JRT-DINA model, which represented by θ_n, τ_n, are assumed to follow a bivariate normal distribution. The distribution with the respective mean vectors and variance and covariance matrix is in equation (12).

$$\boldsymbol{\Theta}_{n=} \begin{pmatrix} \theta_n \\ \tau_n \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\theta \\ \mu_\tau \end{pmatrix}, \Sigma_{person} \right), \Sigma_{person} = \begin{pmatrix} \sigma_\theta^2 & \rho_{\theta\tau}\sigma_\theta\sigma_\tau \\ \rho_{\theta\tau}\sigma_\theta\sigma_\tau & \sigma_\tau^2 \end{pmatrix}. \tag{12}$$

The model calibration was conducted through JAGS (Version 4.3.0; Plummer, 2017) and the R2jags package (Version 0.7-1; Su & Yajima, 2021) in R (Version 4.0.3 64-bit; R Core Team, 2020). Response time has been utilized to help increase estimation precision not only in unidimensional model, but also latest research presented that response time can be incorporated into multidimensional diagnostic model. This integrated model is consisting of two models; the response time model and the probability model of getting an answer correct. By calibrating a specific model using physics response and response time data, we expected to see an improvement in the estimation accuracy.

Purpose

The current literature has proposed some examples of diagnostic classification models that incorporate response-times. However, there are limited examples using authentic diagnostic test applications that also compare the results to a more traditional DCMs that do not incorporate auxiliary information such as response time. The purpose of the current study is to investigate the use of process data (e.g., response time data) for a low-stakes, diagnostic computer-based test in middle-school physics. Three different analyses are conducted for this study. First, response data and response time (representing the process data) from a diagnostic physics test was used to calibrate the JRT-DINA model (Zhan et al. 2018) and evaluate the model fit. Then, only the response data was used to calibrate the DINA model (Junker & Sijtsma, 2001) and evaluate the model fit. Second, an analysis of the change both in individual classification and in posterior probability of mastery for each attribute were conducted by comparing the output from JRT-DINA and DINA model. Lastly, a simulation study compared the estimated correct classification rate (CCR) of profiles when using the estimated JRT-DINA to the CCR when only using the

estimated DINA model. This comparison can provide evidence for whether or not response time is expected to improve accuracy.

Research Questions

1. Does the item response and item response-time data from the physics diagnostic test fit the JRT-DINA model?

2. How do the individual classification and the posterior probability of mastery for each attribute change between the JRT-DINA and DINA models?

3. Does the JRT-DINA model provide higher estimated correct classification rate (CCR) of examinee attribute mastery profile when compared to CCRs when only using the DINA model?

Methods

Data Sources

Response and process data was obtained from a diagnostic physics test. The test is an existing middle-school physics DCM-scored test including 20 items measuring the domain of mechanical efficiency. The test was constructed based on seven attributes with four parallel forms. Test items were administrated through Qualtrics (Qualtrics, Provo, UT). The test was available to 524 ninth-grade students. While no time limit is built into the Qualtrics program, students were asked to complete the test in thirty minutes. Students were informed that they cannot go back and change answers once they have moved to the next item. For each student on each item, the answer and time spent on the item were collected through the testing platform, Qualtrics (Qualtrics, Provo, UT). Only fully completed tests were recorded. Students can take the test anytime during the two days the Qualtrics link is open. The data was collected and downloaded through Qualtrics.

Analyses

This study includes three analyses. In the first analysis, the JRT-DINA model (Zhan et al. 2018) was used to simultaneously fit students' item response data and item response-time data from a diagnostic physics test. Specifically, using the code provided in Zhan et al. (2018). The same procedure in Zhan et al. (2018) was conducted for the first analysis. Two Markov chains with random starting values were used, and each chain had 10,000 iterations. The First 5,000 iterations in each chain were treated as the burn-in, and the remaining 5,000 iterations were used for model parameter inference.

The Gelman–Rubin convergence statistic $\hat{R}$ (Brooks & Gelman ,1998) was computed to assess the convergence of every parameter. Brooks & Gelman (1998) suggests that for the purpose of monitoring convergence, when $\hat{R}$ <1.2 convergence has been reached. The Gelman–Rubin convergence statistic $\hat{R}$ is generated and reported in the JAGS output.

To evaluate the model-data fit, a posterior predictive checking method was used. Posterior predictive checks are a common approach to determining fit in cases where we try to fit the model to an observed dataset. In general, these approaches compared the predicted data from the fitted model to the observed data. In this particular case, I use a method that was describe by Almond et al. (2015), in which the discrepancy measure $D(\boldsymbol{y}, \boldsymbol{\omega})$ for the response model, and $D(\log(\boldsymbol{T}), \boldsymbol{v})$ for the response-time model are computed using the estimated parameters and the observed responses. These two indices in equations (13) and (14) are essentially a measure of the standardized distance between the observed and predicted values of all observations.

$$D(\boldsymbol{y}, \boldsymbol{\omega}) = \sum_{n=1}^{N} \sum_{i=1}^{I} \left(\frac{Y_{ni} - P(Y_{ni} = 1)}{\sqrt{P(Y_{ni} = 1)(1 - P(Y_{ni} = 1))}} \right)^2 \quad (13)$$

$$D(\log(\boldsymbol{T}), \boldsymbol{v}) = \sum_{n=1}^{N}\sum_{i=1}^{I}\left(\frac{\log(T_{ni}) - (\zeta_i - \tau_n)}{\sigma_{\varepsilon_i}}\right)^2 \quad (14)$$

These values will be seen that values corresponding to the estimated model using the observed data.

Secondly, the posterior distribution of these indices in equation (13) were computed under the assumption that the model is true. This posterior distribution is estimated through a process that samples from the posterior of model parameters and simulates data (i.e., the assumed model is known to be true). Specifically, for the response model, a randomly draw from the posterior distribution of all item parameters β_i, δ_i and the attribute probability $P(\alpha_{nk} = 1)$ is obtained. The values of $P(\alpha_{nk} = 1)$ are rounded to compute a simulated sample of examinees mastery profiles α_{nk}. Next, using the randomly drawn model parameters and the simulated examinees (drawn from the distribution of examinees), the values of $P(Y_{ni}^{rep} = 1)$ are computed using the formula specified in Equation (15). Rounded $P(Y_{ni}^{rep} = 1)$ to get Y_{ni}^{rep}. Using the simulated replication, the value $D(\boldsymbol{y}^{rep}, \boldsymbol{\omega})$ is computed and then compared to the original $D(\boldsymbol{y}, \boldsymbol{\omega})$. This process is repeated 1,000 times. The proportion of times that $D(\boldsymbol{y}^{rep}, \boldsymbol{\omega}) \geq D(\boldsymbol{y}, \boldsymbol{\omega})$ is the PPP-value.

