MEASURING STUDENTS' KNOWLEDGE MASTERY PATTERNS IN ENERGY USING COGNITIVE DIAGNOSTIC MODELS

by

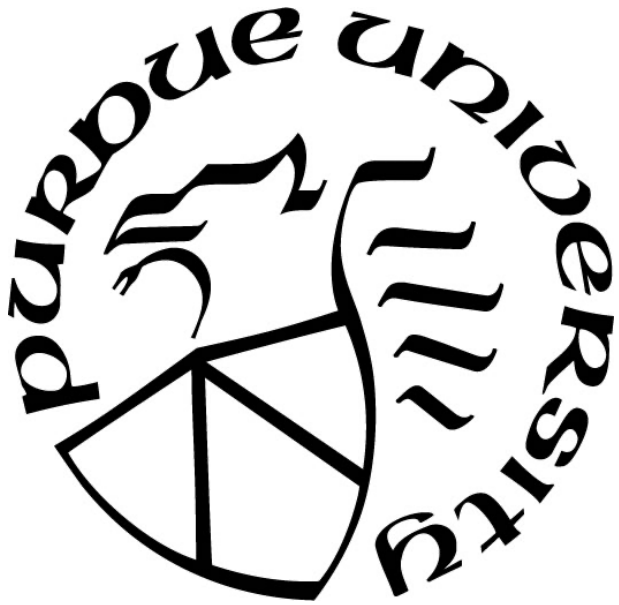
Shuqi Zhou

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy



Department of Educational Studies

West Lafayette, Indiana

August 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Anne Traynor, Chair
Department of Educational Studies

Dr. Hua Hua Chang
Department of Educational Studies

Dr. Kerrie A. Douglas
School of Engineering Education

Dr. Yue Yin
Department of Educational Psychology, the University of Illinois at Chicago

Approved by:
Dr. Janet M. Alsup

To my loving and supportive family

ACKNOWLEDGMENTS

I would like to express my gratitude to Dr. Anne Traynor for her guidance, patience and insight throughout this dissertation. I am also fortunate to have Dr. Anne Traynor as my academic advisor. I greatly appreciate all her kindness, patience, support, and guidance as I progressed through my degree program. I enjoy and benefit a lot from regular meetings and courses with Dr. Traynor, which also contributed greatly to my development as a scholar during my studies.

I would also like to thank my committee members Dr. Hua Hua Chang, Dr. Kerrie Douglas, and Dr. Yue Yin. Their willingness to offer suggestions and ask questions contributed a lot to the finished dissertation. I am particularly grateful for Dr. Chang's encouragement during my study and dissertation writing. I am also appreciative of the five experts who reviewed the Q matrix and their feedback. I am also thankful for Amanda Goodwin Bowman, the graduate coordinator of graduate studies, for her assistance.

I also wish to acknowledge others who contributed to my studies at Purdue. I am grateful to Dr. Ming Ming Chiu for providing me the opportunity to study at Purdue in 2014. I am thankful to Dr. Yukiko Maeda for her guidance and support while I was teaching EDPS 557 Quantitative Methods course. Also, I am thankful to Dr. Wenye Zhou for her encouragement and support during my graduate studies. I am also thankful for the company of my friends during my studies, especially at the difficult pandemic time. I am particularly thankful for Temitope Folasade Adeoye for her encouragement during my dissertation writing. Finally, I am grateful for the support and love of my parents and my sister as I pursued the degree and in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Cognitive diagnostic models can uncover students' mastery of multiple fine-grained skill attributes or problem-solving processes. A number of studies have applied cognitive diagnostic models to detect students' knowledge mastery in mathematics and language testing. However, few studies focus on cognitive diagnostic assessment in K-12 science education, and no studies on the energy topic specifically. This study applied cognitive diagnostic models to Trends in International Mathematics and Science Study (TIMSS) science achievement data to assess students' knowledge mastery in energy. Three TIMSS participating jurisdictions, i.e., Australia, Hong Kong, and Ontario were compared. A Q matrix (i.e., an item attribute alignment table) was proposed based on existing literature about learning progressions of energy in the physical science domain, and the TIMSS assessment framework. The Q matrix was validated through expert review and real data analysis. Then, one of the cognitive diagnostic models, i.e., the deterministic inputs, noisy and-gate (DINA) model was applied to each jurisdiction's data.

Results suggested that the hypothesized learning progression was consistent with Australian and Ontario students' but not Hong Kong students' observed progression in understanding the energy concept. According to overall attribute mastery probabilities and the latent class pattern, most students failed to explain simple electrical systems. Students also performed poorly in recognizing that heating an object can increase its temperature, and that hot objects can heat up cold objects. Identifying sources of energy was found to be easiest to be mastered. I discuss several potential curriculum-related issues that may affect students' mastery patterns in different jurisdictions.

CHAPTER 1. INTRODUCTION

1.1 Purpose of Study and Research Questions

Students' domain-specific concept knowledge has received substantial attention from researchers in science education (Liu & McKeough, 2005). Previous research shows that many students have not mastered an understanding of energy as envisioned in policy documents (Neumann et al., 2013). However, understanding energy is important, since energy concepts are scientifically and academically related to many social, environmental and technological applications (Chen et al., 2014). Although there are extensive studies probing students' understanding of energy (e.g., Duit, 2014; Lacy et al., 2014; Lee & Liu, 2009; Liu & McKeough, 2005), most studies use interviews (e.g., Lacy et al., 2014; Jin & Wei, 2014) and item response theory (IRT) based Rasch analysis (Neumann et al., 2013; Liu & McKeough, 2005). However, there is no strong rationale for using Rasch analysis in these studies, since it has a restrictive model assuming all items are equally discriminating indicators of students' energy understanding, although item discrimination varies in practice. The interview studies are limited by the small sample sizes of participants and their generalizability.

Students' incorrect responses during problem solving can be caused by weaknesses in multiple, distinct underlying skill attributes (e.g., Brown & Burton, 1978; Brown & VanLehn, 1980; Tatsuoka, 1983). Cognitive diagnostic models (CDMs) can uncover students' mastery of multiple fine-grained skill attributes or problem-solving processes. CDMs can diagnose students' performance on a set of multiple discrete skills and provide formative diagnostic information to inform instruction and learning based on students' mastery or non-mastery of these fine-grained skills (Embretson, 1998; Leighton & Gierl, 2007; Nichols, 1994). CDMs can help to diagnose students' mastery of specific energy concepts, which could be useful to validate learning

progressions (an ordered description of students' understanding about a particular concept) that have been proposed in the literature. The aims of the proposed study are a) to measure systematic patterns of students' knowledge mastery and misunderstandings of energy and b) to gain a better understanding of students' learning progression through energy concepts. The study will use CDMs to identify students' knowledge mastery and misunderstanding patterns through the hypothesized learning progressions. CDMs can characterize students' cognitive mastery pattern at a fine-grain size (Rupp et al., 2010), and provide diagnostic feedback about students' mastery or non-mastery of each skill. In addition, since students' opportunity to learn is an essential factor contributing to their learning outcomes (Törnroos, 2005), the study will also examine how the intended curriculum may influence students' understanding of energy across different countries.

Based on previous research (Lacy et al., 2014; Neumann et al., 2013), this study hypothesized that students understand energy through four hierarchical concepts: 1) forms of energy; 2) transfer and transformations of energy; 3) dissipation and degradation of energy; and 4) conservation of energy. The study will use data from a fourth-grade physical science assessment to address the following major questions:

1. To what extent does the hypothesized learning progression match students' observed progression in understanding the energy concept, based on the results of the cognitive diagnostic model?
2. What similarities and differences in students' knowledge mastery patterns are evident for different countries?
3. How does the intended curriculum relate to students' understanding of energy across different countries?

1.2 Significance of the Study

Cognitive diagnostic assessment can illuminate students' knowledge mastery pattern at a fine-grain size level because the latent variables (i.e., attributes) in cognitive diagnostic assessment tend to be more narrowly defined than the constructs in multidimensional factor analysis, another common method of analyzing assessment item responses (Rupp et al., 2010). Many studies have applied cognitive diagnostic assessment in the mathematics (e.g., Lee et al., 2011; Birenbaum et al., 2004) and language testing (e.g., Jang, 2009) fields. However, relatively few studies (Briggs & Alonzo, 2012; Chen et al., 2017; Fumler et al., 2014; Kabiri et al., 2017; Kizil, 2015) focus on cognitive diagnostic assessment in K-12 science education, and no studies have examined the energy topic specifically. This study will use CDM to explore students' mastery patterns for energy concepts. The study will use the literature and experts' judgments to hypothesize a sequence of energy-related concepts likely to be measured by the Trends in International Mathematics and Science Study (TIMSS) achievement test items. Then, I will use CDMs and three countries' TIMSS data to test the hypothesized learning progression, which could allow us to have a better understanding of the sequence in which students tend to master energy concepts and provide more accurate and informative diagnostic assessment results to students and teachers. The study results will have implications for the methodology of validating hypothesized learning progressions through CDM by checking attribute mastery probability, indicating whether CDM will be a feasible method to detect learning progressions. In addition, the study will also provide information about how the intended curriculum affects students' understanding across different countries, which could inform curriculum changes.

CHAPTER 2. REVIEW OF LITERATURE

Before detailing the methods that this study will use, I review the literature related to the current study in this chapter. To provide background for Research Question 1 and Research Question 2, I start by reviewing the definitions of learning progressions. I summarize the methods of developing learning progressions in science disciplines. Then, I introduce specifically the methods that are most widely used. I review students' understanding about energy and learning progressions, particularly those related to the energy topic. Since I will explore how intended curriculum and instruction relate to students' understanding of energy across countries (Research Question 3), I also review the concepts of opportunity to learn and instructional sensitivity. I introduce widely used instructional sensitivity indices. I briefly review the science curricula of Australia, Hong Kong, and the Ontario province of Canada, since this study will investigate these three jurisdictions' Grade 4 science test item performance, as further detailed in Chapter 3. Finally, I introduce cognitive diagnostic models.

2.1 Learning Progressions

Learning progressions (LPs) are descriptions of increasingly sophisticated levels of thinking about or understanding of a topic (National Research Council, 2007). LPs are ordered descriptions of students' understanding of a given concept (Alonzo & Steedle, 2009). LPs describe an upper and lower "anchoring" performance-level description, followed by descriptions of several intermediate levels (Stevens et al., 2010). The lower anchor is defined by students' tentative understanding of a particular idea or concept upon entering the learning progression (Neumann et al., 2013). The level of understanding expected from students once they have mastered the concept or skill defines the upper anchor (Neumann et al., 2013). The upper anchor is often defined by

analysis of policy documents, such as curriculum standards (Chen et al., 2017). LPs also describe different levels of understanding that students have as they move towards the upper anchor (Stevens et al., 2010).

Learning progressions (LPs) are "a promising means of organizing and aligning the science content, instruction and assessment strategies to provide students with the opportunity to develop a deep and integrated understanding of a relatively small set of big ideas of science over an extended period of time" (Stevens et al., 2010, p. 688). LPs may provide a framework that can be used to coordinate standards, assessments, and instruction in a way that advances scientific literacy (Alonzo & Gotwals, 2012). In this way, the development of LPs should not only include increasingly sophisticated levels of thinking about or understanding a topic. LPs also need to include relevant assessment criteria about students' understanding at each level and correspondent instruction to enhance students' learning to more sophisticated levels (Stevens et al., 2010). Intermediate learning progression levels are the levels between the upper and lower anchor. They are informed by two sources of research: (1) "research on how students develop conceptual understanding through an increasingly complex knowledge base" (Neumann et al., 2013, p.168), and (2) "research on how students' understanding of the target concept changes over time" (Neumann et al., 2013, p.168).

2.2 Methods of Developing Learning Progressions in Science Disciplines

2.2.1 Overview

Although there is not a specific widely agreed-upon method for developing LPs (Stevens et al., 2010), the development of LPs is an iterative process of empirical validation and theoretical enhancement (Neumann et al., 2013). LPs use both top-down and bottom-up design approaches

(e.g., Alonzo & Gotwals, 2012): 1) Bottom-up LPs refers to LPs "where the identification of topics and learning pathways are grounded in iterative assessments that obtain evidence of student learning and build on it" (Duschl et al., 2011, p. 125), and 2) "top-down LPs where the selection of topics and pathways is based on a logical task analysis of content domains and personal experiences with teaching" (Duschl et al., 2011, p. 125).

Duschl et al. (2011) summarized and distinguished two types of LPs in the current literature, "validation" and "evolutionary" LPs. Validation LPs aim to validate initial sequences and levels of progression that have been proposed (Duschl et al., 2011). Evolutionary LPs are LPs that refine and define the developmental pathways through identification of intermediate levels that are then used to help instructional interventions (Duschl et al., 2011). Detailed differences between these two types of LPs are presented in Table 1.

Table 1. Validation LPs and Evolutionary LPs

Validation LPs	Evolutionary LPs
(1) LP based on validating a standards-based progression: instruction as intervention	(1) LP based on sequencing of teaching experiments across multi-grades: instruction as refining progression
(2) Theory-driven top/down approach	(2) Evidence-driven bottom/up approach
(3) Upper anchors as college readiness	(3) Upper anchors as targeted literacy
(4) Uses assessments to confirm learning models	(4) Uses assessments to explore learning models
(5) Progress variables steps and targets are fixed	(5) Progress variable steps and targets are flexible
(6) Adopts a misconception-based 'Fix It' view of conceptual change instruction	(6) Adopts an intuition-based 'Work with It' view of conceptual change instruction
(7) Theory building as conceptual change	(7) Model building as conceptual change
(8) Domain general orientation to topic selection	(8) Domain specific orientation to topic selection

Note. Reprinted from "Learning progressions and teaching sequences: A review and analysis", by Duschl et al., 2011, *Studies in Science Education*, *47*(2), p. 173

Grain size is another important issue discussed in the development of LP research (Hokayem & Gotwals, 2016). The grain-size of an LP refers to the extent to which the progression is broadly or finely focused (West et al., 2012). The covered breadth of content and length of time need to be defined in LPs. Alonzo (2012) distinguished coarse-grained and fine-grained LPs. Studies related to LPs vary in the grain size and time length. LPs studies cover from elementary, through middle and high school (e.g., Mohan et al., 2009; Smith et al., 2006). Some studies only concentrate on one grade length (e.g., Johnson & Tymss, 2011; Neumann et al., 2013). The breadth of content in LPs also varies in the existing research. Broader LPs could refine standards and large-scale standards and assessments, while narrower LPs for specific content topics may serve to support the curriculum, instruction and formative assessment in the classroom (Alonzo, 2012; Furtak, 2012; Lehrer & Schauble, 2015).

There are different ways of developing and validating LPs in science education: interview (e.g., Jin & Wei, 2014; Jin & Anderson, 2012; Suzuki et al., 2015), construct map (e.g., Black et al., 2008; Plummer & Maynard, 2014; Wilson, 2009), Rasch-type partial credit model from item response theory (e.g., Lee & Liu, 2009; Neumann et al., 2013; Plummer, & Maynard, 2014), cognitive diagnostic model (Gao et al., 2018; Kizil, 2015; Briggs & Alonzo, 2012), latent class analysis (Steedle & Shavelson, 2009), and Bayes' network (Rupp et al., 2009). Researchers usually combine two of these methods to develop LPs, so that there is a second source of evidence for cross-validation. In the current literature, the most commonly used methods are interviews and Rasch analysis. Appendix A provides a summary of empirical papers about LPs in science (except energy topics), and Appendix B summarizes papers about LPs for energy topics. I will introduce interview, partial credit model, and construct map methods specifically in the following sections, and cognitive diagnostic models in a later section.