A similar process is also used for the response time model. Specifically, the posterior distribution of indices in equation (13) were computed under the assumption that the model is true. Then I randomly drew from the posterior distribution of item parameter ζ_i, σ_{ε_i}, and person parameter τ_n. Computed the $\log(T_{ni}^{rep})$ using formula $\log(T_{ni}^{rep}) = \zeta_i - \tau_n + \varepsilon_{ni}$, $\varepsilon_{ni} \sim N(0, \sigma_{\varepsilon_i}^2)$. Computed $D(\log(\boldsymbol{T}^{rep}), \boldsymbol{v})$ and $D(\log(\boldsymbol{T}), \boldsymbol{v})$using equation (14) and compare them. Repeated the process for 1,000 times. The proportion of times that

$D(\log(\boldsymbol{T}^{rep}), \boldsymbol{v}) \geq D(\log(\boldsymbol{T}), \boldsymbol{v})$ is the PPP-value. The posterior predictive probability (PPP) value for response and response time models are reported.

$$\text{PPP-value} = P(D(\boldsymbol{y}^{rep}, \boldsymbol{\omega}) \geq D(\boldsymbol{y}, \boldsymbol{\omega})) = P(D(\log(\boldsymbol{T}^{rep}), \boldsymbol{v}) \geq D(\log(\boldsymbol{T}), \boldsymbol{v}) \quad (15)$$

PPP-values near .5 indicate that there are no systematic differences between the realized and predictive values, and thus indicate adequate fit of a model (Almond et al., 2015).

In the second analysis, the item parameters including slip, guess and Item discriminating index (IDI) statistics for both JRT-DINA and DINA models are compared. The slip and guess parameters are defined in equations (6) and (9). The IDI statistics is equal to $1 - s - g$. The slip and guess parameters were compared for both model regarding different values. The IDI statistics less than 0.4 indicates less than a 0.4 difference in the probability of getting the item correct when mastered all the required attributes ($1 - s$) and the probability of getting the item correct when not mastered all the required attributes (g). IDI < 0.4 indicates a poor discriminating ability on the students. Moreover, the estimated individual mastery profiles from JRT-DINA and DINA were analyzed. Specifically, the agreement rate between the two estimated models' is computed, which is also used to compute the proportion of the profiles that are different respect to estimated mastery or nonmastery for the two models.

While agreement rates can be used, they do not incorporate the level of "certainty" in classification. For example, a posterior probability of mastery of 0.95 and 0.70 would both provide some evidence for mastery. However, the 0.95 would suggest stronger evidence of mastery compared to the 0.70. Because additional information might be expected to improve the level of certainty, in addition to agreement rates, the extent to which posterior probability are close to 0.00 and 1.00 is computed (i.e., the farther from 0.50 the more evidence of mastery or nonmaster is in the data. As an index to summarize the amount of evidence in the data is

provided for mastery or nonmastery, the average distance d between the posterior probability of mastery for each attribute and 0.50 was calculated for both JRT-DINA and DINA for comparison.

$$d = \frac{\sum_{n=1}^{N} |P(\alpha_{nk} = 1) - 0.5|}{N} \tag{16}$$

It is expected that the value d will be larger for the JRT-DINA because of the use of response time when compared to the results from the DINA model.

Lastly, the estimates of correct classification rate (CCR) on profiles are compared between the JRT-DINA and DINA using a simulation study (Henson et al., 2018; Wang et al., 2015). To estimate CCRs, data is simulated so that truth is known and, this, the CCR can be computed assuming that the model and estimate parameters are truth. The process for computing CCRs based on a simulation is as follows.

1. Use response and response time to calibrate DINA and JRT-DINA model separately. Store the parameters for both models.

2. Randomly simulated 397 students' profile data by sampling with replacement from all the calibrated student profiles. These profiles are considered as the true profile of students. The diagnostic test sample size is 397.

3. The response data set was simulated separately for the DINA and JRT-DINA using the designed Q-matrix and the profile data produced in step 2. Specifically, for DINA model, calculate the probability of getting an item correct ($P(Y_{ni} = 1)$) using DINA model for all the randomly simulated student profiles in step 2; Then sample the response from a Bernoulli distribution with the $P(Y_{ni} = 1)$ that is calculated before. For JRT-DINA model, calculate the probability of getting an item correct ($P(Y_{ni} = 1)$) using JRT-DINA model for all the randomly simulated student profiles in step 1; Then sample the response from a Bernoulli distribution with

the $P(Y_{ni} = 1)$ that is calculated before and sample the response time from equation (5). In the simulated data, the parameters were the same as the calibrated parameters for both models.

4. The simulated response data set was calibrated using the simulated model (DINA or JRT-DINA) and the same item/person parameters obtained in step 1 to compute the estimated profiles.

5. The correct classification rate (CCR) was estimated by comparing the estimated profiles to the true profiles in the second step.

6. The process was repeated 1 to 4 for 50 times. Took the average CCR for both models and compared the value.

Results

The purpose of this study is to explore the usefulness of incorporating additional information in the form of item response time into a model with respect to a low-stakes, diagnostic computer-based test in middle-school physics. Item response and response-time data were collected from a DCM-scored test for middle- school physics mechanical efficiency. The test consists of 20 multiple-choice items, and it was administered through Qualtrics (Qualtrics, Provo, UT) online. Students' responses were recorded and transformed into 0 indicating an incorrect response and 1 for a correct response. Each item was shown on one page for the Qualtrics administration. The time that a student spent on that page was defined based on the moment the student "entered" the page to the moment the student "left" that page and moved to the next item and was recorded through Qualtrics. Both response (correct/incorrect) and response-time data were analyzed for JRT-DINA model in JAGS (Version 4.3.0; Plummer, 2017) and the R2jags package (Version 0.7-1; Su & Yajima, 2021) in R (Version 4.0.3 64-bit; R Core Team, 2020).

To check the status of convergence, Gelman–Rubin convergence statistics $\hat{R}$ (Brooks & Gelman, 1998) were used to evaluate each parameter. Brooks & Gelman (1998) suggested that for the purpose of monitoring convergence, a performance is considered acceptable when $\hat{R}$ is less than 1.2. In this study, all estimates have a $\hat{R}$ =1, indicating convergence.

A posterior-predictive check method was used separately for the response model and the response-time model to evaluate the model fit. In this method, the discrepancy measure $D(\boldsymbol{y}, \boldsymbol{\omega})$ for the response model and $D(\log(\boldsymbol{T}), \boldsymbol{\upsilon})$ for the response-time model were computed and compared to $D(\boldsymbol{y}^{rep}, \boldsymbol{\omega})$ and $D(\log(\boldsymbol{T}^{rep}), \boldsymbol{\upsilon})$ for the posterior predictive probability (PPP) values. The PPP-values for the response and response-time models are 0.595 and 0.577. Almond et al. (2015) suggested that PPP values around 0.5 can indicate adequate data-model fit concerning the characteristics the measure is targeting. The PPP values for both models are close to 0.5, indicating an adequate model-data fit for both the response and response-time model.

After evaluating the convergence status and the model-data fit, estimated item parameters were analyzed. Table 6.1 displays the estimated JRT-DINA item parameters for the DCM-scored test in which β and δ are the reparametrized guessing parameter *g* and the slipping parameter *s*. Table 6.2 compares the item parameters statistics for both JRT-DINA and DINA models. Item discriminating index (IDI) statistics is equal to 1 – *s* – *g*. Items 1, 2, 4, and 8 have IDI statistics less than 0.4 for both models, which indicates a poor discriminating ability on the students. Items 1, 2, 4, 8, 9, and 10 have a guessing parameter estimate higher than 0.50 for both models. These items only measure one attribute, which indicates that students having not mastered the measured attribute still, have a greater than a 50% chance of getting the item correct. Items that measure two or three attributes have lower guessing parameters. Most of the items have a relatively small slipping parameter and this means that students who should get the item right because they have

mastered the measured attribute tend to get the item right. The highest slip parameter is item 12 (0.23).

When comparing the two models, both guess and slip parameters are very similar for the two models. For example, items 1, 2, 4, 8, 9, and 10 have a guessing parameter estimate higher than 0.50 for both models. In addition, the IDI statistics is similar for both models. For example, items 1, 2, 4, and 8 have IDI statistics less than 0.4 for both models, which indicates less than a 0.4 difference in the probability of getting the item correct when mastered all the required attributes ($1 - s$) and the probability of getting the item correct when not mastered all the required attributes (g).The correlation between β and ζ is -0.58, which indicates that more difficult items need more time to solve.

Table 6.1 Estimated Item Parameters for the DCM-scored Test

	β	δ	ζ	σ_ε	IDI
Item 1	0.74	3.70	2.93	0.64	0.31
Item 2	0.90	3.65	2.83	0.59	0.28
Item 3	0.12	3.02	3.16	0.65	0.43
Item 4	1.22	3.88	3.19	0.74	0.22
Item 5	-0.31	3.86	3.23	0.63	0.55
Item 6	-0.54	4.56	3.44	0.73	0.61
Item 7	-1.15	4.24	2.95	0.46	0.72
Item 8	1.88	4.36	2.64	0.69	0.13
Item 9	0.24	3.65	3.18	0.52	0.42
Item 10	0.26	3.57	3.94	0.64	0.41
Item 11	-0.22	4.35	3.38	0.48	0.54
Item 12	-0.86	2.10	3.71	0.62	0.48
Item 13	-0.99	3.76	3.79	0.65	0.67
Item 14	-0.13	3.62	3.43	0.52	0.50
Item 15	0.11	4.09	3.32	0.74	0.46
Item 16	-0.48	3.31	3.48	0.87	0.56
Item 17	-1.10	2.96	3.43	0.76	0.62
Item 18	-0.68	3.52	3.40	0.92	0.61
Item 19	-1.35	3.16	3.40	0.91	0.65
Item 20	-0.58	3.30	3.66	0.96	0.58

Table 6.2 Item Parameters Comparison for JRT-DINA and DINA

	JRT-DINA			DINA		
	g	*s*	IDI	*g*	*s*	IDI
Item1	0.68	0.01	0.31	0.80	0.01	0.18
Item2	0.71	0.01	0.28	0.83	0.01	0.16
Item3	0.53	0.04	0.43	0.40	0.06	0.54
Item4	0.77	0.01	0.22	0.71	0.01	0.28
Item5	0.42	0.03	0.55	0.40	0.03	0.58
Item6	0.37	0.02	0.61	0.32	0.01	0.67
Item7	0.24	0.04	0.72	0.02	0.04	0.94
Item8	0.87	0.00	0.13	0.88	0.00	0.12
Item9	0.56	0.02	0.42	0.57	0.02	0.42
Item10	0.56	0.02	0.41	0.54	0.02	0.44
Item11	0.44	0.02	0.54	0.42	0.01	0.58
Item12	0.30	0.23	0.48	0.32	0.24	0.44
Item13	0.27	0.06	0.67	0.27	0.05	0.69
Item14	0.47	0.03	0.50	0.47	0.03	0.50
Item15	0.53	0.01	0.46	0.50	0.01	0.49
Item16	0.38	0.06	0.56	0.38	0.08	0.54
Item17	0.25	0.14	0.62	0.25	0.17	0.58
Item18	0.34	0.06	0.61	0.34	0.06	0.60
Item19	0.21	0.14	0.65	0.20	0.16	0.65
Item20	0.36	0.06	0.58	0.36	0.09	0.55

In addition to comparing the item parameters, the attribute profiles of mastery can be compared. Specifically, Table 6.3 presents the proportion of mastery on each attribute for both models. With the exception of attribute 1, the proportion of mastery estimated by the JRT-DINA is approximately equal or smaller than the DINA estimated proportion of mastery. One potential explanation for the exception with attribute 1 could be due to the fact that items 1 and 2 only measured attribute 1, and the IDI statistics showed that JRT-DINA model has higher IDI for item 1 and 2 compared to DINA model. The higher IDI suggested a higher discrimination on item 1 and 2 for JRT-DINA model. So it is possible that based on item 1 and 2, JRT-DINA model classifies more students as mastery on attribute 1 than DINA model.

Table 6.3 Proportion of Mastery on Each Attribute for Both Models

	Proportion of Mastery	
	JRT-DINA	DINA
Attribute 1	0.89	0.74
Attribute 2	0.61	0.73
Attribute 3	0.55	0.68
Attribute 4	0.63	0.62
Attribute 5	0.60	0.72
Attribute 6	0.68	0.78
Attribute 7	0.56	0.66

Table 6.4 presents the agreement rates on attribute and vector levels when comparing the DINA to the JRT-DINA models. Results show that both models agree, on average, 85% of the time across all attributes, and the agreement rate is 44% when comparing perfect agreement of mastery profiles. Of the profiles for complete agreement between profiles, 84% were classified as masters on all attributes.

In addition to general agreement rates, it was believed that the modeling response-time would provide additional information that could be used to indicate attribute mastery or non-mastery. In diagnostic models, the increased information was expected to result in posterior probabilities of mastery that are closer to 0 or 1. To assess this, quantity *d* is computed. Table 6.4 presents the distance *d* between the posterior probability of mastery for each attribute on both models. Ideally, by adding in the response-time variable, distance *d* should increase because the model is more certain about the estimation. However, the results show that with the exception of attribute 3, all other attributes were higher *d* when using the DINA model for estimation.