2.2.2 Interview

Interviews have been used to develop learning progressions in numerous studies (Alonzo & Steedle, 2007; Draney, 2009; Hokayem & Gotwals, 2016; Jin & Anderson, 2012; Jin et al., 2013; Lacy et al., 2014; Paik et al., 2017; Plummer & Krajck, 2010; Shin et al., 2009; Stevens et al., 2010). An interview is a good way to truly understand students' reasoning process about a particular topic when they solve the questions or tasks to ensure the substantive validity of an assessment (Paik et al., 2017; Jin et al., 2013). However, interview results may be restricted by the small interview sample size and cannot be used for statistical generalizations (Jin & Anderson, 2012). There are mainly two kinds of interviews that are applied in LP development and validation: think-aloud interviews and traditional clinical interviews (e.g., Stevens et al., 2009; Alonzo & Steedle, 2007). Generally, students are asked to think aloud while they answer the items. After a student completes the test, the interviewer also may ask the student to talk about each item to understand students' responses to the items (Alonzo & Steedle, 2007). In addition, interviews have been implemented before and after instructional intervention in some studies (Lacy et al., 2014; Jin et al., 2013; Plummer & Krajcik, 2010) to track students' understanding and progression.

2.2.3 Partial Credit Model

In item response theory, an item discrimination parameter indicates the strength of the relationship between the item and latent trait score. The partial credit model (PCM) is a Rasch-type item response model that constrains the item discrimination parameter values to be equal across all items when responses are in two or more ordered categories or levels (Masters, 1982). PCM is based on a unidimensional probabilistic model assuming a student's probability on each item is merely decided by "the difference between the student's latent trait status (i.e., academic

ability) and the difficulty of the task involved" (Liu & McKeough, 2005, p. 501). The equation of the PCM is *Equation 1* as follows (Masters & Wright, 1997):

$$\frac{P_{ijx}}{P_{ijx-1}+P_{ijx}} = \frac{exp\,(\theta_i - \delta_{jx})}{1+ exp\,(\theta_i - \delta_{jx})} \qquad (1)$$

where P_{ijx} is the probability of person i scoring x on item j, P_{ijx-1} is the probability of person j scoring x-1, and δ_{ix} is an item parameter governing the probability of scoring x rather than x-1 (Masters & Wright, 1997, p. 102)

2.2.4 Construct Map Approach

A construct is an unobservable human trait or abstract personal attribute (e.g., motivation, ability, opinion, agreeableness) that is given meaning by a specific theoretical framework (Peak, 1953). A construct map defines a particular construct and different levels of student performance on it. It was the first building block of the Berkeley Evaluation and Assessment Research Center (BEAR) assessment system (BAS; Wilson, 2005; Wilson & Sloane, 2000), which has been used to evaluate students, schools, and educational policy in some parts of the US.

A construct map is used to represent a cognitive theory of learning from a development perspective (Draney, 2009). It is developed only after progress variables are determined and defined (Masters et al., 1990; Wilson, 1990). A construct map consists of different progress variables. The levels of the progress variables are linked to the construct map levels (Wilson, 2009). Progress variables represent a range of student thinking about a particular knowledge domain or construct, and they describe the construct or core idea researchers want to track (Merritt & Krajcik, 2013) through the learning activities associated with a curriculum (Wilson & Sloane, 2000). Progress variables often come from studies examining representative domains of core science

topics (Draney, 2009). The main purpose of developing progress variables is to serve as a framework for assessment and diagnosis (Wilson, 2008).

Construct map is a way to structure both measurement and diagnosis and to make sure that the two are aligned (Wilson, 2008). Construct maps also reflect the learning goals and instructional sequencing of the curriculum (Kennedy et al., 2005). The importance of embedded assessments tied to the learning goals of a curriculum is also highlighted through construct maps by assessing what students know and can do at several levels (Kennedy et al., 2005). When a learning progression only has one construct, the learning progression is identical to a construct map (e.g., Plummer & Maynard, 2014). A set of construct maps constitute a learning progression (Draney, 2009).

I introduced different kinds of learning progressions in science education in this section. According to different classifications, there are top-down and bottom-up design approaches of LPs, Validations LPs and Evolutionary LPs, and coarse- and fine-grained LPs. I also summarized different ways of developing and validating LPs in science education. Three widely used methods, i.e., interview, partial credit model, and construct map methods, were specifically introduced. CDM is another method of developing LPs, which I will specifically describe in Section 2.7. In section 2.3, first, I will introduce LPs related to the energy topic, beginning from the conceptual framework of energy and existing research about students' understanding of energy.

2.3 Students' Conceptual Understanding about Energy

2.3.1 The Conceptual Framework of Energy

It is widely agreed that the concept of energy is a central idea in science education. Energy is a core idea in science education since it is the basis to foster students' ability to learn about a

variety of scientific topics with coherence and increasing depth (National Research Council; NRC, 2012; cited in Opitz et al., 2015). Energy is a crosscutting concept connecting all science disciplines and we experience it in our everyday life situations (Saglam-Arslan & Kurnaz, 2009). Energy is also a core idea proposed in the US Next Generation Science Standards (NGSS; the NGSS Lead States, 2013).

Duit (1984) proposed five basic aspects of the energy concept as a potential framework for energy teaching: conceptions of energy, energy transfer, energy conversion, energy conservation, and energy degradation. Duit (1984) explained each basic aspect specifically: energy transfer refers to the energy that can be transferred from one system or place to another; energy conversion refers to the energy that can be converted from one form to another; energy conservation recognizes the amount of energy does not change while it is transferred or converted; and energy degradation refers to the "value" of energy that is lost from transferring from one form to another, although the total amount of energy does not change.

2.3.2 Students' Understanding about Energy

Since each student has some prior knowledge of energy concepts from their life experience, they have different understandings about energy, which may include some erroneous ideas. Students' prior knowledge may affect their success in learning about energy concepts (Trumper & Gorsky, 1993). Previous literature shows that energy is often defined as "the ability to do work" by students and students' understanding about energy tends to be superficial (Boylan, 2017). Many researchers (e.g., Watts, 1983; Gilbert & Pope, 1986; Kirkwood & Carr, 1988; Trumper & Gorsky, 1993; Duit, 2014) have investigated students' conceptual understanding about energy, using several distinct conceptual frameworks for talking about energy that can be classified as follows:

1. Energy is associated with human beings (anthropocentric framework).
2. Things possess and expend energy (depository framework).
3. Energy causes things to happen (cause framework).
4. Energy is a dormant ingredient in things and can be released by a trigger (ingredient framework).
5. Energy is associated with activity (activity framework).
6. Energy is a product of certain processes (product framework).
7. Energy is a general kind of fuel associated with making life comfortable (functional framework).
8. Energy is a kind of fluid which is transferred in some processes (flow-transfer framework).
9. A scientific conception in which energy is transferred from one system to another. (Trumper & Gorsky, 1993, p. 639)

Other researchers have also explored students' understanding of energy in detail. Chabalengula et al. (2012) summarized five main kinds of erroneous ideas that students hold about energy and energy-related concepts: energy is force; energy is work; energy is electricity; energy is power; and energy is an entity. In Chabalengula et al. (2012)'s study, about half of the students (44%) gave a correct definition of energy (i.e., energy is the ability to do work), but a large percentage of these students (24%) did not write any additional statements even though students were required to write. Their results showed that many students had problems in understanding energy and energy-related concepts. Trumper (1990) also summarized most students as holding the following alternative frameworks of energy before they studied physics: anthropocentric frameworks, cause frameworks, and product frameworks. After students studied physics, they typically still retained the same alternative frameworks. Thus, Trumper's (1990) study used a model viewing children's minds as a rich and varied network of ideas from day-to-day experiences, and non-scientific language to help students to change their conceptions about energy. Trumper

(1991) also tried pupil/teacher dialogue in small groups to help students change their misconceptions.

Researchers also use learning progressions to examine students' understanding about energy as an educational trajectory and I will introduce the learning progression of energy in the following section.

2.3.3 Learning Progression of Energy

In the past two decades, as a core science concept, energy has received a lot of attention in the research on LPs across different grades or grade bands (e.g., Lee & Liu, 2010; Liu & McKeough, 2005; Neumann et al., 2013; Yao et al., 2017). These studies aim to develop corresponding assessments, examine students' progression in understanding energy, and improve instruction and curriculum related to energy topics. A summary of recent studies on LPs related to energy is presented in Appendix B. Similar to approaches of developing learning progressions in other concepts in science, the development of LPs on energy mainly has used interviews (Lacy et al., 2014; Dawson-Tunik, 2006) and Rasch type partial credit models (Herrmann-Abell & DeBoer, 2011; Lee & Liu, 2010; Neumann et al., 2013; Yao et al., 2017). The studied grades have ranged from third grade to twelfth grade. These studies include not only small samples but also large-scale samples, such as participants in the TIMSS (Liu & McKeough, 2005; Lee & Liu, 2009). Though studies may use different terms to refer to the same concepts, most of these studies propose LPs of energy from four strands: energy sources and forms, transfer and transformation, degradation, and conservation. In summary, the current studies on LPs of energy cover different grade levels, sample sizes, and development methods. I will introduce the main studies on LP of energy specifically next.

Studies have explored students' LPs for energy from different perspectives. Dawson-Tunik (2006) explored students' progression in understanding energy based on Fischer's (1980) skill theory across three levels (i.e., representational systems level, single abstractions level, abstract mappings level) using both interviews and Rasch analysis. Their results concluded that many students did not have a sufficient understanding about the energy concept. Liu and McKeough's (2005) study hypothesized five levels of an energy concept sequence (i.e., activity/work, source/form, transfer, degradation, conservation). Correspondingly, they analyzed three different populations from the third TIMSS database using Rasch partial credit models: students aged 9 years at the time of testing, typically grades 3 and 4; students of age 13 typically 7 and 8; students at the final year of their secondary education, grade 12. The results showed that their hypothesized sequence of energy concept development was supported. Their study also showed that third- and fourth-grade students can develop an understanding of the first two levels, i.e., energy does work, and sources or forms of energy. They also concluded that energy degradation should be an important component for understanding energy conservation (Liu & McKeough, 2005). Herrmann-Abell and DeBoer (2011) examined grade six to college students' understanding about energy transformation, energy transfer and conservation of energy using Rasch analysis. Their study supported that knowledge of forms of energy was important for students to successfully answer questions about energy transformation. They found the idea of conservation of energy was much more difficult than the ideas of energy transformation and energy transfer to students. They concluded that it is easier for students to know general principles than to apply them in real life. Herrmann-Abell and DeBoer (2011) also found that there are some misconceptions about energy that are widespread at all grade levels. For example, there is a misconception held by students at all grade levels that both force and energy are transferred during mechanical interactions.

Lee and Liu (2010) explored students' progression in understanding energy from a knowledge integration approach applying the Rasch partial credit model. Their analysis showed that items about advanced energy concepts such as conservation are related to the highest knowledge integration levels, followed by transformation and source items at lower knowledge integration levels. The difficulty is partially associated with the increased demand for integrating many scientifically relevant ideas. Furthermore, students' knowledge integration level differs by grade and subject. Eighth-grade students' mean energy knowledge integration level is significantly higher than that of sixth- or seventh-grade students, and the mean knowledge integration level at the end of school year of students who took a physical science course is significantly higher than that of students who took a life or earth science course after a school year. Lee and Liu's (2010) study suggests that to help students develop an understanding of energy, science curricula should address the relevant instructional sequence of energy concepts as well as encourage students to integrate ideas.

Neumann et al. (2013) explored four hierarchical energy topics: forms, transfer, degradation, and conservation, each of which was conceptualized as having four hierarchical levels of complexity: facts, mappings, relations, and concept. They confirmed a general progression of the four levels for energy conceptions (forms and sources, transfer and transformation, dissipation, conservation). But they did not confirm the distinct levels of these conceptions. Their Rasch analysis and analysis of variance (ANOVA) suggest that students may develop an understanding of energy transfer and transformation in parallel with an understanding of energy degradation.

Following Neumann et al.'s (2013) approach, Yao et al.'s (2017) study examined eighth- to twelfth-grade students' developing understanding of energy in mainland China to collect evidence for national standard revision and build a foundation for future instructional research.

Their study took both ideas about energy and levels of conceptual development into account. Ideas about energy was their first progress variable, and levels of conceptual development was their second progress variable. There were four key ideas of energy: form, transfer and transform, dissipation, and conservation; and four conceptual development levels: fact, mapping, relation, and systematic. Although their study followed the same sequence of four ideas about energy as previous studies (i.e., forms, transfer and transformation, dissipation, and conservation) (Neumann et al., 2013), their Rasch analysis results did not support the hypothesis that students actually progress along this sequence in their understanding of energy. Their findings showed that although "energy forms" is a foundational idea for developing a deeper understanding of energy, other ideas may not necessarily be developed in a distinct sequence (Yao et al., 2017).

In order to allow students to accomplish understanding by the end of the elementary grades, Lacy et al. (2014) proposed a detailed learning progression for energy from four strands, focusing on grades 3-5: forms of energy, transfer and transformations, dissipation and degradation, and conservation. Their proposed learning progression was established on the "aligned development of a network of interconnected and interdependent foundational ideas" (p. 265). Their proposed progression was also based on students' intuitive ideas. The progression also takes students' misinterpretations and hurdles in previous research into account. Their exploratory interviews and teaching interventions have supported that relevant instruction could increasingly enhance, transform, and integrate students' knowledge toward a scientific understanding of energy (Lacy et al., 2014).

The studies on LPs of energy also provide some suggestions about instruction for the energy topic. For instance, McKeough (2005) argued for a multi-faceted and holistic approach to introducing the energy concept. This holistic and multi-faceted approach means that teachers

should expose students to as many aspects of the energy concept as developmentally appropriate and continue to incorporate additional aspects through grade 12 and beyond. At all grades, instruction of energy should focus not only on developing students' understanding of the energy concept itself but also on the application of their understanding in various contexts. Nordine et al. (2010) explored the effectiveness of an approach to middle school energy instruction that was consistent with the principles of learning-goals-driven design (Krajcik et al., 2008) and curricular coherence (Roseman et al., 2008; Shwartz et al., 2008) to address energy concepts. For students who were taught using this approach, they gained a more integrated understanding of energy than students who were not taught using this approach.

This section summarized current studies on LPs of energy. Most of these studies propose LPs of energy from four separate strands: energy sources and forms, transfer and transformation, degradation, and conservation. Some studies only used partial credit model (PCM)s' results and did not compare other models, while PCM constrains the item discrimination parameter values to be equal across all items and some original information of the item is lost. In addition, the results of energy LPs appear different across studies: although some studies (Liu & McKeough, 2005; Neumann et al., 2013) confirmed the four proposed hierarchical levels, i.e., 1) energy sources and forms, 2) transfer and transformation, 3) degradation, and 4) conservation, one study (Yao et al., 2017) did not confirm four sequential levels. Since studies use different grades' student samples, this may lead to differences in results. For the current study, since I will focus on fourth grade students, I will hypothesize the learning progression following Lacy et al. (2014)'s four-strand LP: forms of energy, transfer and transformations, dissipation and degradation, and conservation.

2.4 Opportunity to Learn

Previous studies (e.g., Lee & Liu, 2010; Plummer & Krajcik, 2010; Yao et al., 2017) on learning progressions found that educational environments play important roles in students' learning, seeming to affect their learning trajectories or rates of progress. Curriculum and instruction are both important factors contributing to an educational environment. Frankenberg et al. (2016) also emphasize that schools should ensure high-quality instruction for all students. These findings implicate curriculum and instruction as playing an important role in students' learning. Whether the curriculum and instruction provide students an opportunity to learn related topics would influence students' understanding and their progress through stages of a learning progression. Having an opportunity to learn is essential for learning, but a learning opportunity cannot guarantee students truly learn.

Opportunity to learn (OTL) refers to "whether or not the students have had the opportunity to study a particular topic or learn how to solve a particular type of problem" (Husen, 1967a, p. 162; cited in Törnroos, 1993). The concept of OTL was first introduced in the early 1960s to ensure the validity of cross-national comparisons in studies of mathematics achievement (McDonnell, 1995). In order to interpret differences in achievement within or across countries, topics included in a country's implemented curriculum at a particular grade level for a particular population, and excluded or given minimal attention, must be considered (Törnroos, 2005). OTL variables are often measured by large-scale cross-national assessments such as TIMSS, and the Teacher Education and Development Study in Mathematics (TEDS-Math). OTL is associated with the study of educational equity and fairness related to the adequacy of educational experiences, which include the availability of resources across classrooms, teacher quality differences, and other aspects of schooling related to learning (D'agostino et al., 2007).