These results indicate that by adding in the response-time model, the JRT-DINA did not improve the estimation of classification. Ideally, by adding the response-time into the model, person speed parameter τ should have some correlation with general ability variable θ. Each student's attribute pattern is influenced by the θ. This indirect connection is the only way that the

response-time would affect the estimation of the attribute mastery. If the correlation between the τ and θ is weak, it limits the influence between the response-time and the general ability, and no additional information would be provided by adding the response-time into the diagnostic model. It also means that the response model and the response-time model are independent of each other. Because the result is not as expected, the relation between the ability variable θ and the time variable τ was inspected. The result shows that the covariance between these two variables is -0.08, which is almost negligible. This result explains why adding the response-time variable did not significantly improve the classification estimation.

Table 6.4 JRT-DINA and DINA Agreement Rate

	Agreement Rate
Attribute Level	0.85
Vector Level	0.44

Table 6.5 Distance *d* for Each Attribute on Both Models

	Distance *d*	
	JRT-DINA	DINA
Attribute 1	0.34	0.41
Attribute 2	0.38	0.41
Attribute 3	0.37	0.35
Attribute 4	0.37	0.45
Attribute 5	0.38	0.48
Attribute 6	0.34	0.39
Attribute 7	0.38	0.43

Table 6.6 shows the average CCR for the two models from the simulation. The simulation shows that both models have high CCR, whereas the DINA has higher CCR. Considering that both model-data fit statistics for JRT-DINA and DINA show adequate fit, it is reasonable that both models present high CCR. Additionally, the ability variable θ and the time variable τ have a negligible covariance, which means the additional time variable did not help in

predicting the classification and even introduced some noise into the model estimation, which is why JRT-DINA has a slightly lower CCR compared to DINA.

Table 6.6 Average CCR for JRT-DINA and DINA

	JRT-DINA	DINA
Average CCR	0.928	0.947

Discussion

The results of this study demonstrated that the JRT-DINA model successfully converged and has an adequate model-data fit. On both items 1 and 2, the JRT-DINA had lower guessing parameters and higher IDI. Thus, compared to the DINA estimation, the JRT-DINA performed better on the estimation of items 1 and 2 based on the item quality. The JRT-DINA also estimated a higher proportion of mastery of attribute 1 compared to the DINA; however, the DINA estimated a higher proportion of mastery for the rest of the attributes. The profile estimation agreement rate for both models is high, and the distance *d* statistics show that the JRT-DINA model did not improve the estimation of the classification. Both models have high CCR, and the DINA has a slightly higher CCR. All the results indicate that the response-time variable did not improve the classification estimation because the ability variable θ and the time variable τ have a negligible covariance. By adding the time variable into the model, no extra information was gained on estimating a person's ability.

Three potential reasons could be found to explain why the covariance of the ability variable θ and the time variable τ was negligible. Firstly, this test is a low-stakes test, and some students were not trying very hard to solve the problem in the given time. The response-time did not reflect their true response-time on each item. Secondly, the test is too easy, and students had more than enough time to finish it, which leads to some lagging behavior. Lastly, the relationship

between the θ and τ might not be linear, and that relation is not captured well by the model. Future studies could investigate the other type of relationship between the θ and τ.

The purpose of the current study is to investigate the potential benefits of incorporating response-time on classification accuracy for a low-stakes, diagnostic computer-based test in middle-school physics. Adding the response-time would ideally result in shorter tests that maintain the same amount of precision for final student classifications. Results of the study showed that adding the response-time into the model did not improve the classification accuracy. While no significant results were found, other studies have proven response-time variables useful in increasing the precision of ability estimates (De Boeck & Jeon, 2019; Ferrando & Lorenzo-Seva, 2007; Zhan et al., 2018). The abovementioned potential reasons for the test being a low-stakes, easy test, and the relationship between the θ and τ is not linear may explain the non-significant results. Other ancillary variables could also be investigated in the future.

CHAPTER VII: FINAL DISCUSSION

The goal of this dissertation was to provide a set of guidelines for researchers and educators interested in DCM and classroom education by illustrating and verifying the process of development, administration, and intervention for an online middle-school physics test using a DCM framework, in addition to showing the potential of adding additional information such as time spent on each item. While this process was demonstrated using a specific physics domain, the intent was to give information in a way that could generalize to other domains where appropriate. Note that, while providing this process, significant contributions were made.

In the first paper, I illustrated the process of implementing DCM into a middle-school physics test. First, the content domain was carefully selected using multiple standards. Multiple subject matter experts (SMEs) then identified and agreed upon the possible combinations of attributes that could be assessed using a single item. A simulation-based approach on the best combination of possible items was then conducted to decide the final Q-matrix, and a physics test on the final Q-matrix was developed, administered, and analyzed by the author and the SMEs.

An existing DCM-scored physics test is demonstrated in Paper 1, while the second paper is a narration of a step-by-step validation process for the effects of the DCM-scored assessment on learning and teaching. In the second paper, I developed a validity argument based on multiple sources of evidence (Standards, AERA, APA, & NCME, 2014). Various qualitative methods were then used to provide evidence to back up the validity arguments. Finally, paper two provided a small follow-up intervention as well as various methods used to report and evaluate the effect of the diagnostic feedback on student learning.

The third paper investigates the potential benefit of incorporating response-time into the diagnostic model as a way to add additional information and potentially improve classification accuracy. In the third paper, I estimated the model fit of the JRT-DINA (Zhan et al., 2018) and the DINA (Junker & Sijtsma, 2001), utilizing the item responses and response-times from the test. The change of individual classification and the posterior probability of mastery for each attribute were then analyzed. Finally, I used a simulation study to compare the predicted correct classification rate (CCR) of the DCM model with the JRT-DINA and DINA models.

The results of the first paper supported the possibility of applying a DCM to a middle-school physics test with careful selection of the content domain and a simulation approach for a Q-matrix construction. It was determined that the results were promising despite some items that showed inadequate fit and quality. In addition, the correlation between attributes showed that most could be distinguished, although the attributes may be similar enough that in follow-up adjustments, they could be combined. Further discussion with SMEs could help refine the items and improve the test quality. Items with poor quality or poor fit could be appropriately examined for potential improvement or elimination before an actual test.

Paper 2 showed that some evidence was found to support multiple validity arguments based on the sources of test content, internal structure, relations to other variables, and the usefulness of the diagnostic feedback. However, regarding the validity arguments based on the source of response processes, validity threats were found because one of the students identified multiple strategies in solving some of the questions, while the other two students identified the same required attributes as the test design. This means that the Q-matrix might not capture some of the students' problem-solving process and create a misfit. However, this misfit has less impact because that student identified an extra attribute for some items rather than a different set of

attributes. Further study could analyze the multiple strategies and refine the Q-matrix or these items. In addition, paper two provided an intervention study for which the students were given study guides based on the attributes they had not mastered (based on the estimated DINA model). It was demonstrated that this additional study material significantly improved some of the attributes. Future research should include more students and consider alternative versions of study guides (the intervention).

The results of the third paper demonstrated a specific approach to adding additional information using the response-time. Although the item parameters for a few items appeared to be better, the rest of the parameters are not. Unfortunately, in this case, the response-time variable did not improve the classification estimation because the ability variable θ and the time variable τ have a negligible relationship. By adding the time variable into the model, no extra information was gained on estimating a person's ability. In the future, different types of relations between ability and the time variable or other ancillary variables should be investigated to improve the classification estimation.

This dissertation demonstrates the process of developing, implementing, and validating a diagnostic assessment in a domain in physics. This process could be used in many other domains across other content areas. It also demonstrated the need for multiple steps for domain selection and test construction. In addition, this dissertation addressed the fact that it is not enough to just retrofit a data set or demonstrate the fit of a diagnostic model. There must also be a follow-up on the diagnostic test in which various pieces of evidence for validity are collected and the usefulness of the instrument explored.

This study showed how steps across the entire process could suggest potential areas that should be addressed. For example, it is critical to consider the construct to be assessed and

constraints prior to determining that a DCM could be used. In this case, it was shown that the diagnostic test marginally fits the response data, while there were a few items that should be improved or addressed. Despite an acceptable fit, when the validity of the test was investigated, there were concerns about the multiple strategies in which some students answered the questions differently than the SMEs' expectations. In total, five of the items showed multiple potential strategies. While items 14 and 20 showed multiple strategies also identified as poor fit (RMSEA > .10) items, the remining three items were not identified as such. In addition, while intuitively, it was thought that the amount of time spent on an item may also be an indication of the attributes mastered, our results did not show a benefit. Thus, no validity evidence was found to support the use of response-time on improving the estimation of classification accuracy. That said, the consideration of the "right" additional information may ultimately improve estimation when using reasonably short tests. In the future, validity evidence on other auxiliary variables could be inspected to support the use of a shorter test with the same level of accuracy.

This dissertation provides a start-to-finish process of development, administration, and validation for an online middle-school physics test using a DCMs framework with response-time. It presents the broken-down steps and detailed description for researchers and educators interested in DCMs and classroom education. Although limitations were found in the dissertation and multiple actions could be taken refine this process in future research, the process could still be generalized into other domains and provide guidelines for researchers and educators interested in DCMs application.

REFERENCES

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015a). *Bayesian Networks in Educational Assessment*. Springer New York. https://doi.org/10.1007/978-1-4939-2125-6

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015b). *Bayesian Networks in Educational Assessment*. Springer New York. https://doi.org/10.1007/978-1-4939-2125-6

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Arican, M., & Sen, S. (2015). Diagnostic Comparison of Turkish and Korean Students' Mathematics Performances on the TIMSS 2011 Assessment. *Unpublished*. https://doi.org/10.13140/rg.2.1.1262.5362

Bolsinova, M., & Maris, G. (2016). A test for conditional independence between response time and accuracy. *British Journal of Mathematical and Statistical Psychology*, *69*(1), 62–79. https://doi.org/10.1111/bmsp.12059

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing Teachers' Understandings of Rational Numbers: Building a Multidimensional Test Within the Diagnostic Classification Framework. *Educational Measurement: Issues and Practice*, *33*(1), 2–14. https://doi.org/10.1111/emip.12020

Bresnock, A. E., Graves, P. E., & White, N. (1989). Multiple-Choice Testing: Question and Response Position. *The Journal of Economic Education*, *20*(3), 239–245. https://doi.org/10.1080/00220485.1989.10844626

Brooks, S. P., & Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

Chen, J., & de la Torre, J. (2013). A General Cognitive Diagnosis Model for Expert-Defined Polytomous Attributes. *Applied Psychological Measurement*, *37*(6), 419–437. https://doi.org/10.1177/0146621613479818

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and Absolute Fit Evaluation in Cognitive Diagnosis Modeling: Relative and Absolute Fit Evaluation in CDM. *Journal of Educational Measurement*, *50*(2), 123–140. https://doi.org/10.1111/j.1745-3984.2012.00185.x

Chen, Y.-H., Ferron, J. M., Thompson, M. S., Gorin, J. S., & Tatsuoka, K. K. (2010). Group comparisons of mathematics performance from a cognitive diagnostic perspective. *Educational Research and Evaluation*, *16*(4), 325–343. https://doi.org/10.1080/13803611.2010.527760

Cohen, J. (1988). *Statistical power analysis for the social sciences*.

Cui, Y., Gierl, M. J., & Chang, H.-H. (2012). Estimating Classification Consistency and Accuracy for Cognitive Diagnostic Assessment: Classification Consistency and Accuracy for CDA. *Journal of Educational Measurement*, *49*(1), 19–38. https://doi.org/10.1111/j.1745-3984.2011.00158.x

Culpepper, S. A. (2015). Bayesian Estimation of the DINA Model With Gibbs Sampling. *Journal of Educational and Behavioral Statistics*, *40*(5), 454–476. https://doi.org/10.3102/1076998615595403

Davier, M. von, & Lee, Y.-S. (2019). *Handbook of diagnostic classification models: Models and model extensions, applications, software packages*. https://doi.org/10.1007/978-3-030-05584-4

De Boeck, P., & Jeon, M. (2019). An Overview of Models for Response Times and Processes in Cognitive Tests. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00102

de la Torre, J. (2009a). DINA Model and Parameter Estimation: A Didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115–130. https://doi.org/10.3102/1076998607309474

de la Torre, J. (2009b). Improving the Quality of Ability Estimates Through Multidimensional Scoring and Incorporation of Ancillary Variables. *Applied Psychological Measurement*, *33*(6), 465–485. https://doi.org/10.1177/0146621608329890

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

de la Torre, J., & Minchen, N. (2014). Cognitively Diagnostic Assessments and the Cognitive Diagnosis Model Framework. *Psicología Educativa*, *20*(2), 89–97. https://doi.org/10.1016/j.pse.2014.11.001

Ekici, E. (2016). “Why Do I Slog Through the Physics?” Understanding High School Students’ Difficulties in Learning Physics. *Journal of Education and Practice*, 13.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380–396. WorldCat.org. https://doi.org/10.1037/1082-989X.3.3.380

Fay, R. H., Dedrick, R. F., Chen, Y.-H., Ferron, J. M., & Van Ingen, S. (2018). *Application of the fusion model for cognitive diagnostic assessment with non-diagnostic algebra-geometry readiness test data* [University of South Florida]. WorldCat.org. https://scholarcommons.usf.edu/etd/7285

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532–538. https://doi.org/10.1037/a0015808

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An Item Response Theory Model for Incorporating Response Time Data in Binary Personality Items. *Applied Psychological Measurement*, *31*(6), 525–543. https://doi.org/10.1177/0146621606295197

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2014). *Bayesian Data Analysis (CRC, Boca Raton, FL)*.