OTL variables are multidimensional and related studies investigate different dimensions of students' educational experiences. Among those dimensions, content coverage, content exposure, and content emphasis are the most measured aspects (Stevens, 1993; Wang, 1998). Content coverage is the most frequently used OTL variable and even is the only indicator of OTL in some studies (Wang, 1998; Schmidt et al., 2011). Content coverage refers to whether the topics tested are covered in the instruction or not. Content exposure refers to allowing and devoting time to instruction and the depth of the teaching provided (Wang, 1998). Content emphasis refers to whether a certain area was treated as a major topic, a minor topic, a review topic, or not taught at all (Wang, 1998).

Researchers have explored the relation between OTL and students' achievement using large scale assessments. There is a substantial correlation between learning achievement and OTL when achievement has been examined across countries (Törnroos, 2005). Mo et al. (2013) found that OTL is an important factor in students' science achievement in TIMSS 2002 using hierarchical linear modeling methodology. Students' science achievement was higher in the class whose teachers had a full science teaching license or certificate (Mo et al., 2013). Topic coverage is also related to science achievement (Mo et al., 2013). Specifically, there is an interaction effect of topic coverage and students' emotional engagement. There is a positive relation between topic coverage and science achievement among students who are not interested in science, while the relation is slightly negative among students who are interested in science. When OTL was measured using an item-based approach, i.e., students' OTL for each test item, it has higher correlations between OTL and students' achievement than measured by aggregated values that cover broader mathematical topics (Törnroos, 2005). Törnroos (2005) found that the association between OTL

and student achievement must be established on learning opportunity data covering a longer period than only the most recent year (Törnroos, 2005).

To sum up, OTL is essential to students' performance. Different dimensions of OTL can be measured. Content coverage, content exposure, and content emphasis are perhaps the most prevalent dimensions of OTL measured. These dimensions could help us to understand students' learning progression in energy through comparing knowledge mastery patterns for students who have and have not had the opportunity to learn.

2.5 Instructional Sensitivity

Another concept related to OTL is instructional sensitivity, called "instructional validity" in some studies. Haladyna and Roid's (1981) definition of instructional sensitivity is "the tendency for an item to vary in difficulty as a function of instruction" (p. 40). Popham (2006) defines instructional sensitivity as "the degree to which students' performances on a test accurately reflect the quality of instruction specifically provided to promote students' mastery of what is being assessed" (p. 1). D'Agostino et al. (2007) related instructional validity to OTL and referred to instructional validity as "the ability of a test to detect instructional differences that might arise due to OTL" (p. 4). In summary, these definitions all refer to the degree that a test reflects student ability as the result of the instruction. In this study, I will use the term instructional sensitivity to maintain consistency. Through instructional sensitivity, we can see how the instructional opportunity can influence students' learning progression and attribute mastery in the energy domain.

Instructional sensitivity could be observed through instruction-focused methods and expert judgment (Polikoff, 2010). There are also different indices measuring instructional sensitivity in psychometrics, such as pre-to-post difference index (PPDI) (Cox & Vargas, 1966), percent of

possible gain (PPG) (Brennan & Stolurow, 1971), the Brennan index (Brennan, 1972), and ZDIFF (Haladyna & Roid, 1981). Researchers also use contingency table indices, Bayesian methods (Helmstadter, 1974), Hedge's g, and Cohen's d to measure instructional sensitivity. Hedges' g (Hedges, 1981) is also widely used as an effect size index for approximately continuous data. It is relatively simple to compute and interpret. Similarly, Cohen's d is also an effect size indicating the standardized difference between two means. The difference between Hedge's and Cohen's d is that Cohen's d uses standard deviation that is divided by N, while Hedge's g is divided by N-1.

Differential item function (DIF) methodology is also applied to measure instructional sensitivity. DIF is a statistical characteristic of an item representing whether different subgroups perform differently to a particular item. In measuring instructional sensitivity, the subgroups could be divided into the group that has received instruction and the group that has not received instruction. DIF has two primary types: uniform and nonuniform. For uniform DIF, the magnitude and direction of the item difficulty difference between the groups are constant across the entire range of observed scores (Hanson, 1998). For nonuniform DIF, item difficulty favors one group across part of the score range and another group across other parts of the score range. The logistic regression method has many strengths in detecting DIF: it can accommodate continuous conditioning variables (Li et al., 2017), can model uniform and nonuniform DIF simultaneously (Swaminathan, 1994); logistic regression for ordinal items is flexible in model specification and it is especially efficient for simultaneous conditioning on multiple variables (Li et al., 2017, p. 3). When the data has nested structure, hierarchical logistic regression has advantages over logistic regression since it accounts for the nested structure in the data set. It performs better than simple logistic regression when the data has a nested structure (French & Finch, 2010). Hierarchical logistic regression is also used to detect instructional sensitivity (Li et al., 2017).

2.6 Science Curriculum of Primary Schools across Three Jurisdictions

Different countries and regions have different science curricula. The detailed expectations specified in the curriculum may also vary by country or region. I will explore how the intended curriculum relates to students' understanding of energy in Research Question 3 of this study. In this section, I will specifically introduce the science curriculum of the three jurisdictions (i.e., Australia, Hong Kong, Ontario) that will be included in this investigation. Australia, Hong Kong, and Ontario are chosen since their curricula have changed or been updated before 2011, and these jurisdictions participated in the TIMSS assessment, which included items measuring understanding of energy, and item-level curriculum coverage information.

2.6.1 Science Curriculum of Australia

The current Australian science curriculum was initially released in 2010. Some states started implementing the current science curriculum in 2011 and full implementation across Australia was scheduled for 2014. This is the first national curriculum for Australia. Before 2011, each state in Australia had its own curriculum. The curriculum before the current national curriculum was implemented between 1993 and 2009. Different states may have had different grade coverages for the science curriculum before 2011. In most states, the curriculum covered a number (2 or 3) of grade levels at a time before the current curriculum.

The current Australian science curriculum aims to provide students with a solid foundation in science knowledge, understanding, skills, and values on which further learning and adult life can be built (ACARA, 2009, p. 5). The Australian science curriculum is formed around three strands: science understanding, science as a human endeavor, and science inquiry skills (ACARA, 2009). As described in the following figure, there are six key ideas in the science

curriculum: Patterns, order and organization; Form and function; Stability and change; Scale and measurement; Matter and energy; and Systems.

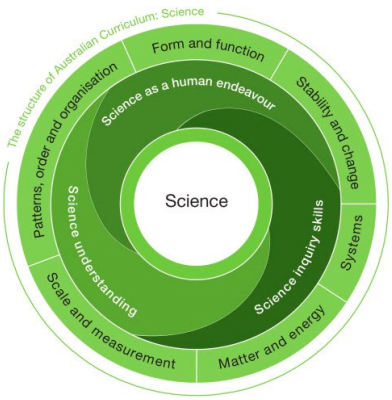


Figure 1. Key Ideas of the Science Curriculum (ACARA, 2020a)

The Australian science curriculum is described year by year. It also provides guidelines for year groupings. The curriculum focuses on years K-2 is "awareness of self and the local world;" the curriculum focus of years 3-6 is "recognizing questions that can be investigated scientifically and investigating them" (ACARA, 2009). "Forms use and transfer of energy" is listed as an essential part of the science understanding strand of years 3-6, while it is not listed in the years K-2. The specific content description and elaborations of each year's physical science curriculum are listed in Table 2. Energy is an important concept in year 1, year 3, year 5, and year 6. As we can see in the table, the energy concepts presented in the science curriculum are similar to those in the learning progression for the energy topic. The source of energy is covered in year 1 and year 5. Energy transfer in the heat is covered in year 2. Energy transfer in electricity is covered in year 6. Energy conservation and degradation are not covered in primary schools' science curriculum.

Table 2. Curriculum Content Description and Elaborations from Year 1- Year 6 (ACARA, 2020b)

Year	Curriculum content description	Elaborations
Year 1	Light and sound are produced by a range of sources and can be sensed	• recognizing senses are used to learn about the world around us: our eyes to detect light, our ears to detect sound, and touch to feel vibrations • identifying the sun as a source of light • recognizing that objects can be seen when light from sources is available to illuminate them • exploring different ways to produce sound using familiar objects and actions such as striking, blowing, scraping, and shaking • comparing sounds made by musical instruments using characteristics such as loudness, pitch and actions used to make the sound
Year 2	A push or a pull affects how an object moves or changes shape	• exploring ways that objects move on land, through water and in the air • exploring how different strengths of pushes and pulls affect the movement of objects • identifying toys from different cultures that use the forces of push or pull • considering the effects of objects being pulled towards the Earth
Year 3	Heat can be produced in many ways and can move from one object to another	• describing how heat can be produced such as through friction or motion, electricity, or chemically (burning) • identifying changes that occur in everyday situations due to heating and cooling • exploring how heat can be transferred through conduction

Table 2 Continued

Year	Curriculum content description	Elaborations
		• recognizing that we can feel the heat and measure its effects using a thermometer
Year 4	Forces can be exerted by one object on another through direct contact or from a distance	• observing qualitatively how speed is affected by the size of a force • exploring how non-contact forces are similar to contact forces in terms of objects pushing and pulling another object • comparing and contrasting the effect of friction on different surfaces, such as tires and shoes on a range of surfaces • investigating the effect of forces on the behavior of an object through actions such as throwing, dropping, bouncing, and rolling • exploring the forces of attraction and repulsion between magnets
Year 5	Light from a source forms shadows and can be absorbed, reflected and refracted	• drawing simple labelled ray diagrams to show the paths of light from a source to our eyes • comparing shadows from point and extended light sources such as torches and fluorescent tubes • classifying materials as transparent, opaque or translucent based on whether light passes through them or is absorbed • recognizing that the color of an object depends on the properties of the object and the color of the light source • exploring the use of mirrors to demonstrate the reflection of light • recognizing the refraction of light at the surfaces of different transparent materials, such as when light travels from air to water or air to glass

Table 2 Continued

Year	Curriculum content description	Elaborations
Year 6	Electrical energy can be transferred and transformed in electrical circuits and can be generated from a range of sources	• recognizing the need for a complete circuit to allow the flow of electricity • investigating different electrical conductors and insulators • exploring the features of electrical devices such as switches and light globes • investigating how moving air and water can turn turbines to generate electricity • investigating the use of solar panels • considering whether an energy source is sustainable

Note. The descriptions and elaborations are direct quotations from ACARA (2020b).

2.6.2 Science Curriculum of Hong Kong

Aiming to stimulate students' thinking and develop their capabilities to "Learn to Learn," the Education Bureau of Hong Kong launched a curriculum reform in 2000. The current curriculum of Hong Kong is based on the General Studies for Primary Schools (GS) curriculum, which was introduced in 2002. The Curriculum Development Council (CDC) of Hong Kong updated the GS curriculum guide in 2011 and then in 2017. Science is taught as part of the subject General Studies at the elementary level. "Science and Technology in Everyday Life" is an important strand in the GS curriculum. The aim of this strand is to "arouse students' curiosity and interest in science and technology through hands-on and minds-on activities and help them develop basic science process skills and technology learning skills" (CDC, 2017a, p. 29).

The curriculum framework describes what students should know, value, and be able to do from three interconnected perspectives: Knowledge and understanding; Skills; and Values and attitudes. (The relations among the three perspectives are depicted in Figure 2.) The curriculum

specifically describes the learning objectives from these three perspectives of each strand. In addition, there are two stages, i.e., key stage 1 and key stage 2, in the learning objectives. The curriculum framework of GS continues to be updated responding to the changes and challenges in society and around the world. For instance, the curriculum remains open and flexible, with the following new emphases added in 2017: developing STEM education, and deepening values education (CDC, 2017a).

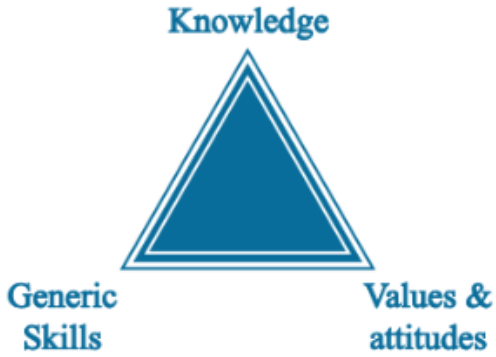


Figure 2. Interconnection among Three Perspectives (CDC, 2017a, p. 12)

Energy-related topics are an important part of the learning objectives. For instance, "to recognize sources of energy and know their uses in everyday life" is the first learning objective presented in the "Science and Technology in Everyday Life" strand. At key stage one, "sources of energy and uses of energy in everyday life (e.g., light and electricity)" is listed as one of the core learning elements. At key stage two, "examples of energy and conversion of energy (e.g., light, sound, electricity)" is listed as one of the core learning elements. The examples of themes for the GS primary curriculum provided in the guide also list learning elements related to energy at primary 1, primary 4, primary 5, and primary 6. "Energy and Change" is one of six strands in the major learning elements of the Science Education curriculum. The specific learning objectives in

the primary grades 1-6 are listed in Table 3. Unlike Australia's science curriculum, Hong Kong's curriculum description is much simpler. However, the first two stages of learning progressions of energy are also covered in Hong Kong's curriculum: energy sources and energy transfer.

Table 3. Learning Objective of Energy and Change by Stage (CDC, 2007b, p. 29)

Stage	Energy and Change
Learning Objectives at Key Stage 1 (Primary 1 - 3)	To recognize sources of energy and know their uses in daily life; To recognize heat transfer and some related phenomena; To understand the need for saving energy; To describe energy use at home and in school.
Learning Objectives at Key Stage 2 (Primary 4 - 6)	To recognize some patterns or phenomena related to light, sound, electricity and object movement; To recognize different forms of energy involved in energy change; To use energy wisely and save energy in daily life; To recognize the safety measures in using energy of different forms in daily life.

2.6.3 Science Curriculum of Ontario

The Ontario province of Canada was one of the benchmarking participants in TIMSS across the years of 2007, 2011 and 2015. Benchmarking participants are the states and districts that participated in TIMSS with the opportunity to assess their students' achievement from an international comparative perspective and view their curriculum and instruction within an international context. Canada does not have a uniform national curriculum, and each province is responsible for developing its own ministry-established common curriculum. The current science curriculum of Ontario was originally developed in 1998. The official curriculum document about the science discipline is *The Ontario Curriculum, Grades 1–8: Science and Technology*. The curriculum was updated in 2007 and implemented in September 2008. There are three major goals

outlined in the curriculum: 1) “to relate science and technology to society and the environment” (OME, 2008, p. 3).; 2) “to develop the skills, strategies, and habits of mind required for scientific inquiry and technological problem solving” (OME, 2008, p. 3).; and 3) “to understand the basic concepts of science and technology” (OME, 2008, p. 3).

Understanding matter and energy is one of the strands of the Ontario science curriculum. The curriculum describes the big ideas for each strand’s fundamental concepts by different grades. Students broaden and deepen their understanding about the fundamental concepts as they progress through the grades in the curriculum (OME, 2008). Energy is one of the fundamental concepts in the curriculum. In addition, there are two sets of expectations for each grade’s strand: overall expectations and specific expectations. The specific expectations are described from three perspectives: relating science and technology to society and the environment; developing investigation and communication skills; understanding basic concepts. Table 4 lists the big ideas and overall expectations corresponding to each grade and strand. Ontario science curriculum’s description about energy is the most specific among three selected jurisdictions. Due to their length, specific expectations are not included here. The first two stages of learning progressions of energy: “energy sources” and “energy transfer and transformation” are covered in the Ontario science curriculum from lower to upper grades.

Table 4. Big Ideas and Overall Expectations of Energy from the Ontario Curriculum (OME, 2008, p. 50, p. 61, p. 63, p. 76, p. 86, p. 90, p. 104, p. 107, p. 118)

Grade & Strand	Big ideas	Overall expectations
Grade 1 Understanding matter and energy (Energy in our lives)	Everything that happens is a result of using some form of energy.	1. Assess uses of energy at home, at school, and in the community, and suggest ways to use less energy; 2. Investigate how different types of energy are used in daily life; demonstrate an understanding that energy is something that is needed to make things happen, and that the sun is the principal source of energy for the earth.
Grade 2 Understanding structures and mechanisms (Movement)	Simple machines help objects to move. *(Overall expectations 1, 2, and 3)* Mechanisms are made up of one or more simple machines. *(Overall expectation 2)* Simple machines and mechanisms make life easier and/or more enjoyable for humans. *(Overall expectation 1)*	1. Assess the impact on society and the environment of simple machines and mechanisms; 2. Investigate mechanisms that include simple machines and enable movement; 3. Demonstrate an understanding of movement and ways in which simple machines help to move objects.
Grade 2 Understanding matter and energy (Properties of liquid and solids)	Materials that exist as liquids and solids have specific properties. *(Overall expectations 2, and 3)*	1. Assess ways in which the uses of liquids and solids can have an impact on society and the environment; 2. Investigate the properties of and interactions among liquids and solids; 3. Demonstrate an understanding of the properties of liquids and solids.

Table 4 Continued

Grade & Strand	Big ideas	Overall expectations
Grade 3 Understanding matter and energy (Forces causing movement)	There are several types of forces that cause movement. *(Overall expectations 1, 2, and 3)*	1. Assess the impact of various forces on society and the environment; 2. Investigate devices that use forces to create controlled movement; 3. Demonstrate an understanding of how forces cause movement and changes in movement.
Grade 4 Understanding structures and mechanisms (Pulleys and gears)	Pulleys and gears make it possible for a small input force to generate a large output force. (*Note*: Grade 4 students need to understand mechanical advantage only in its qualitative sense). *(Overall expectation 1)* Gears are specialized wheels and axles that are used daily in many machines. *(Overall expectations 1, 2, and 3)*	1. Evaluate the impact of pulleys and gears on society and the environment; 2. Investigate ways in which pulleys and gears modify the speed and direction of, and the force exerted on, moving objects; 3.Demonstrate an understanding of the basic principles and functions of pulley systems and gear systems.
Grade 5 Understanding matter and energy (Light and sound)	Light and sound are forms of energy with specific properties. *(Overall expectations 2 and 3)* Sound is created by vibrations. *(Overall expectations 2 and 3)* Light is required to see. *(Overall expectation 3)* Technological innovations involving light and sound have an impact on the environment. *(Overall expectation 1)*	1.Assess the impact on society and the environment of technological innovations related to light and sound; 2. Investigate the characteristics and properties of light and sound; 3.Demonstrate an understanding of light and sound as forms of energy that have specific characteristics and properties.

Table 4 Continued

Grade & Strand	Big ideas	Overall expectations
Grade 5 Understanding matter and energy (Properties and changes in matter)	Matter that changes state is still the same matter. *(Overall expectations 2 and 3)*	1. Conduct investigations that explore the properties of matter and changes in matter; 2. Demonstrate an understanding of the properties of matter, changes of state, and physical and chemical change.
Grade 5 Understanding earth and space system (Conservat-ion of energy and resources)	Energy sources are either renewable or non-renewable. *(Overall expectation 3)*	1. Demonstrate an understanding of the various forms and sources of energy and the ways in which energy can be transformed and conserved.
Grade 6 Understanding matter and energy (Electricity and electrical devices)	Electrical energy can be transformed into other forms of energy. *(Overall expectations 2 and 3)*	1. Investigate the characteristics of static and current electricity, and construct simple circuits; 2. Demonstrate an understanding of the principles of electrical energy and its transformation into and from other forms of energy.

2.7 Cognitive Diagnostic Models

This study proposes to apply cognitive diagnostic models (CDMs) to detect students' knowledge mastery patterns for energy concepts to address Research Questions 1 and 2. I briefly mentioned that CDMs are also one of the methods used to develop learning progression in section 2.2. I will introduce these models specifically in this section. I will start with an overview of CDMs. Then, I will define the *attribute* and *Q matrix*, which are two important terms in CDMs. I will also briefly introduce the classification of CDMs. The specific statistical models of CDMs that this study intends to use are not covered in this section. Instead, I will introduce the statistical models in Chapter 3.

2.7.1 Overview of Cognitive Diagnostic Models

CDMs are designed for differentiating a large number of skills or attributes at a fine cognitive grain size level to provide diagnostic feedback to learners (Rupp et al., 2010). They are special cases of latent class models that characterize the relationship of observable data to a set of categorical latent ability attributes (typically dichotomous) (Templin & Henson, 2006). CDMs can diagnose the presence or absence of each attribute for every student and illuminate different mastery patterns. In achievement testing contexts, the presence or absence of attributes is referred to as skills mastery and non-mastery, which are represented by a vector of binary latent variables. CDMs also provide diagnostic feedback about test-takers' or learners' master or non-mastery of the subskills.

2.7.2 Attributes

Conceptually, the term attributes refer to "skills, dispositions, or any other constructs that are related to behavioral procedures or cognitive processes that a learner must engage in to solve an assessment item" (Carragher et al., 2019). Psychometrically, attributes refer to unobserved (latent) variables in a statistical model, which are measured through assessment items and encoded in a Q matrix (Carragher et al., 2019). (I will discuss the Q matrix in the following section.) In CDM contexts, latent attributes can be binary, categorical polytomous, or ordinal polytomous. For instance, Tatsuoka (1983) defined eight attributes for solving fraction subtraction items and these attributes were all binary. These eight attributes are: convert a whole number to a fraction; separate a whole number from a fraction; simplify before subtracting; find a common denominator; borrow from whole number part; column borrows to subtract the second numerator from the first; subtract numerators; and reduce answers to simplest form.

CDM could classify test-takers into latent classes (i.e., attribute mastery patterns). If k is the number of attributes and each attribute is assigned into two levels (i.e., mastery or non-mastery), CDMs will generate 2^k possible latent classes. In Tatsuoka's (1983) study, for example, there are 2^8, i.e., 128 latent classes. Test-takers who are likely to have mastered corresponding attributes (probabilities above 0.5) are coded as 1, otherwise (probabilities below 0.5) they are coded as 0.

2.7.3 Q Matrix

The specification of attributes hypothesized to be measured by each item is done numerically in a table called a Q matrix (e.g., Tatsuoka, 1983; de la Torre, 2009). A well-designed Q matrix is fundamental to CDMs. The construction of the Q matrix needs to be developed from theory and empirical investigations (Rupp et al., 2010), which requires joint input from content experts, cognitive and learning theorists, and psychometricians (Liu et al., 2014). From a statistical perspective, the Q matrix is the loading matrix or pattern matrix that shows the relation of items and latent variables (Rupp et al., 2010). Generally, the items are in the rows and attributes are in the columns of the Q matrix. A cell is coded as one if item j involves attribute k for answering item j correctly, otherwise, that cell is coded as zero in the Q matrix. The Q matrix shows the cognitive specification for each test item explicitly (de la Torre, 2009). An example of a Q matrix is presented in Table 5. There are 10 items and 6 attributes measured by this example assessment. If the attribute is measured by the item, its cell is coded as one, otherwise, it is coded as zero. When an item can be solved using different strategies, the most dominant attributes or skills should be used to define the Q matrix (Lee et al., 2011). Researchers (Rupp et al., 2010) divide Q matrices into three categories: adjacency matrix, reachability matrix and reduced Q matrix. For an adjacency matrix, attributes are directly hierarchically dependent on one another. Attributes are both directly and indirectly hierarchically dependent on one another in a reachability matrix. A

reduced matrix is derived from both adjacency and reachability matrices. It is reduced because some attribute combinations that would be permissible if all attributes were independent are not permissible if an attribute hierarchy is specified (Rupp et al., 2010, p. 62).

Table 5. Sample Q Matrix of an Assessment

	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6
Item 1	0	0	0	1	0	0
Item 2	0	0	0	1	1	0
Item 3	1	0	0	1	0	1
Item 4	1	1	1	1	0	0
Item 5	1	0	0	0	0	1
Item 6	0	1	0	1	1	0
Item 7	1	1	1	0	0	0
Item 8	0	1	0	1	0	1
Item 9	0	1	0	1	1	1
Item 10	0	0	0	0	0	1

2.7.4 Classification of CDMs

According to whether the latent attributes are additive in predicting the probability of correct response or not, there are two kinds of models: compensatory latent-variable models and non-compensatory latent-variable models. In compensatory latent-variable models, a low value on one latent variable can be compensated for a high value on another latent variable to yield a correct response. For instance, the generalized deterministic inputs, noisy and-gate (G-DINA) and higher-order deterministic inputs, noisy and-gate (HO-DINA) models are compensatory latent-variable models. In non-compensatory latent-variable models, a low value on one latent

variable cannot be compensated by a high value on another latent variable. For instance, the deterministic inputs, noisy and-gate (DINA) model is non-compensatory. Whether particular observed response data is more consistent with a compensatory or non-compensatory model can be investigated using Akaike information criterion (AIC) or Bayesian information criterion (BIC) values (Rupp et al., 2010, p. 93). But theory is also needed to interpret any model results.

2.7.5 Retrofitting CDM to Non-diagnostic Framework Based Assessment

Retrofitting normally refers to the practice of fitting CDMs to responses obtained from assessments that are not designed under a diagnostic measurement framework (Liu et al., 2018, p. 359). The main differences between diagnostic and non-diagnostic measurement frameworks can be described as follows: Non-diagnostic framework assessments such as item response theory-based assessment assign scores to test-takers on a trait continuum or continua; however, diagnostic framework assessment specifies multiple categorical traits and classifies examinees on each trait as mastered or non-mastered (Liu et al., 2018). For instance, in a non-diagnostic framework, reading ability could be represented and assessed as a unidimensional latent trait (e.g., overall reading ability) or multiple latent traits (e.g., make inferences, evaluate skills), with scores assigned on the latent trait continuum(s). Retrofitting CDMs to much of the existing achievement testing data is likely to yield unsatisfactory diagnostic classification results (Gierl & Cui, 2008). However, it is still possible to retrofit CDMs to assessments developed under a non-diagnostic framework and obtain satisfactory results. In CDMs, for instance, reading ability would be assessed by multiple categorical traits (e.g., make inferences, evaluate skills) and learners are scored in a series of dichotomous or polytomous (e.g., mastery or non-mastery) latent categories. The nondiagnostic framework assessment would also have skill-level considerations during test development (Liu et al, 2018). The test's content development often breaks the theoretical larger

construct into subdomains and those subdomains could be treated as multiple attributes when retrofitting CDMs (Liu et al, 2018). Thus, it is possible to retrofit CDM to current nondiagnostic assessment to get skill-level information about test-takers, although non-diagnostic measurement frameworks normally do not provide fine-grain skill-level diagnostic feedback (Liu et al., 2018).

Retrofitting has been used widely as "an add-on to simulation studies addressing different research questions in diagnostic measurement" (Liu et al., 2018, p. 360). However, few studies focusing on the methodology of retrofitting CDMs to nondiagnostic framework besides Liu et al. (2018). Liu et al. (2018) proposed an iterative process of the methodology of retrofitting CDMs based on their review of published retrofitted examples and their experiences. There are four stages in this process (Liu et al., 2018): 1) Gathering information about the assessment, end users, and item responses, 2) specifying attributes and attribute-item relationships, 3) modeling item responses through evaluating fit statistics, and examining attribute correlations and reliability, and 4) interpreting results.

Studies have applied CDMs to existing non-diagnostic assessments to provide diagnostic feedback in different content areas. A number of applied studies have retrofitted CDM to existing reading comprehension tests (e.g., Chen & Chen, 2016; Javidanmehr & Sarab, 2019; Kasai, 1997; Jang, 2009; Lee & Sawaki, 2009b; Li et al., 2015; Mirzaei, Vincheh & Hashemian, 2020; Ravand & Robitzsch, 2018; Yi, 2012; Wang & Gierl, 2011), mathematics tests (e.g., Gierl et al., 2008; Gierl et al., 2010; Lee et al., 2011; Toker & Green, 2012; Yamaguchi & Okada, 2018; Wu et al., 2020), and listening tests (e.g., Aryadoust, 2018; Effatpanah, 2019), while few studies (Kabiri et al., 2017) have applied CDM to existing nondiagnostic assessment in science disciplines. The normal practice of retrofitting CDM of these studies is to start with model selection among different CDMs, then choose the best fitting model to conduct the analysis and provide diagnostic

feedback based on the chosen model. If it's assumed that there are hierarchical relations between attributes, an attribute hierarchy method model is often retrofitted to the non-diagnostic tests without model comparisons. In summary, CDMs have been retrofitted to non-diagnostic framework assessments in different subject areas, but the applications in science disciplines are few.

This chapter reviewed literature related to the current study. I reviewed definitions of learning progressions, methods of developing learning progressions in science disciplines, students' understanding about energy, and learning progressions, particularly related to the energy topic. Then, the opportunity to learn and instructional sensitivity were reviewed. I also briefly reviewed the science curricula of Australia, Hong Kong, and the Ontario province of Canada. Finally, I introduced CDM-related concepts and retrofitting CDM to non-diagnostic framework-based assessment. In Chapter 3, I will introduce the methods this study will use.

CHAPTER 3. METHOD

This study hypothesizes a learning progression of energy based on previous research (Lacy et al., 2014; Neumann et al., 2013) study sequenced as 1) forms of energy; 2) transfer and transformations of energy; 3) dissipation and degradation of energy; and 4) conservation of energy. I will examine the extent to which the hypothesized learning progression matches students' observed progression in understanding the energy concept using CDM (Research Question 1). I will also compare the differences in students' knowledge mastery patterns for different countries (Research Question 2). Finally, since different jurisdictions implement different science curricula, I will also explore how the intended curriculum may relate to students' understanding of energy across different countries (Research Question 3). Specifically, I will explore how OTL may affect students' understanding and assessment items' instructional sensitivity. The items' instructional sensitivity analysis will examine the validity of the items and will inform the discussion of the differences in students' knowledge mastery patterns for different countries.

This chapter will describe the cross-national science assessment data, specific variables, and statistical models related to the four research questions. For Research Questions 1 and 2, I will describe the Q matrix development and validation in this chapter. I will introduce the deterministic inputs, noisy and-gate (DINA) model. Finally, since I will investigate Question 3 by applying logistic regression, I will present that model.

3.1 Data

This study will use Trends in International Mathematics and Science Study (TIMSS) student achievement test data and curriculum data from Grade 4 and Year 2011. TIMSS applies a two-stage random sample design: in the first stage, a sample of schools was drawn; in the second

stage, one or more intact classes of students were selected from sampled schools (Martin et al., 2016). The science assessment framework is organized around two dimensions: content and cognitive. Specifically, the Grade 4 science assessment framework is designed from three major content domains, i.e., life science, physical science, and earth science. There are three cognitive domains of TIMSS science assessment: knowing, applying and reasoning. The TIMSS data sets are suitable for the current investigation because: 1) they provide reliable data on students' science achievement, including performance on the energy topic, which is the main focus of the study; and 2) it also provides curriculum data from different countries, which allows me to analyze and compare how the science curriculum may relate to students' understanding of energy across countries.

Three jurisdictions, Australia, the Ontario province of Canada, and Hong Kong, are chosen for analysis since their curricula have changed or been updated before 2011. In 2011, Australia had 6,146 students who participated, Ontario had 4,568 students who participated, and Hong Kong had 3,957 students who participated in TIMSS.

3.2 Variables

This section will introduce the variables that will be used in the study. I will describe the student level and teacher level variables. The specific variables that will be used in this study are listed in Table 6.

Student level variables. In the proposed study, I will focus on achievement test item variables assessing each student's knowledge mastery of energy topics under the physical science domain in the year 2011. The cognitive domain of each item is specified in the assessment's framework. Specific item IDs are listed in Table 6. Some items had several subitems. For instance, item S031197 had two subitems, resulting in two response variables: S031197A, and S031197B.

These subitems will be treated as independent items since they provide unique information about students' responses. There are 28 items included in this study counting all the subitems. Twelve items only had two score categories and all other items had more than two score categories. It should be noted here that the multiple-choice items will be classified into two categories (correct will be coded as 1 and incorrect will be coded as 0) in CDM. For open-ended questions with more than two response categories, all the correct response categories are coded as 1 and all the incorrect response categories are coded as 0. (Items have multiple types of correct answers and/or multiple types of incorrect answers, but do not have 'partially correct' answers.)