Gierl, M. J., Alves, C., & Majeau, R. T. (2010). Using the Attribute Hierarchy Method to Make Diagnostic Inferences about Examinees’ Knowledge and Skills in Mathematics: An Operational Implementation of Cognitive Diagnostic Assessment. *International Journal of Testing*, *10*(4), 318–341. https://doi.org/10.1080/15305058.2010.509554

Gierl, M. J., & Cui, Y. (2008). Defining Characteristics of Diagnostic Classification Models and the Problem of Retrofitting in Cognitive Diagnostic Assessment. *Measurement: Interdisciplinary Research & Perspective*, *6*(4), 263–268. https://doi.org/10.1080/15366360802497762

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education*, *46*(8), 757–765. https://doi.org/10.1111/j.1365-2923.2012.04289.x

Groß, J., Robitzsch, A., & George, A. C. (2015). Cognitive diagnosis models for baseline testing of educational standards in math. *Journal of Applied Statistics*, *43*(1), 229–243. https://doi.org/10.1080/02664763.2014.1000841

Hart, R., Casserly, M., Uzzell, R., Palacios, M., Corcoran, A., & Spurgeon, L. (2015). *Student Testing in America's Great City Schools: An Inventory and Preliminary Analysis*. Council of the Great City Schools. https://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/87/Testing%20Report.pdf

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. University of Illinois at Urbana-Champaign.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a Family of Cognitive Diagnosis Models Using Log-Linear Models with Latent Variables. *Psychometrika*, *74*(2), 191–210. https://doi.org/10.1007/s11336-008-9089-5

Henson, R., DiBello, L., & Stout, B. (2018). A Generalized Approach to Defining Item Discrimination for DCMs. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 18–29. https://doi.org/10.1080/15366367.2018.1436855

Hodson, D. (1984). The effect of changes in item sequence on student performance in a multiple-choice chemistry test. *Journal of Research in Science Teaching*, *21*(5), 489–495.

Hou, L., la Torre, J. de, & Nandakumar, R. (2014). Differential Item Functioning Assessment in Cognitive Diagnostic Modeling: Application of the Wald Test to Investigate DIF in the DINA Model: Applying Wald Test to Investigate DIF in DINA Model. *Journal of Educational Measurement*, *51*(1), 98–125. https://doi.org/10.1111/jedm.12036

Huff, K., & Goodman, D. P. (2007). *The demand for cognitive diagnostic assessment*.

Izsak, A., Remillard, J., & Templin, J. (Eds.). (n.d.). *Psychometric methods in mathematics education: Opportunities, challenges, and interdisciplinary collaborations*.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to *LanguEdge* assessment. *Language Testing*, *26*(1), 031–073. https://doi.org/10.1177/0265532208097336

Junker, B. W., & Sijtsma, K. (2001). Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory. *Applied Psychological Measurement*, *25*(3), 258–272. https://doi.org/10.1177/01466210122032064

Kabiri, M., Ghazi-Tabatabaei, M., Bazargan, A., Shokoohi-Yekta, M., & Kharrazi, K. (2016). Diagnosing Competency Mastery in Science: An Application of GDM to TIMSS 2011 Data.

Applied Measurement in Education, *30*(1), 27–38. https://doi.org/10.1080/08957347.2016.1258407

Kim, A.-Y. (Alicia). (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, *32*(2), 227–258. https://doi.org/10.1177/0265532214558457

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation*, *35*(2–3), 64–70. https://doi.org/10.1016/j.stueduc.2009.10.003

Lane, S. (1999). *Validity Evidence for Assessments*. 20.

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*(2), 144–177. https://doi.org/10.1080/15305058.2010.534571

Lee, Y.-W., & Sawaki, Y. (2009). Cognitive Diagnosis Approaches to Language Assessment: An Overview. *Language Assessment Quarterly*, *6*(3), 172–189. https://doi.org/10.1080/15434300902985108

Li, H., Hunter, C. V., & Lei, P.-W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, *33*(3), 391–409. https://doi.org/10.1177/0265532215590848

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. https://doi.org/10.1080/10627197.2013.761522

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting Diagnostic Classification Models to Responses From IRT-Based Assessment Forms. *Educational and Psychological Measurement*, *78*(3), 357–383. https://doi.org/10.1177/0013164416685599

Luecht, R. M. (2013). *Assessment Engineering Task Model Maps, Task Models and Templates as a New Way to*. 38.

Mejía-Ramos, J. P., Lew, K., de la Torre, J., & Weber, K. (2017). Developing and validating proof comprehension tests in undergraduate mathematics. *Research in Mathematics Education*, *19*(2), 130–146. https://doi.org/10.1080/14794802.2017.1325776

Mislevy, R. J. (1986). Exploiting Auxiliary Information About Examinees in The Estimation of Item Parameters. *ETS Research Report Series*, *1986*(1), i–49. https://doi.org/10.1002/j.2330-8516.1986.tb00173.x

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A Brief Introduction to Evidence-centered Design*. 37.

Peterson, C. H., Peterson, N. A., & Powell, K. G. (2017). Cognitive Interviewing for Item Development: Validity Evidence Based on Content and Response Processes. *Measurement and Evaluation in Counseling and Development*, *50*(4), 217–223. https://doi.org/10.1080/07481756.2017.1339564

Plummer, M. (2017). *JAGS Version 4.3.0 user manual*. https://people.stat.sc.edu/hansont/stat740/jags_user_manual.pdf

Qualtrics. (2021). *Qualtrics* (Version 09/2021) [Computer software]. Qualtrics. https://www.qualtrics.com

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ravand, H. (2016). Application of a Cognitive Diagnostic Model to a High-Stakes Reading Comprehension Test. *Journal of Psychoeducational Assessment*, *34*(8), 782–799. https://doi.org/10.1177/0734282915623053

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2020). *CDM: Cognitive Diagnosis Modeling*. R package version 7.5-15. https://CRAN.R-project.org/package=CDM

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-Size Indices for Dichotomized Outcomes in Meta-Analysis. *Psychological Methods*, *8*(4), 448–467. https://doi.org/10.1037/1082-989X.8.4.448

Sessoms, J., & Henson, R. A. (2018). Applications of Diagnostic Classification Models: A Literature Review and Critical Commentary. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 1–17. https://doi.org/10.1080/15366367.2018.1435104

Sinharay, S., & Haberman, S. J. (2009). How Much can we Reliably Know About what Examinees Know? *Measurement: Interdisciplinary Research & Perspective*, *7*(1), 46–49. https://doi.org/10.1080/15366360802715486

Spurgeon, S. L. (2017). Evaluating the Unintended Consequences of Assessment Practices: Construct Irrelevance and Construct Underrepresentation. *Measurement and Evaluation in Counseling and Development*, *50*(4), 275–281. https://doi.org/10.1080/07481756.2017.1339563