Country level variables. Country level variables came from TIMSS Test-Curriculum Matching Analysis (TCMA). TCMA was conducted to investigate the appropriateness of the TIMSS mathematics and science assessments for the fourth and eighth grade students in the participating countries (Foy et al., 2013, p. 102). Binary coding (Yes/No) indicated whether items in the assessment were included in the national curriculum, or not, for a particular participating jurisdiction. (It should be noted here that there is no existing variable indicator for this information. Only a table is presented showing the binary coding, through which I can code the table into variables).

Table 6. Selected TIMSS Variable List of TIMSS 2011 Grade 4

Variable	Cognitive Domain	Question type	Response Category
S031273	Applying	Multiple choice	4
S031076	Reasoning	Open-ended	3
S031077	Applying	Multiple choice	4
S031197A	Knowing	Open-ended	7
S031197B	Knowing	Open-ended	7
S031298	Applying	Multiple choice	4
S031299	Knowing	Open-ended	5

Table 6 Continued

Variable	Cognitive Domain	Question type	Response Category
S041311	Applying	Multiple choice	4
S041120	Knowing	Multiple choice	4
S041067	Knowing	Open-ended	2
S041069	Applying	Multiple choice	4
S041070	Applying	Multiple choice	4
S041191	Knowing	Multiple choice	4
S041195	Applying	Open-ended	3
S051119	Reasoning	Open-ended	3
S051074	Applying	Open-ended	3
S051179	Applying	Multiple choice	4
S051201	Applying	Multiple choice	2
S051121A	Knowing	Multiple choice	2
S051121B	Knowing	Multiple choice	2
S051121C	Knowing	Multiple choice	2
S051121D	Knowing	Multiple choice	2
S051121E	Knowing	Multiple choice	2
S051188A	Knowing	Multiple choice	2
S051188B	Knowing	Multiple choice	2
S051188C	Knowing	Multiple choice	2
S051188D	Knowing	Multiple choice	2
S051188E	Knowing	Multiple choice	2

3.3 Analysis

The data analysis will be divided into four steps. First, I will report the descriptive analysis of test item variables. Second, I will analyze the achievement test items of each jurisdiction using CDM to obtain students' mastery patterns. CDM analysis will be conducted using the R software CDM package (Robitzsch et al., 2020). Since most of the items only measure one attribute in this study (see sections 3.5 and 4.3, 85.71% items in the proposed Q matrix and 91.66% items in the

final Q matrix), the results of CDM analysis are expected to be similar across different models. Thus, the parsimonious DINA model is chosen. I will use the parsimonious and interpretable "deterministic, inputs, noisy, 'and' gate" (DINA; see Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1977) model, with results obtained by weighted maximum likelihood estimation. Third, I will compare the differences between students' mastery patterns in different countries from step two. Fourth, I will analyze how the intended curriculum may influence students' mastery or understanding of the particular topic domain, i.e., energy. I will use logistic regression to see if students' performance on each item differs depending on whether it was covered or not covered in the national curriculum, i.e., detect the instructional sensitivity of each item, using Mplus software. The nested structure of the data will be accounted for by using complex adjustment analysis (Stapleton, 2016). The missing data will be handled through full information maximum likelihood estimation.

3.4 Analysis Model

3.4.1 DINA Model

Among CDMs, the deterministic inputs, noisy and-gate (DINA) model is widely used for its simplicity and interpretability (de la Torre & Douglas, 2004). The DINA model (e.g., Haertel, 1989; Junker & Sijtsma, 2001) is a noncompensatory model with a conjunctive condensation rule. The respondent needs to master all the attributes required for a particular item (Rupp et al., 2010) in the DINA model to have a high probability of answering the item correctly. A latent variable η_{ij} represents whether or not respondent *i* has all of the required attributes to resolve item *j* in the DINA model (Hsu & Wang, 2015). The latent variable η_{ij} is a function of the determinist input which is defined as *Equation 2*:

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

where η_{ij} = 1 when respondent *i* masters all of the required attributes for item *j*, η_{ij} = 0 when respondent *i* lacks at least one of the required attributes, α_{ik} is attribute vector for respondent *i* and attribute *k*. If an attribute is not measured by an item, then q_{ik} = 0, which means that α_{jk}^{0}=1. If an attribute is measured by an item, then q_{ik} = 1, which means that whether the respondent masters the attribute or not matters for the probability of correct response (Rupp et al., 2010).

DINA model accounts for the noise (i.e., random error) introduced in the underlying stochastic process with *slip* and *guessing* parameters. Even respondents who have mastered all measured attributes for an item can slip and miss the item. The respondents who have not mastered at least one of the measured attributes can guess and answer a question correctly (Rupp et al., 2010). The probability of respondent *i* with the skill vector α_i answering item *j* correctly in the DINA model is defined as *Equation 3*:

$$P_j(\alpha_i) = P(\alpha_i) = g_j^{1-\eta_{ij}}(1 - s_j)^{\eta_{ij}} \quad (3)$$

where g_j is the guessing parameter, s_j is the slipping parameter, and all other terms are as defined previously. Given local independence among items and examinees, the joint likelihood function of the DINA model is defined as *Equation 4*:

$$L(s, g, \alpha|X) = \prod_{i=1}^{N} \prod_{j=1}^{J} [(1 - s_j)^{X_{ij}} s_j^{1-X_{ij}})]^{\eta_{ij}} [\, g_j^{X_{ij}} (1 - g_j^{1-X_{ij}})]^{1-\eta_{ij}} \quad (4)$$

where *s* is the vector that consists of all slip parameters and *g* is the vector that consists of all guessing parameters in the test; N is the examinee sample size; and the others have been defined previously.

DINA model is one of the most parsimonious and interpretable CDM models with only two parameters (i.e., guessing parameter and slipping parameter). de la Torre and Lee's (2008)

study found that DINA keeps item-level information generated from item response theory models. TIMSS assessment items are selected based on IRT item statistics (Martin & Mullis, 2011). The DINA model is consistent with generalizations of standard IRT assumptions (e.g., local independence, monotonicity) (Junker & Sijtsma, 2011), and has been shown to be sensitive to attributes even if the items were designed to fit by an item response theory model (Junker & Sijtsma, 2011). Retrofitting TIMSS item response data with the DINA model could reveal important cognitive and content attributes, providing diagnostic information on students' attribute mastery (Choi, et al., 2015; Lee et al., 2011).

3.4.2 Logistic Regression

In this study, I will use logistic regression to detect items' instructional sensitivity. The outcome variable Y_{ij} indicates the natural log odds of a correct response for student i on item j. For items with more than two ordered score categories, I will combine different correct score categories into one category as correct. Similarly, I will combine all incorrect score categories into one category as incorrect. This coding procedure matches the item responses' coding in the CDM analysis. Whether the item is covered in the national curriculum, from the TCMA results, will be the independent variable ($Curriculum_{ij}$). Items covered in the national curriculum will be coded as 1, otherwise they will be coded as 0. I will use full information maximum likelihood estimation to deal with missing data on the outcome variables. The equation I will use for logistic regression, with one model for each item, is defined as *Equation 5*.

$$Y_{ij} = \beta_{0j} + \beta_{1j}\, Curriculum_{ij} \qquad (5)$$

where Y_{ij} is the log odds of a correct response for student *i* on item *j*, $_{0j}$ is the log odds when the independent variable is zero, and β_{1j} is the coefficient indicating instructional sensitivity regarding curriculum.

Some studies controlled for students' ability in exploring items' instructional sensitivity (e.g., D'Agostino et al., 2007; Li et al., 2017). The rationale is that students' ability would relate to students' performance in each item, while performance on an instructionally sensitive item is expected to increase with effective teaching (Baker, 1994). Thus, I will also examine the instructional sensitivity of selected items, controlling for students' ability ($Ability_{ij}$). Students' ability is indicated by the number of attributes each student mastered from CDM analysis. The equation for the second part of instructional sensitivity analysis, controlling students' ability is defined as *Equation 6*

$$Y_{ij} = \beta_{0j} + \beta_{1j}\, Curriculum_{ij} + \beta_{2j}\, Ability_{ij} \quad (6)$$

3.5 Q Matrix Development and Validation

3.5.1 Development of the Draft Q matrix

A well-designed Q matrix is essential in CDMs. I developed a draft Q matrix based on the literature related to learning theory, learning progressions of energy in the physical science domain, and the TIMSS assessment framework. The Q matrix reflected the learning progression I proposed based on Lacy et al. (2014)'s study, i.e., 1) forms of energy; 2) transfer and transformations of energy; 3) dissipation and degradation of energy; and 4) conservation of energy. Since I will use the Grade 4 items, none of which are related to "dissipation and degradation of energy" or "conservation of energy," I deleted these two strands in the Q matrix. The proposed Q matrix (see Table 7) thus has two higher-order strands ("forms of energy" and "transfer and transformations

of energy"). According to the TIMSS science assessment framework (Mullis et al., 2009) and Quebec *Progression of Learning Science and Technology* (Quebec Education Program [QEP], 2009), I then classified the existing items into six specific attributes in the proposed Q matrix. Most of the attribute descriptions came from objectives proposed in the TIMSS science assessment framework and Quebec *Progression of Learning Science and Technology* for elementary school except the last attribute "Understands heat transfer" which was most clearly characterized in Hong Kong's curriculum. There are two attributes under the "forms of energy" strand: 1) "Describes different forms of energy (mechanical, electrical, light, chemical, heat, sound, nuclear)"(QEP, 2009, p.5); and 2) "Identifies sources of energy in his/her environment (e.g., moving water, the chemical reaction in a battery, sunlight)"(QEP, 2009, p.5). There are four attributes under the "transfer and transformations of energy" strand: 1) Distinguish between substances that are conductors and those that are insulators; 2) "Explain that simple electrical systems, such as a flashlight, require a complete (unbroken) electrical pathway" (Mullis et al., 2009, p.59); 3) "Relate familiar physical phenomena to the behavior of light (e.g., reflections, rainbows, shadows)" (Mullis et al., 2009, p.59); and 4) Understand heat transfer. According to the hypothesized learning progression from Lacy et al. (2014) and Neumann et al. (2013), I hypothesize that these six attributes in this study are not necessarily fully ordered, but the four attributes for the strand "transfer and transformations of energy" are followed by the two attributes for the strand "forms of energy" in learning sequence, and attributes are not ordered within the strands. There are 28 items in the proposed Q matrix, and 24 out of 28 items (85.71%) are only measuring one attribute.

Table 7. Proposed Q matrix

Items	A1	A2	A3	A4	A5	A6
S031273	0	0	1	0	0	1
S031076	0	1	0	0	0	0
S031077	0	1	1	0	0	0
S031197A	1	0	0	0	0	0
S031197B	1	0	0	0	0	0
S031298	0	1	0	0	0	1
S031299	1	0	0	0	0	0
S041311	1	0	0	0	0	1
S041120	0	1	0	0	0	0
S041067	1	1	0	0	0	0
S041069	0	0	0	0	1	0
S041070	0	0	0	0	1	0
S041191	0	0	1	0	0	0
S041195	0	0	0	1	0	0
S051119	0	1	0	0	0	0
S051074	0	0	0	1	0	0
S051179	0	0	0	0	1	0
S051201	0	0	1	0	0	0
S051121A	0	0	1	0	0	0
S051121B	0	0	1	0	0	0
S051121C	0	0	1	0	0	0
S051121D	0	0	1	0	0	0
S051121E	0	0	1	0	0	0
S051188A	0	1	0	0	0	0
S051188B	0	1	0	0	0	0
S051188C	0	1	0	0	0	0
S051188D	0	1	0	0	0	0
S051188E	0	1	0	0	0	0

Note. A1 = Describes different forms of energy (mechanical, electrical, light, chemical, heat, sound, nuclear); A2 = Identifies sources of energy (e.g. moving water, the chemical reaction in a battery, sunlight); A3 = Distinguishes between substances that are conductors and those that are insulators; A4 = Explain that simple electrical systems, such as a flashlight, require a complete (unbroken) electrical pathway; A5 = Relate familiar physical phenomena to the behavior of light (e.g., reflections, rainbows, shadows); A6 = Understand heat transfer

3.5.2 Expert Review

Then, five experts from science education were invited to review the draft matrix and the proposed attributes. Among the five experts, two experts were K-9 physical science teachers. One of the science teachers had five years' teaching experience and another one had one year's teaching experience. One expert was a faculty member of physical science education and he had been a physical science teacher for two years. One expert was a faculty member of science education and he had been a physical science teacher for five years. Another expert was a fourth-year doctoral candidate in science education. All the five experts had obtained their bachelor's degree and master's degree in science education. I conducted an interview with each expert by discussing each item's endorsed attributes. The length of each interview was about an hour to an hour and a half. Experts were asked whether each endorsed attribute was correct or not, and what revisions needed to be made. Experts were also asked whether new attributes needed to be added to fully describe the available items' content. The Q matrix was revised after the expert review.

3.5.3 Q Matrix Validation Using Real Data

The revised Q matrix was analyzed and validated using CDMs and two split data sets. The current dataset was divided into halves. I randomly drew half of the data within each jurisdiction and combined those into one dataset for the first validation of the Q matrix. I combined the rest of each jurisdiction's data for the second validation. I conducted the validation analysis using the DINA model. I revised the Q matrix according to the first analysis result, again referring to the expert review information. Then, I used the second half of the combined data set to do the second validation analysis. I used weighted maximum likelihood estimation to deal with specific sampling features and missingness in the survey data. Maximum likelihood estimation allows us to "estimate a set of parameters that maximize the probability of getting the data that was observed" (Newman,

2003, p. 332), and it is an effective way to treat missingness on outcome variables. Adding sampling weights to the analysis allows the sample results to reconstruct those that would be obtained if it was a random draw from the total population and leads to accurate population parameter estimates (Friedman, 2013). The absolute model fit will be identified using the Standardized Mean Square Root of Squared Residuals (SRMSR), mean of absolute deviations in observed and expected correlations (MADcor), mean of absolute values of Q3 statistic (MADQ3), and a maximum of all chi-square statistics (max(X2)). To calculate MADQ3, residuals $\varepsilon_{ni} = X_{ni} - e_{ni}$ of observed and expected responses for respondents *n* and items *i* are constructed (Robitzsch et al., 2020, p. 167). Then, the average of the absolute values of pairwise correlations of these residuals is computed for MADQ3 (Robitzsch et al., 2020, p. 167). The max(X2) statistic is the maximum of all item pair x_{jj}^2, statistics, and a statistically significant *p*-value shows that some item pairs violate statistical independence (Robitzsch et al., 2020). Thus, a non-significant value for max(X2) ($p > 0.05$) indicates a good fit. For all other model fit indices, the model fits the data better if these fit indices are close to zero (Ravand & Robitzsch, 2015).

Item level fit will be evaluated using the item fit Root Mean Square Error of Approximation (RMSEA) and item discrimination index (IDI). The criteria for interpreting item-fit RMSEA are as follows: item-fit RMSEA below 0.05 indicates good fit, item-fit RMSEA below 0.10 indicates moderate fit, and item-fit RMSEA above 0.10 indicates poor fit (Kunina-Habenicht et al., 2009). IDI for each item is calculated as $IDI_j = 1 - s_j - g_j$ (Lee et al., 2012), where s_j is the slipping parameter and g_j is the guessing parameter. IDI can be used as a diagnostic index about how an item discriminates between students having a response probability of $1 - s_j$ possessing all skills, and students guessing with probability (g_j) without possessing any skills (George et al., 2016). IDIs close to 1 indicates a good discrimination of the item, and IDI values close to 0 indicate items

with a low discrimination (George et al., 2016). The matrix may be revised according to the analysis result. Similar attributes may be combined to reduce the number of attributes according to the analysis results. For items that do not have a good item fit, their attribute classifications will be reconsidered, or the item may be deleted from the model if it violated the assumptions of the model's assumption. The final cognitive diagnostic assessment analysis will be based on the validated Q matrix.

This chapter described the data, variables, selected models, statistical models, and planned analytic strategies to address my research questions. This chapter also described the proposed Q matrices and validation process of the Q matrices for the cognitive diagnostic models. In the next chapter, I will present the results.