Su, Y.-S., & Yajima, M. (2021). *R2jags: Using R to Run "JAGS."* 0.7-1. https://cran.r-project.org/web/packages/R2jags/index.html

Templin, J., & Bradshaw, L. (2013a). Measuring the Reliability of Diagnostic Classification Model Examinee Estimates. *Journal of Classification*, *30*(2), 251–275. https://doi.org/10.1007/s00357-013-9129-4

Templin, J., & Bradshaw, L. (2013b). Measuring the Reliability of Diagnostic Classification Model Examinee Estimates. *Journal of Classification*, *30*(2), 251–275. https://doi.org/10.1007/s00357-013-9129-4

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

Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal*, *26*(2), 237–255. https://doi.org/10.1007/s13394-013-0090-7

van der Linden, W. J. (2006). A Lognormal Model for Response Times on Test Items. *Journal of Educational and Behavioral Statistics*, *31*, 24.

von Davier, M. (2005). A General Diagnostic Model Applied to Language Testing Data. *ETS Research Report Series*, *2005*(2), i–35. https://doi.org/10.1002/j.2333-8504.2005.tb01993.x

von Davier, M. (2007). HIERARCHICAL GENERAL DIAGNOSTIC MODELS. *ETS Research Report Series*, *2007*(1), i–19. https://doi.org/10.1002/j.2333-8504.2007.tb02061.x

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. https://doi.org/10.1002/ets2.12043

Wang, C., & Gierl, M. J. (2011). Using the Attribute Hierarchy Method to Make Diagnostic Inferences about Examinees' Cognitive Skills in Critical Reading: Using the AHM in Reading. *Journal of Educational Measurement*, *48*(2), 165–187. https://doi.org/10.1111/j.1745-3984.2011.00142.x

Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2018). Tracking Skill Acquisition With Cognitive Diagnosis Models: A Higher-Order, Hidden Markov Model With Covariates. *Journal of Educational and Behavioral Statistics*, *43*(1), 57–87. https://doi.org/10.3102/1076998617719727

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-Level and Pattern-Level Classification Consistency and Accuracy Indices for Cognitive Diagnostic Assessment. *Journal of Educational Measurement*, *52*(4), 457–476. https://doi.org/10.1111/jedm.12096

Warner, Z. B. (2013). *COMPARING COGNITIVE MODELS OF DOMAIN MASTERY AND TASK PERFORMANCE IN ALGEBRA: VALIDITY EVIDENCE FOR A STATE ASSESSMENT*. University at Albany, State University of New York.

Wells, M., Hestenes, D., & Swackhamer, G. (1995). A modeling method for high school physics instruction. *American Journal of Physics*, *63*(7), 606–619. https://doi.org/10.1119/1.17849

Williams, C., Stanisstreet, M., Spall, K., Boyes, E., & Dickson, D. (2003). Why aren't secondary students interested in physics? *Physics Education*, *38*(4), 324–329. https://doi.org/10.1088/0031-9120/38/4/306

Yang, X., & Embretson, S. E. (2007). Construct Validity and Cognitive Diagnostic Assessment. In J. Leighton & M. Gierl (Eds.), *Cognitive Diagnostic Assessment for Education* (pp. 119–145). Cambridge University Press. https://doi.org/10.1017/CBO9780511611186.005

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 262–286. https://doi.org/10.1111/bmsp.12114

Zhang, susu. (2018). *COGNITIVE DIAGNOSIS MODELING AND APPLICATIONS TO ASSESSING LEARNING*.

朱邦芬. (2016). "减负"误区及我国科学教育面临的挑战. 物理与工程, *3–6*.

APPENDIX A: DETAIL DESCRIPTION OF THE SEVEN ATTRIBUTES

Attribute	**Explanation**
1	$\eta = \frac{W_{output}}{W_{input}}$ η: Mechanical efficiency W_{output} : Output work W_{input} : Input work
2	$W_{output} = G_{load} \times h$ W_{output} : Output work G_{load} : Gravity of the object h : Height change of the object (moving pulley)
3	$W_{input} = F \times S$ W_{input} : Input work F : Tension at the free end of the rope S : Distance the free end of the rope moves
4	$F = \frac{1}{n}(G_{load} + G_{pulley})$ F : Tension at the free end of the rope n : Number of the ropes on the moving pulley G_{load}: Gravity of the object G_{pulley}: Gravity of the moving pulley
5	$S = n \times h$ S : Distance the free end of the rope moves n : Number of the ropes on the moving pulley h : Height change of the object (moving pulley)
6	$W_{input} = W_{output} + W_{loss}$ W_{input} : Input work W_{output} : Output work

	W_{loss} : Loss work
7	$W_{loss} = G_{pulley} \times h$ W_{loss} : Loss work G_{pulley}: Gravity of the moving pulley h : Height change of the object (moving pulley)

APPENDIX B: ALL-POSSIBLE Q-MATRIX

attribute 1	attribute 2	attribute 3	attribute 4	attribute 5	attribute 6	attribute 7
1	0	0	0	0	0	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	0	0	1	0	0	0
0	0	0	0	1	0	0
0	0	0	0	0	1	0
0	0	0	0	0	0	1
1	1	0	0	0	0	0
1	0	1	0	0	0	0
1	0	0	0	0	1	0
1	1	0	1	0	0	0
0	1	0	0	1	0	0
0	1	0	0	0	1	0
0	1	0	0	0	0	1
0	0	1	1	0	0	0
0	0	1	0	1	0	0
0	0	1	0	0	1	0
0	0	0	1	0	0	1
0	0	0	0	1	0	1
0	0	0	0	0	1	1
1	1	1	0	0	0	0
1	1	0	1	0	0	0
1	1	0	0	1	0	0
1	0	1	1	0	0	0
1	0	1	0	1	0	0
1	0	1	0	0	0	1
1	0	0	0	0	1	1
0	1	0	1	1	0	0
0	1	0	1	0	0	1
0	1	0	0	1	1	0
0	1	0	0	0	1	1
0	0	1	1	1	0	0
0	0	1	1	0	1	0
0	0	1	1	0	0	1
0	0	1	0	1	1	0
0	0	1	0	1	0	1
0	0	1	0	0	1	1

APPENDIX C: TRANSLATED PRE-TEST

1. It is known that the useful work of a pulley group is 24J, and the total work done by pulling force is 30J. Find the mechanical efficiency of the pulley group ( )

A.24% B.80% C.60% D.125%

2. Lift an object at a uniform speed for a distance, the total work is 20J, the mechanical efficiency is 80%, the useful work is ( )

A.18J B.16J C.4J D.8J

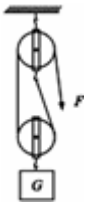
3.Xiaoming uses the pulley group to lift a 5N object to 4 meters high with 4N force. The useful work done by the pulley group is （ ）

A.20J B.30J C.9J D.16J

4. When Xiao Ming experimented with the pulley set as shown in the figure, he lifted an object up at a constant speed of 0.2m, and the useful work was 2.4J. The gravity of the object was （ ）

A.0.12N B.0.04N C.12N D.4N

5. A student uses the pulley set on the right to lift a weight of 6N, and the moving pulley weighs 2N. Without the rope weight and friction, the student 's pulling force is （ ）



A.6N B.4N C.3N D.8N

6. As shown in the figure, a 100N pulling force is used to lift an object at a uniform speed upward for a distance. The gravity of the moving pulley is known to be 60N, excluding rope weight and friction. The gravity of the object is （）

A.240N B.100N C.40N D.300N

7. A student lifts a heavy object with the pulley set on the right. If the object is lifted by 3 meters, the spring dynamometer must move （ ）

A.3m B.6m C.9m D.12m

8. Use a pulley set to lift an object at a uniform speed for a distance, the useful work is 100J, the extra work is 20J, and the total work is （）

A.100J B.20J C.120J D.80J

9.A classmate lifted a heavy object with a pulley set. He did a total of 10J of work, and moved the pulley to do 2J of work to overcome gravity, regardless of rope weight and friction. How much work does the heavy object do to overcome gravity? （ ）

A.10J B.8J C.6J D.2J

10.Someone lifted a weight of 800N with a pulley set. Under the action of pulling force F, the object moved up 12m within 30S. If the gravity of the moving pulley is 200N, the rope weight and friction resistance are not taken into account, the extra work done is ( )

A.6000J B.2400J C.4800J D.9600J

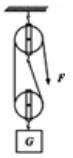
11. Xiaoming uses a pulley group to lift an object weighing 6N to a height of 5 meters. The mechanical efficiency of the pulley group is 60%. Then Xiaoming did a total of ( )

A.30J B.40J C.50J D.60J

12、Someone slowly lifts a heavy object by 10m under the pull of F, the useful work done is 300J, the extra work is 100J excluding the rope weight and frictional resistance, then F is ( )

A.40N B.30N C.20N D.10N

13. Xiaoming raised the weight of 9N to a height of 3m with the pulley set on the right. The moving pulley overcomes gravity to do 6J of work, excluding rope weight and friction. The pulling force is （ ）

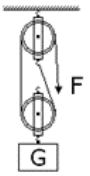


A.10N B.4.5N C.5.5N D.6N

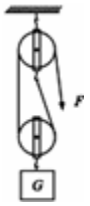
14．Someone slowly lifts a heavy object by 10m with a pulley set, the total work done is 500J, and the gravity of the moving pulley is 10N, if the rope weight and friction resistance are not taken into account, the useful work is ( )

A、100J B、400J C、300J D、500J

15.A student uses a pulley set to lift a 6N weight to 4m, and the pulling force of the rope is 5N. The free end moves 8m. The mechanical efficiency of the pulley set is （ ）

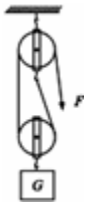


A.50%　B.60%　C.75%　D.80%



16. A student lifted a heavy object to 5m with the pulley set shown on the right. The heavy object did 30J of work to overcome the frictional force, and the pulling force of the rope was 4N. The mechanical efficiency of the pulley set（ ）

A.50%　B.60%　C.75%　D.80%



17. As shown in the figure, someone uses a 4N pulling force to slowly lift a heavy object 5m, the mechanical efficiency is 75%, and the effective work is（）

A．45J B.30J C.20J D.15J

18、Someone slowly lifts a heavy object by 2m with a pulley set, the useful work done is 80J, and the gravity of the moving pulley is 10N, if the rope weight and friction resistance are not taken into account, the mechanical efficiency of the pulley unit is ( )

A.75%　B.87.5%　C.66.7%　D.80%

19. Xiaoming lifted a heavy object to 4m with the pulley set on the right. The heavy object did 24J of work to overcome gravity and moved the pulley to do 12J of work to overcome gravity. If the rope weight and friction resistance are not taken into account, the pulling force is（ ）

A.3N　B.4.5N　C.6N　D.9N

20. Someone slowly lifted a heavy object by 3m with a pulley set, the free end of the rope moved 9m, the useful work done was 270J, and the gravity of the moving pulley was 30N. If the rope weight and friction resistance are not taken into account, the tensile force applied to the free end of the rope is ( )

A.30N　B.40N　C. 90N　D.10N

APPENDIX D: FULL HANDOUT

知识点 1：

η：机械效率　　W 有：有用功　　W 总：总功

例题：用某一滑轮组提升重物，重物克服重力做的功为 6J，拉力做的总功为 10J，则该滑轮组的机械效率是多少？

解：重物克服重力做的功即为有用功，则

知识点 2：

W 有=G 物 h

W 有：有用功　　G 物：物体的重力　　h：物体（动滑轮）上升的高度

例题：若将一重 10N 的物体，用滑轮组提升了 4m，则有用功是多少？

解：W 有=G 物 h =10N×4m=40J

知识点 3：

W 总=Fs

W 总：总功　F：绳子自由端的拉力　s：绳子自由端移动的距离

例题：若用 5N 的拉力拉动滑轮组，绳子自由端移动了 3m，则该滑轮组做的总功是多少？

解：W 总=Fs=5N×3m=15J

知识点 4：

F：绳子自由端的拉力 n：动滑轮上绳子的根数 G 物：物体的重力

G 动：动滑轮的重力

例题：小明用下图所示滑轮组提升一重为 4N 的物体，动滑轮重 2N，忽略绳重和摩擦，则小明所用的拉力是多少？

解：由图可知，动滑轮上有 3 根绳子，n=3

则

知识点 5：

s=nh

s：绳子自由端移动的距离 n：动滑轮上绳子的根数

h：物体（动滑轮）上升的高度

例题：某人用右图所示滑轮组提升某一重物上升了 5m，则绳子拉动了几米？

解：由图可知，动滑轮上有 2 根绳子即 n=2

则 s=nh =2×5m=10m

知识点 6：

W 总=W 有+W 额

W 总：总功 W 有：有用功 W 额：额外功

例题：某同学用滑轮组提升一重物，该重物克服重力做了 10J 的功，动滑轮克服重力做了 5J 的功，不计绳重和摩擦，则该同学做的总功是多少？

解：有用功就是重物克服重力做的功，W 有=10J；在不计绳重和摩擦的情况下，额外功就是动滑轮克服重力做的功，W 额=5J

则　W 总=W 有+W 额=10J+5J=15J

知识点 7：

W 额=G 动 h

W 额：额外功　G 动：动滑轮的重力　h：物体（动滑轮）上升的高度

例题：某同学用滑轮组将一重物提升了 3m，其中动滑轮重 2N，忽略绳重和摩擦，该滑轮组做的额外功是多少？

解：动滑轮提升高度和重物上升高度一样，即 h=3m

则　W 额=G 动 h=2N×3m=6J