CHAPTER 4. RESULTS

In this chapter, I will document the Q matrix validation results from both expert reviews and using real data. The final Q matrix will be presented. I will also present descriptive statistics of items. The overall attribute mastery probability and latent class mastery pattern profile from cognitive diagnostic models across the three jurisdictions will also be presented. The overall attribute mastery probability could help answer the first research question of the study, i.e., the extent to which the hypothesized learning progression matches students' observed progression in understanding the energy concept using the cognitive diagnostic model. The latent class mastery pattern profile would help answer the second research question, i.e., the similarities and differences in students' knowledge mastery patterns for different questions. I will also present the instructional sensitivity of each item, which helps answer the study's third research question.

4.1 Descriptive Statistics of Items

Table 8 presents descriptive statistics of each item for each jurisdiction, i.e., the proportion-correct item difficulty. Item S041195 and item S051074, both about simple electrical circuits, are two items with the lowest proportion of correctness for all the three selected jurisdictions. Item S051188C and item S051188D about sources of energy are two items with the highest proportion of correctness.

Table 8. Item Statistics: Proportion-correct Item Difficulty

Item	Proportion-correct Item Difficulty		
	Australia	Hong Kong	Ontario
S031273	0.66	0.87	0.63
S031076	0.37	0.50	0.53
S031077	0.76	0.84	0.80
S031197A	0.86	0.81	0.85
S031197B	0.77	0.69	0.77
S031298	0.29	0.44	0.26
S031299	0.45	0.54	0.57
S041311	0.94	0.96	0.94
S041120	0.45	0.26	0.47
S041067	0.65	0.66	0.63
S041069	0.60	0.73	0.57
S041070	0.63	0.51	0.62
S041195	0.14	0.21	0.20
S051119	0.26	0.32	0.39
S051074	0.18	0.23	0.12
S051179	0.85	0.74	0.88
S051201	0.55	0.17	0.55
S051121A	0.84	0.89	0.90
S051121B	0.82	0.84	0.78
S051121C	0.76	0.89	0.69
S051121E	0.72	0.94	0.72
S051188A	0.84	0.89	0.89
S051188B	0.75	0.89	0.77
S051188C	0.93	0.93	0.95
S051188D	0.90	0.93	0.94
S051188E	0.69	0.92	0.75

4.2 Q Matrix Validation Results: Expert Review

I summarized the four experts' feedback and revised the proposed Q matrix according to their feedback. One attribute's description was revised, some items were deleted from the matrix, and some items' attribute correspondence was changed. First, one attribute's description was revised according to experts' review. Experts commented that Attribute 6 was broad, and we narrowed down Attribute 6 "Understand heat transfer" to "Recognize that heating an object can increase its temperature and that hot objects can heat up cold objects" based on the content of TIMSS items.

Second, experts also suggested deleting some items. Item S031076 was about magnets repelling or attracting because of North/South poles repelling or attracting rather than about Attribute 2 "Identifies sources of energy." Since only one item assessed the attribute about magnetism proposed by the reviewer, which could not provide adequate estimation, this item was deleted. There were still 9 items assessing Attribute 2 in the Q matrix after item S031076 was deleted, which would not affect the testing of the hypothesized learning progressions. Item S051119 was about the reasoning of magnet property that magnets can attract pins and only this item measured this content. Thus, this item was also deleted. Although item S041311 was under the "source and effects of energy" topic area, it is about the reading of a thermometer and was not related to any attributes proposed. Since only one item was related to thermometer reading, this item was deleted. Item S041120 about the objects that produce their own light was also deleted since this item is not related to any attributes proposed.

Some items' attribute assignment changed according to experts' review. For item S031077, Attribute 2 was not endorsed since all the experts agreed that this item did not involve identifying sources of energy as proposed in Attribute 2. For item S031298, Attribute 2 was also not endorsed since students did not need to identify sources of energy to solve this problem. For item S031299,

Attribute 5 was added since it's about light rays as proposed in Attribute 5. For item S041067, Attribute 2 was deleted. For item S051201, it assessed whether students understand sweaters are insulators or not. Thus, Attribute 3 was endorsed, and Attribute 6 was not endorsed. In summary, four items were deleted and there remained six attributes after the expert reviewers' feedback. Eight items assessed the attributes of the first strand of the proposed learning progressions and seventeen items assessed the attributes of the second strand of the proposed learning progressions. The revised Q matrix is presented in Table 9.

Table 9. Revised Q Matrix

Items	A1	A2	A3	A4	A5	A6
S031273	0	0	1	0	0	1
S031077	0	0	1	0	0	0
S031197A	1	0	0	0	0	0
S031197B	1	0	0	0	0	0
S031298	0	0	0	0	0	1
S031299	1	0	0	0	1	0
S041067	1	0	0	0	0	0
S041069	0	0	0	0	1	0
S041070	0	0	0	0	1	0
S041191	0	0	1	0	0	0
S041195	0	0	0	1	0	0
S051074	0	0	0	1	0	0
S051179	0	0	0	0	1	0
S051201	0	0	1	0	0	0
S051121A	0	0	1	0	0	0
S051121B	0	0	1	0	0	0
S051121C	0	0	1	0	0	0
S051121D	0	0	1	0	0	0
S051121E	0	0	1	0	0	0
S051188A	0	1	0	0	0	0

Table 9 Continued

Items	A1	A2	A3	A4	A5	A6
S051188B	0	1	0	0	0	0
S051188C	0	1	0	0	0	0
S051188D	0	1	0	0	0	0
S051188E	0	1	0	0	0	0
total	3	5	9	2	4	2

Note. A1 = Describes different forms of energy (mechanical, electrical, light, chemical, heat, sound, nuclear); A2 = Identifies sources of energy (e.g. moving water, the chemical reaction in a battery, sunlight); A3 = Distinguishes between substances that are conductors and those that are insulators; A4 = Explain that simple electrical systems, such as a flashlight, require a complete (unbroken) electrical pathway; A5 = Relate familiar physical phenomena to the behavior of light (e.g., reflections, rainbows, shadows); A6 = Recognize that heating an object can increase its temperature and that hot objects can heat up cold objects

4.3 Q Matrix Validation Results: Using Real Data

I divided the current data into two datasets, and I did the validation based on the first half of the data first. I validated the Q matrix using the DINA model and weighted maximum likelihood estimation method as stated earlier. I checked the indices of the DINA model as the following procedure. First, I examined item-level fit indices to check how well the model fit each item's observed response data. The item level indices for the first validation are presented in Table 10. Item S041191 had a negative item discrimination index (IDI) index, -0.072, which violated the constraint of the DINA model that $g_j < 1 - s_j$ (George et al., 2016). Item S041191 was a multiple-choice item inquiring which material was the best conductor of heat. Then I double-checked this item's attribute classification (the endorsed attribute was Attribute 3 "Distinguishes between substances that are conductors and those that are insulators") and consulted with the experts again, who indicated that no further changes of this item's attribute should be made based on its content. Since there were still multiple items measuring Attribute 3, this item was deleted due to its negative IDI. All other item-level indices were good. The IDI indices ranged from 0.104 to 0.879. Item

S051121A and item S051188E were two items with the lowest IDI index 0.104 and 0.159. Item S051121A and item S051188E were also found to have local dependence with other items (as presented in the next paragraph). Thus, these two items would be deleted. All items' RMSEA values were below 0.05.

Then I checked the absolute model fit indices. All other absolute model fit indices were good: SRMSR = 0.053, MADcor = 0.039, MADQ3 = 0.076. However, the max(X2) statistic was not good: max(X2) = 42.681, $p < 0.05$. Max(X2) statistics' p-value was significant, which indicated a violation of statistical independence of the item pair. Then I checked the item pairs' local independence. Item S051121A had significant local dependence with another two items (S051121D and S051188B). Item S051121D had significant local dependence with another two items (S051121A, S051121B). S051188E had significant local dependence with another two items (S051188B, S051201). Item S051121A, S051121B, and S051121D came from the same set of items, which distinguish between substances that are conductors and those that are insulators. Item S051188B and S051188E came from the same set of items about sources of energy. Item S051201 was also about sources of energy. In addition, item S051121A and item S051188E had the lowest IDI indices among all the items. Thus, I deleted one item that had multiple local dependence with other items and the two items with a lower IDI index, i.e., S051121A, S051121D, and S051188E.

Then I used the second half of the data to check the revised Q matrix. The absolute model fit indices were all good SRMSR =0.040, MADcor = 0.034, MADQ3 = 0.080, max(X2) = 8.787, $p = 0.134$. The item-level fits were also all good. The IDI indices ranged from 0.162 to 0.722. Item S051188D had the lowest IDI index 0.162. Item S041067 and item S041069 had the highest IDI index 0.746. RMSEA statistics were all below 0.05. Table 11 presents item-level fit results for the

second validation. Most items' IDI values increased at the second validation. Table 12 presents the final Q matrix of this study. In total, there are 20 items and six attributes in the final matrix.

Table 10. Item-level Fit Indices of the First Validation

Item	Guess	Slip	IDI	RMSEA
S031273	0.538	0.124	0.337	0.027
S031077	0.492	0.017	0.491	0.015
S031197A	0.595	0.000	0.405	0.005
S031197B	0.329	0.000	0.670	0.007
S031298	0.046	0.429	0.525	0.003
S031299	0.407	0.417	0.176	0.034
S041067	0.127	0.097	0.776	0.004
S041069	0.101	0.020	0.879	0.004
S041070	0.240	0.133	0.627	0.009
S041191	0.508	0.564	-0.072	0.015
S041195	0.003	0.618	0.378	0.012
S051074	0.012	0.640	0.348	0.006
S051179	0.692	0.046	0.261	0.007
S051201	0.212	0.135	0.652	0.013
S051121A	0.811	0.085	0.104	0.016
S051121B	0.676	0.121	0.203	0.003
S051121C	0.414	0.024	0.562	0.018
S051121D	0.524	0.191	0.284	0.019
S051121E	0.367	0.000	0.632	0.016
S051188A	0.314	0.016	0.669	0.009
S051188B	0.301	0.144	0.555	0.015
S051188C	0.720	0.019	0.262	0.019
S051188D	0.696	0.048	0.256	0.028
S051188E	0.630	0.211	0.159	0.030

Table 11. Item-level Fit Indices of the Second Validation

Item	Guess	Slip	IDI	RMSEA
S031273	0.544	0.028	0.428	0.019
S031077	0.407	0.057	0.536	0.006
S031197A	0.554	0.000	0.446	0.009
S031197B	0.278	0.000	0.722	0.003
S031298	0.004	0.453	0.543	0.006
S031299	0.410	0.405	0.185	0.034
S041067	0.279	0.075	0.746	0.004
S041069	0.179	0.075	0.746	0.004
S041070	0.247	0.093	0.660	0.003
S041195	0.001	0.636	0.363	0.015
S051074	0.024	0.646	0.330	0.003
S051179	0.652	0.006	0.342	0.006
S051201	0.022	0.222	0.756	0.007
S051121B	0.653	0.087	0.260	0.003
S051121C	0.318	0.000	0.682	0.002
S051121E	0.330	0.023	0.647	0.007
S051188A	0.503	0.000	0.497	0.016
S051188B	0.536	0.098	0.366	0.047
S051188C	0.775	0.006	0.219	0.031
S051188D	0.808	0.030	0.162	0.025

Table 12. Final Q matrix

Items	A1	A2	A3	A4	A5	A6
S031273	0	0	1	0	0	1
S031077	0	0	1	0	0	0
S031197A	1	0	0	0	0	0
S031197B	1	0	0	0	0	0
S031298	0	0	0	0	0	1
S031299	1	0	0	0	1	0
S041067	1	0	0	0	0	0
S041069	0	0	0	0	1	0
S041070	0	0	0	0	1	0
S041195	0	0	0	1	0	0
S051074	0	0	0	1	0	0
S051179	0	0	0	0	1	0
S051201	0	0	1	0	0	0
S051121B	0	0	1	0	0	0
S051121C	0	0	1	0	0	0
S051121E	0	0	1	0	0	0
S051188A	0	1	0	0	0	0
S051188B	0	1	0	0	0	0
S051188C	0	1	0	0	0	0
S051188D	0	1	0	0	0	0
total	3	4	6	2	4	2

Note. A1 = Describes different forms of energy (mechanical, electrical, light, chemical, heat, sound, nuclear); A2 = Identifies sources of energy in (e.g. moving water, the chemical reaction in a battery, sunlight); A3 = Distinguishes between substances that are conductors and those that are insulators; A4 = Explains that simple electrical systems, such as a flashlight, require a complete (unbroken) electrical pathway; A5 = Relates familiar physical phenomena to the behavior of light (e.g., reflections, rainbows, shadows); A6 = Recognizes that heating an object can increase its temperature and that hot objects can heat up cold objects

Then, I also double-checked the model fit and the item fit indices for subset datasets for Australia, Hong Kong, and Ontario to ensure the model fits were all good. The absolute model fit indices were good for each jurisdiction: none of the p values of max(X2) were significant, which means that the current model fit well for each jurisdiction's data. MADcor and SRMSR values were below 0.05; MADQ3 were below 0.1 (see Table 13). The item-level fit for each jurisdiction is also acceptable: item-level RMSEA statistics are all below 0.05; IDI values are all above zero, ranging from 0.214 to 0.819.

Table 13. Absolute Model Fits Statistics for Australia, Hong Kong, and Ontario

Jurisdiction	max(X2)	MADcor	SRMSR	MADQ3
Australia	9.330 (p = 0.099)	0.033	0.043	0.089
Hong Kong	5.325 (p = 0.925)	0.037	0.049	0.072
Ontario	5.002 (p= 1.000)	0.036	0.046	0.078

4.4 Item Covered in National Curriculum or Not

A Test-Curriculum Matching Analysis (TCMA) was conducted by the International Association for the Evaluation of Educational Achievement (IEA) to investigate the appropriateness of the TIMSS 2011 mathematics and science assessments for the fourth and eighth grade students in the participating countries (Foy et al., 2013). Participating countries were asked to indicate whether the corresponding items on the TIMSS 2011 assessments were included in their national curricula or not (Foy et al., 2013). Table 14 presents the results of whether the selected items in this study were covered in the selected jurisdiction's national curriculum or not. As presented in Table 14, most items were covered in Australia and Hong Kong's national

curriculum. Ontario has fewer items that were covered in the national curriculum compared to Australia and Hong Kong.

Table 14. Item Covered in National Curriculum or Not

Items	Item Covered in National Curriculum or Not		
	Australia	Hong Kong	Ontario
S031273	not	yes	yes
S031076*	yes	yes	not
S031077	not	not	not
S031197A	yes	not	not
S031197B	yes	not	not
S031298	yes	yes	not
S031299	yes	yes	yes
S041311*	yes	yes	not
S041120*	yes	yes	yes
S041067	yes	not	not
S041069	yes	yes	yes
S041070	yes	yes	yes
S041191*	yes	yes	not
S041195	not	not	not
S051119*	yes	yes	yes
S051074	not	yes	not
S051179	yes	yes	yes
S051201	yes	yes	not
S051121A*	yes	yes	not
S051121B	yes	yes	not
S051121C	yes	yes	not
S051121D*	yes	yes	not
S051121E	yes	yes	not
S051188A	yes	yes	not
S051188B	yes	yes	not
S051188C	yes	yes	not
S051188D	yes	yes	not
S051188E*	yes	yes	not

Note. This data is from the TIMSS TCMA result (IEA, 2013).
* Items were not included in the final Q matrix.

4.5 Attribute Mastery Profile Across Three Jurisdictions

Table 15 and Figure 3 present the attribute mastery probabilities of the three jurisdictions. They show each participant population's mastery probability for each attribute, which is the relative difficulty levels of different sub-skills underlying the energy topic for each jurisdiction. As the results show, Attribute 1 "Describes different forms of energy" and Attribute 2 "Identifies sources of energy," which were mastered by 71.26% and 77.5% of the test-takers from Australia, were easiest for the Australian students. Attribute 4 "Explain that simple electrical systems, such as a flashlight, require a complete (unbroken) electrical pathway" mastered by 44.04%, was the most difficult for Australian students.

As to Hong Kong, Attribute 2 "Identifies sources of energy", which was mastered by 87.54 % of the test-takers from Hong Kong, was easiest for the Hong Kong students. Attribute 3 "Distinguishes between substances that are conductors and those that are insulators" came after Attribute 2 with a relatively high level of mastery, 80.46%. Attribute 4 "Explain that simple electrical systems, such as a flashlight, require a complete (unbroken) electrical pathway" mastered by 46.04% was also the most difficult for Hong Kong students.

As to Ontario, Attribute 1 "Describes different forms of energy" and Attribute 2 "Identifies sources of energy", which were mastered by 66.07% and 74.82% of the test-takers from Ontario, were easiest to the Ontario students. Attribute 3 "Distinguishes between substances that are conductors and those that are insulators" comes after Attribute 2 with a relatively high level of mastery, 80.46%. Attribute 4 "Explain that simple electrical systems, such as a flashlight, require a complete (unbroken) electrical pathway," mastered by 46.04%, was also most difficult to Ontario students. Overall, Ontario students' mastery probabilities for each attribute are the lowest among the three selected jurisdictions, while Hong Kong students' are the highest.

Table 15. Attribute Mastery Probabilities across Three Jurisdictions

Attribute	Attribute Mastery Probability		
	Australia	Hong Kong	Ontario
A 1	0.7126	0.5981	0.6607
A 2	0.7750	0.8754	0.7482
A 3	0.6517	0.8046	0.5970
A 4	0.4404	0.4604	0.4339
A 5	0.5690	0.6057	0.5991
A 6	0.4868	0.5913	0.4997

Note. A1 = Describes different forms of energy (mechanical, electrical, light, chemical, heat, sound, nuclear); A2 = Identifies sources of energy (e.g. moving water, the chemical reaction in a battery, sunlight); A3 = Distinguishes between substances that are conductors and those that are insulators; A4 = Explains that simple electrical systems, such as a flashlight, require a complete (unbroken) electrical pathway; A5 = Relates familiar physical phenomena to the behavior of light (e.g., reflections, rainbows, shadows); A6 = Recognizes that heating an object can increase its temperature and that hot objects can heat up cold objects

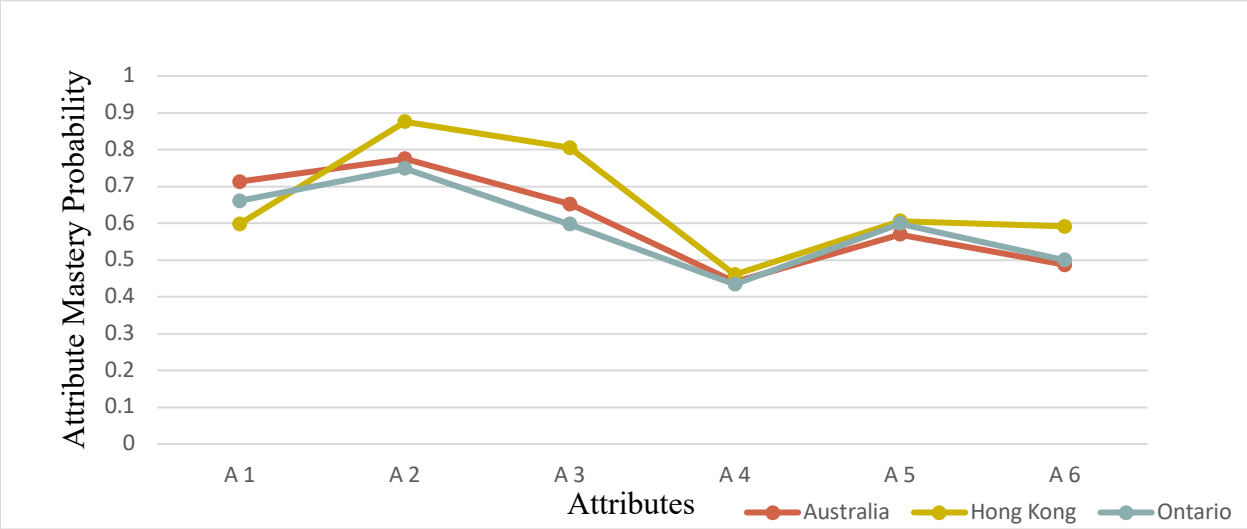


Figure 3. Attribute Mastery Probabilities across Three Jurisdictions

4.6 Latent Class Profiles

The DINA model defines 2^k possible latent classes for each cognitive domain, where k is the number of attributes. In the current study, I have 6 attributes. Thus, there are 64 latent classes. As I noted in Chapter 2, test-takers who are likely to have mastered corresponding attributes are coded as 1, otherwise, it is coded as 0. Test-takers who master all the underlying attributes are categorized to the "111111" latent class. Test-takers who are not considered as masters of any attributes belong to "000000." The latent class "110000" indicates a group of test-takers who are estimated to have mastered the first two attributes presented in the final Q matrix.

Table 16 demonstrates 64 latent class profiles and their attribute mastery pattern probabilities in the three jurisdictions. As is presented in Table 16, for Australia, the latent class "111111" had the highest latent class probability (0.13871), which means about 13.87% of the overall test-takers were estimated to have mastered all attributes. The latent class "111011", to which 8.11% of the test-takers belong, came second. About 8.11% of students could not master Attribute 4 "Explains that simple electrical systems" while they could master all other attributes. The third dominant latent class is "111110" (0.06661), which means that about 6.66% of the test-takers belong to this latent class. This latent class represents mastery of the first five defined attributes. About 6.66% of students did not master Attribute 6 "recognizes that heating an object can increase its temperature and that hot objects can heat up cold objects" while they could master all other attributes. The next seven dominant classes are: "111010" (3.92%), "111000" (3.86%), "111100" (3.718%), "011000" (3.025%), "100000" (2.692%), "110000" (2.583%), and "010000" (2.569%). Finally, about 2.537% of test-takers did not master any attribute.

For Hong Kong, the highest probability is also class "111111" and about 13.19% of test-takers were estimated to have mastered all attributes. As for Australia, the second-highest class

probability of Hong Kong was also a latent class “111011” (0.12016), which means that about 12.02% of test-takers belong to this latent class. About 12.02% of Hong Kong students could not master Attribute 4 “Explains that simple electrical systems,” while they could master all other attributes. The percentage was relatively higher than for Australian students. The third highest class probability is “011010”, with about 7.13% of test-takers possessing this latent class. The next seven dominant classes were: “111101” (5.96%), “011110” (5.47%), “111110” (5.45%), “111010” (5.148%), “010001” (3.823%), “111100” (3.198%), and “000001” (2.831%). In addition, the percentage of Hong Kong test-takers in the latent class that had none of the attributes mastered “000000” was very low (0.665%).

For Ontario, the highest probability is class “111011” and about 8.233% of test-takers were categorized to this class. Test-takers mastered all the attributes except Attribute 4 (i.e., Explain that simple electrical systems, such as a flashlight, require a complete electrical pathway). The second highest class probability of Ontario is the latent class “111111” (0.07489), which means that about 7.489% of test-takers are masters of all the attributes. The third highest class probability is “111010”, with about 6.422% of test-takers not mastering Attribute 4 “Explain that simple electrical systems, such as a flashlight, require a complete electrical pathway” and Attribute 6 “Recognizes that heating an object can increase its temperature and that hot objects can heat up cold objects”. The next seven dominant classes are “111110” (5.80%), “011110” (5.47%), “111101” (4.083%), “111001” (3.34%), “110010” (3.24%), “011100” (3.19%), and “010011” (2.817%). In addition, the percentage of Ontario test-takers that did not master any attributes (i.e., in the latent class “000000”) is slightly higher than that of Australia and Hong Kong.

Table 16. Latent Class Probabilities

Latent Class	Attribute Mastery Pattern	Australia	Hong Kong	Ontario
1	000000	0.02537	0.00665	**0.02789**
2	100000	**0.02692**	0.00364	0.01766
3	010000	**0.02569**	0.00894	0.02090
4	001000	0.00980	0.00617	0.00891
5	000100	0.00473	0.00031	0.01230
6	000010	0.00358	0.00169	0.00436
7	000001	0.01505	**0.02831**	0.01288
8	110000	**0.02583**	0.00454	0.01251
9	101000	0.01309	0.00433	0.00478
10	100100	0.00917	0.00031	0.00807
11	100010	0.00407	0.00216	0.01372
12	100001	0.01435	0.00805	0.00934
13	011000	**0.03025**	0.02243	0.02717
14	010100	0.00782	0.00637	0.01585
15	010010	0.01461	0.00295	0.01116
16	010001	0.01555	**0.03823**	0.00987
17	001100	0.00141	0.00043	0.00569
18	001010	0.00456	0.01021	0.00192
19	001001	0.00311	0.00299	0.00101
20	000110	0.00102	0.00005	0.00187
21	000101	0.00273	0.00136	0.00555
22	000011	0.00281	0.00239	0.01088
23	111000	**0.03855**	0.01535	0.01408
24	110100	0.01451	0.00590	0.00993
25	110010	0.01549	0.00348	**0.03242**
26	110001	0.01399	0.01006	0.00673
27	101100	0.00348	0.00054	0.00315
28	101010	0.01117	0.00763	0.01455
29	101001	0.00620	0.00812	0.01142
30	100110	0.00236	0.00009	0.00438
31	100101	0.00478	0.00069	0.00418
32	100011	0.00490	0.00243	0.01121
33	011100	0.01597	0.02598	**0.03187**

Table 16 Continued

Latent Class	Attribute Mastery Pattern	Australia	Hong Kong	Ontario
34	011010	0.01731	**0.07134**	0.00932
35	011001	0.00979	0.01074	0.00318
36	010110	0.00716	0.00138	0.00854
37	010101	0.00462	0.02724	0.00734
38	010011	0.01170	0.00421	**0.02817**
39	001110	0.00104	0.00046	0.00121
40	001101	0.00044	0.00021	0.00064
41	001011	0.00302	0.00382	0.00244
42	000111	0.00079	0.00007	0.00458
43	111100	**0.03718**	**0.03198**	0.01724
44	111010	**0.03925**	**0.05148**	**0.06422**
45	111001	0.01842	0.02810	**0.03342**
46	110110	0.01532	0.00228	0.01870
47	110101	0.00771	0.01302	0.00525
48	110011	0.01885	0.00393	0.02688
49	101110	0.00516	0.00047	0.00686
50	101101	0.00163	0.00104	0.00753
51	101011	0.02370	0.01845	0.01886
52	100111	0.00280	0.00010	0.00349
53	011110	0.01476	**0.05471**	0.01117
54	011101	0.00513	0.01269	0.00372
55	011011	0.01150	0.02660	0.01193
56	010111	0.00564	0.00196	0.02120
57	001111	0.00068	0.00017	0.00153
58	111110	**0.06661**	**0.05446**	**0.05795**
59	111101	0.01779	**0.05962**	**0.04083**
60	111011	**0.08110**	**0.12016**	**0.08233**
61	110111	0.01835	0.00257	0.01520
62	101111	0.01115	0.00120	0.00898
63	011111	0.00981	0.02087	0.01423
64	111111	**0.13871**	**0.13187**	**0.07489**

Note. The top 10 most frequent mastered latent class patterns for each jurisdiction are in bold.

4.7 Individuals' Performance on the Energy Topic

Test-takers with the same total score on the energy topic could have different attribute mastery patterns. Besides the population attribute mastery patterns, it is also possible to interpret each individual's attribute mastery pattern. I randomly selected two test-takers who received the same total score from the population to take a closer look at each individual's performance on energy. Table 17 presents the attribute mastery patterns for two individual students. From Table 17, we can see that although the two students earned the same score in the energy topic, their mastery patterns differed. Test-taker 1 did not master Attribute 4 and Attribute 5, while test-taker 2 did not master Attribute 3 and Attribute 4. This means that Test-taker 1 could not explain a simple electrical system and failed to relate familiar physical phenomena to the behavior of light (e.g., reflections, rainbows, shadows). Test-taker 2 could not distinguish between substances that are conductors and those that are insulators or explain a simple electrical system. We could create diagnostic performance reports for each test-taker based on individual attribute mastery patterns to inform students' weaknesses and strengths so that we could help each student to promote their performance on energy tasks.

Table 17. Attribute Mastery Pattern for Individual Test-taker

Test-taker	Score in Energy Topic	Attribute Mastery Pattern
1	5	111001
2	5	110011

4.8 Instructional Sensitivity of Selected Items

4.8.1 Instructional Sensitivity of Selected Items without Controlling Student Ability

I also checked the instructional sensitivity of items in the final diagnostic model using logistic regression. I ran each logistic regression based on a separate dataset only including

students' achievement results for each item, and an indicator for whether the corresponding item is covered in each student's national curriculum. Whether the item was covered in each student's national curriculum was treated as the independent variable. Items covered in the national curriculum were coded as 1, and items not covered in the national curriculum were coded as 0. Students' performance on each item (correct or incorrect) was treated as the dependent variable. Six items without any variation in national curriculum coverage across the three jurisdictions, S031077, S031299, S041069, S041070, S041195, and S051179, were not included in the analysis. Item S031077 measures Attribute 3 "Distinguishes between substances that are conductors and those that are insulators". Item S031299 measures Attribute 1"Describes different forms of energy" and Attribute 5 "Relate familiar physical phenomena to the behavior of light". Item S041069, S041070, and S051179 measure Attribute 5 "Relate familiar physical phenomena to the behavior of light". S041195 measures Attribute 4 "Explain simple electrical systems require a complete (unbroken) electrical pathway".

Table 18 presents instructional sensitivity for all items. Item S031273, S030298, S051074, S051121B, and S051121E showed instructional sensitivity. Item S051201 has a negative regression coefficient. Hong Kong students had a much lower proportion of correctness on this item while they performed better on most other questions, so this may cause a negative regression coefficient. The possible reason may be that Hong Kong students took the TIMSS test in Mandarin and the translation of item S051201 may increase this item's difficulty. Item S051188D also has a negative regression coefficient, while the *p*-value of this item is marginal. P values are influenced by the sample size. When the sample size is large, *p*-values are more likely to be significant, which may be the case in this study.

Table 18. Results of the Instructional Sensitivity for All Items without Controlling Student Ability

Item	Regression Coefficient of the Grouping Variable (Log Odds)	*p*
S031273*	0.413	0.000
S031197A	0.187	0.141
S031197B	0.196	0.063
S030298*	0.426	0.000
S041067	0.029	0.766
S051074*	0.464	0.000
S051201	-0.564	0.000
S051121B*	0.313	0.009
S051121C*	0.595	0.000
S051121E*	0.504	0.000
S051188A	-0.212	0.161
S051188B	0.237	0.058
S051188C	-0.349	0.116
S051188D	-0.435	0.022

Note. * Items that were found to show instructional sensitivity.

4.8.2 Instructional Sensitivity of Selected Items after Controlling Student Ability

Some previous studies have controlled for students' ability in exploring items' instructional sensitivity (e.g., D'Agostino et al., 2007; Greer, 1995; Ing, 2018; Li et al., 2017). The rationale is that students' ability prior to instruction would relate to their performance in each item, while the performance of an instructionally sensitive item is expected to increase with effective teaching (Baker, 1994). Thus, I also examined the instructional sensitivity of all items with variation in national curricular coverage, controlling for students' ability. As I noted in section 4.7, we can get the attribute mastery pattern for each student. I calculated the number of attributes each student mastered and treated this as an estimate of their overall competence in the energy domain. Table 19 presents the results for the items showing instructional sensitivity after controlling student's ability. Compared to the results without controlling student ability (see Table 18), items S031273 and S030298 no longer showed instructional sensitivity, while items

S031197A, S031197B, S041067 now appeared to be instructional sensitive. All other items were still showing instructional sensitivity. Items that appeared to be instructional sensitive after controlling students' ability all measured attributes in the first strand of the hypothesized learning progression. We can see that the regression coefficients of items S051188A, S051188B, S051188C, and S051188D were negative after controlling for students' ability at the time of testing. I noted a possible reason that the regression coefficient of whether the curriculum covered the item or not was negative might be due to translation issues. The score of these items may be influenced by other instruction issues, but that cannot be inferred from this study.

Table 19. Results of the Instructional Sensitivity for All Items after Controlling Student Ability

Items	Regression Coefficient of the Grouping Variable (Log Odds)	*p*
S031273	1.202	0.160
S031197A*	1.720	0.003
S031197B*	2.926	0.000
S030298	0.952	0.696
S041067*	0.673	0.000
S051074*	1.017	0.000
S051201	-1.577	0.000
S051121B*	0.240	0.047
S051121C*	0.937	0.000
S051121E*	2.319	0.000
S051188A	-1.366	0.000
S051188B	-0.300	0.035
S051188C	-1.317	0.000
S051188D	-1.299	0.000

Note. * Items that were found to show instructional sensitivity.

CHAPTER 5. DISCUSSION AND CONCLUSION

5.1 Research Question 1

This study hypothesized students' learning progression in energy across four strands: 1) forms of energy; 2) transfer and transformations of energy; 3) dissipation and degradation of energy; and 4) conservation of energy. Since Grade 4 items only related to the first two strands, I only focused on the first two strands (i.e., forms of energy; transfer and transformations of energy) of the learning progression in this study. According to the TIMSS assessment framework and Quebec *Progression of Learning Science and Technology* (Quebec Education Program [QEP], 2009) for elementary school, six attributes were identified in the energy domain. I hypothesized that these six attributes are not necessarily ordered but that the four attributes for the strand "transfer and transformations of energy" are followed by the two attributes from the strand "forms of energy", and attributes are not necessarily consistently ordered across student populations within either strand.

The first research question of this study is to what extent the hypothesized learning progression matches students' observed progression in understanding the energy concept, based on results from the cognitive diagnostic model. This research question could be answered by this study's attribute mastery probability results. According to the results, Attribute 1 "describes different forms of energy (mechanical, electrical, light, chemical, heat, sound, nuclear)" and Attribute 2 "identifies sources of energy (e.g., moving water, the chemical reaction in a battery, sunlight)" from Strand 1 of the hypothesized learning progression had the highest mastery probabilities for Australia and Ontario. Attribute 2 had the highest mastery probability for Hong Kong, while Attribute 1 was ranked as the third-highest mastered attribute for Hong Kong. The highest mastery probability of the two attributes from Strand 1 for Australia and Ontario indicates

that the hypothesized learning progression could be matched to Australia and Ontario students' observed progression in understanding the energy concept using cognitive diagnostic models by detecting the attribute mastery probability. This was consistent with previous research about the learning progression in energy (Lacy et al., 2014; Neumann et al., 2013) that showed the stand "forms of energy" learned before the Strand "transfer and transformations of energy" in the learning progression. However, the hypothesized learning progression could not be matched to Hong Kong students' results. The possible reason leading to differences was that two items (item 031197A and 031197B) assessing Attribute 1 were not covered in Hong Kong's national curriculum according to the TCMA report. "Sources of energy and uses of energy in everyday life" is listed as one of the core learning elements at the Key Stage One in Hong Kong's General Studies for Primary Schools (GS) curriculum document, while forms of energy were not mentioned in Hong Kong's national curriculum. Hong Kong students may be lacking the opportunity to learn Attribute 1. Although according to the TCMA result these two items were also not covered in Ontario's national curriculum, the Ontario curriculum described specifically that "everything that happens is a result of using some form of energy" was a big idea and it expects students could investigate how different types of energy are used in daily life in Grade 1. In addition, these two items were covered in Australia's national curriculum. These may lead to the differences between students' mastery probabilities across countries in describing forms of energy for Attribute 1 and may explain why Hong Kong students have lower mastery probability in Attribute 1.

The content of a country's curriculum (i.e., the intended curriculum) has been shown to affect students' performance (Schmidt et al., 2001; Ramírez, 2006). Schmidt et al. (2005) also found that curricular coherence was the most dominant predictive factor for Grade 1 to Grade 8 students' academic performance in science and mathematics, where the curricular coherence is

defined as curriculum standards sequenced progressively towards the understanding of the deeper structure of each topic both within and across grades. This study reemphasized the importance of the curriculum for students' performance, which is consistent with earlier studies (Ramírez, 2006; Schmidt et al., 2001; Schmidt et al., 2005). Furthermore, this study suggests that the students' observed LPs are dependent on the curriculum they have. Previous learning progression studies have examined results for only one curriculum at a time (e.g., Plummer & Krajcik, 2010), while this study compared diagnostic model results for different curricula. In addition, though previous studies (Gunckel et al., 2012; Liu &Tang, 2004) found differences in LPs for students from different countries or contexts, they did not identify how the curriculum may influence students' LPs specifically. Modeled LPs in this study suggest that curriculum may affect students' performance by modifying their learning trajectory, i.e., students from three jurisdictions likely differed in their learning progressions on the energy topic since some jurisdictions' curricula did not cover Attribute 1 "describes different forms of energy" of the Strand 1 "forms of energy" in the proposed LPs. Besides curriculum, other factors such as classroom instruction and schools' systems may also influence students' LPs. However, this cannot be inferred from this study. In future research, we can look into how classroom level and school level factors relate to students' observed LPs.

LPs can provide a framework to coordinate standards, assessments, and instruction (Alonzo & Gotwals, 2012). The alignment of standards, assessments and instruction could be achieved through LPs. LPs are essential in designing curricula materials that allow learners to develop integrated understandings of key scientific ideas and practices across time (Fortus & Krajcik, 2012). However, currently, not all curricula are designed based on students' LPs. It is common that the curriculum was not built to coherently help learners make connections between ideas

within and among disciplines nor help learners develop an integrated understanding (Fortus & Krajcik, 2012). The development of coherent curriculum materials calls for multiple cycles of design and development, testing and revising the materials, aligning materials, assessments, and teacher support with learning progressions (Fortus & Krajcik, 2012, p.796).

Additionally, I also found that Attribute 4 from Strand 2 was learned latest by students. The mastery probability of the Attribute 4 "Explains that simple electrical systems, such as a flashlight, require a complete electrical pathway" from Strand 2 "transfer and transformation of energy" is the lowest among all the attributes for all the three selected participating jurisdictions: Australia (0.4404), Hong Kong (0.4604), and Ontario (0.4339). In addition, the latent class pattern (111011) was ranked as the first mastered pattern for Ontario (8.2%), and the second for Australia (8.1%) and Hong Kong (12.02%). These results show that no matter which jurisdiction students came from, they performed worse in mastering Attribute 4, and more than half of the students in each jurisdiction failed to acquire Attribute 4. There were mainly two items assessing Attribute 4: item S041195 and item S051074. For item S041195, none of the three jurisdictions' curriculum covered this item. Item S051074 showed instructional sensitivity, which means that the performance of the item was related to whether the item was covered in the curriculum or not. Students performed better on this item if it was covered in the national curriculum than if it was not. However, neither Ontario's nor Australia's curriculum covered this item. When I examined the description of energy for each jurisdiction in the curriculum carefully, "electrical circuits" was highlighted in Australia's curriculum in Grade 6. Similarly, the Ontario curriculum described "simple circuits" in Grade 6. Though Hong Kong reported covering this item in their curriculum, the grade band structure of the national curriculum makes it difficult to identify whether circuits are generally covered in Grade 4, 5 and/or 6. Thus, there was still a large possibility that students

in Grade 4 had not had the opportunity to learn this attribute in school. In addition, students' misconceptions about circuits are common across the world and have been well documented (Moodley & Gaigher, 2019, p. 74). This may explain students' lowest mastery of Attribute 4 in this study. Studies have shown that students have many different misconceptions about electric circuits (e.g., Çepni & Keleş, 2006; Pesman & Eryılmaz, 2010). For instance, Çepni and Keleş (2006) summarized four models used by students that resulted in misunderstanding circuits: a unipolar model; the clashing currents model; the current consumed model, and the scientist model with current conserved. For example, in the unipolar model, students believe that only one cable is enough to complete a circuit, which would hinder their mastery of Attribute 4. Science teachers should get to know different misconceptions that students have in mastering Attribute 4 and utilize these misconceptions to help students to change their misconceptions and enhance their conceptual understanding, for example, by asking students to demonstrate that one cable is not enough to complete a circuit.

5.2 Research Question 2 and 3

The second research question of the study was to examine the similarities and differences in students' knowledge mastery patterns for different countries. The third research question of the study was to explore how the intended curriculum relates to students' understanding of energy across different countries. The second question could be answered from the overall attribute mastery probabilities and attribute mastery pattern profiles obtained during this study. To answer the third question, I examined the TCMA data and the curriculum descriptions of the three jurisdictions. Since these two questions are related, I will discuss them together in this section.

Overall, this study's results showed that Australia and Hong Kong had higher percentages of students mastering all the attributes, while lower percentages of Ontario students mastered all

the attributes and most individual attributes. These indicate that Ontario students perform relatively worse than Australian and Hong Kong students on the energy topic. Among 20 selected items assessing the attributes in the Q matrix, Hong Kong had 15 items that were reported to be covered in their curriculum according to the TCMA data and Australia had 16 items. However, Ontario only had 5 items, many fewer than were covered in the comparison countries. Ontario students' relatively poor performance in energy learning may be caused by their much lesser curriculum exposure to learn these attributes.

This study found some other similarities in students' knowledge mastery patterns across the selected jurisdictions using cognitive diagnostic models. There were high proportions of students in latent class pattern "111010" for all three jurisdictions. Ontario had the highest proportions of students possessing this pattern, then followed by Hong Kong. Australia had a relatively smaller proportion of students. This implicated that most students had weakness in mastering both Attribute 4 and Attribute 6. The latent class pattern "111110" was another dominant knowledge mastery pattern among three jurisdictions: Australia (0.06661), Hong Kong (0.05446), and Ontario (0.05795). In addition, the overall mastery probability of Attribute 6 was the second-lowest next to Attribute 4's. These results indicated that Attribute 6 "Recognize that heating an object can increase its temperature and that hot objects can heat up cold objects" was also difficult for all the participants from three jurisdictions. Students in primary schools always hold some misconceptions about heat and temperature. Students may believe the temperature of an object is related to its physical properties, that is, the object's temperature differs by its material properties (Choi et al., 2001; Erickson & Tiberghien, 1985), and may confuse it with heat (Paik et al., 2007). For instance, some students in primary schools thought that objects of different material in the same room were at different temperatures, and there was a misconception of the students that wood

objects were hotter than metal objects (Erickson & Tiberghien, 1985). These misconceptions about the temperature of objects may lead to students' poor mastery of Attribute 6.

There were also some other differences in students' knowledge mastery patterns. Hong Kong and Ontario students had a much higher proportion of students possessing the latent class pattern "111101" than Australian students, which means that Hong Kong and Ontario students tend to not master Attribute 5 while they could master all the other attributes. This means that Attribute 5 "Relate familiar physical phenomena to the behavior of light" was relatively more difficult for some Hong Kong and Ontario students. However, as to the overall attribute mastery probability, Hong Kong and Ontario students had a slightly larger proportion of students in mastering Attribute 5 than Australian students. Among 64 latent mastery patterns, there were 32 patterns that Attribute 5 was not acquired. Thus, caution is needed in interpreting that more Hong Kong and Ontario students failed to acquire Attribute 5 only from the probability of latent mastery pattern "111101". But we can conclude there was a fairly large number of students from Hong Kong and Ontario who did not master Attribute 5 while they could acquire all other attributes. There are four items measuring Attribute 5 (S031299, S041069, S041070, and S051179). From the TCMA results, I can see all three jurisdictions reported that all four items were covered in their national curriculum, while students' performance differed in mastering this attribute. Thus, I cannot judge how the curriculum may relate to students' understanding of this attribute. I may more closely examine how teachers' implemented curriculum and instruction may relate to students' mastery of Attribute 5 in future research.

5.3 Limitations and Future Directions

Since the current study used existing TIMSS Grade 4 science datasets, I could only detect students' proficiency on attributes of the energy topic that were measured by the test's items, and

only two strands of the hypothesized learning progression could be tested due to the limited number of items that TIMSS administered on the energy topic. In addition, some attributes (e.g., A3 "Distinguishes between substances that are conductors and those that are insulators") conceptually could have been divided into more specific attributes in this study if there had been multiple items measuring these more specific attributes (but there were not). In future research, we could develop an assessment from a cognitive diagnostic model approach to include more attributes, so we can separately detect more abilities and skills of students' energy mastery learning progression (Neumann et al., 2013). For instance, we could add attributes and items related to another two strands that were not included in this study, i.e., dissipation and degradation of energy, and conservation of energy. Since this research only used the DINA model future research could also explore other CDM models (e.g., Attribute Hierarchy Models) to probe students' learning in energy and other science topics.

Second, in the Q matrix validation process, I invited experts in physical science education to review the Q matrix while Grade 4 students were not interviewed to talk through their problem-solving method for each item. In future research, we could also include students' think-aloud process for each item to validate the Q matrix (e.g., Kabiri et al., 2017; Mirzaei et al., 2020). It would be more comprehensive to include both experts' and students' views.

Thirdly, this study only included whether each item was covered in the national curriculum as an independent variable in detecting each item's instructional sensitivity. Although teachers were asked to report when each particular topic assessed in TIMSS (e.g., energy) was taught in the teacher questionnaire, the survey question topics were general and item-level information about whether students received instruction was not available. In order to have a better understanding

about how implemented curriculum and instruction may relate to students' mastery of knowledge in energy, it would be helpful to collect item-level information about instruction in the future.

Fourthly, I only added the sampling weights to improve the estimation accuracy, but I did not take the multilevel structure of the TIMSS data with students nested in classrooms into account due to constraints of the R CDM package. In future research, if we continue to retrofit CDMs to large-scale survey data, we could try to take both their multilevel structure and weights into consideration.

5.4 Conclusion

This study aimed to gain a better understanding of students' learning progression of energy concepts through cognitive diagnostic models. An initial Q matrix was constructed based on the literature related to learning progressions of energy in the physical science domain, and the TIMSS assessment framework. A well-validated Q matrix is crucial to CDM analysis. The initial Q matrix consisted of six attributes and 28 items. Then, the initial Q matrix was reviewed by experts from the physical science education domain. Four items were deleted after the review. After that, the Q matrix was validated by applying the DINA model to two sets of TIMSS data. Item fit indices and overall model fit statistics were computed. This application led to the refinement of the Q matrix, ultimately yielding 20 items that effectively measured six attributes of students' energy understanding. From the result of students' overall attribute mastery probability, it showed that the hypothesized learning progression of learning "forms of energy" first, followed by learning "transfer and transformation of energy" matches students' observed progression in understanding the energy concept using cognitive diagnostic models for Australia and Ontario students, but not for Hong Kong students.

The second aim of the study set out to identify students' knowledge mastery patterns of energy across jurisdictions. For the overall test-takers, the most difficult attribute was to explain that simple electrical systems require a complete (unbroken) electrical pathway, and the attribute identifying sources of energy was found to be the easiest. Students also performed poorly in recognizing that heating an object can increase its temperature and that hot objects can heat up cold objects. There was also a large portion of students having difficulty with both the simple electrical systems and concepts related to heat and temperature.

In conclusion, cognitive diagnostic models are a feasible method to detect students' learning progression since they can be used to identify mastery or non-mastery of fine-grained attributes corresponding to each strand in a learning progression. CDM has the potential to provide insights about fine skills underlying the performance of test-takers. The CDM results may allow learners and teachers to recognize learners' weaknesses, which could guide teachers to adjust their instruction and promote students' learning of energy.

APPENDIX A. LIST OF STUDIES ON LEARNING PROGRESSIONS IN SCIENCE

Study	Learning progression domain/topic	Grade level	Methods/ Models
Alonzo & Steedle, 2007	Force and motion	Grade 8	The standard error of measurement (SEM) analysis; the reliability analysis (Cronbach's α); interview
Black et al. (2011)	Molecular theory of matter	Grade 8	Rasch-based partial credit model
Breslyn et al. (2016)	Sea level rise, a Major Impact of climate change; Causes and mechanisms; Scale and representations Impacts of sea level rise;		
Briggs & Alonzo (2012)	Earth and the solar system	High school students in Iowa	Attribute Hierarchy Method; Classical test theory (CTT)
Chen et al. (2017).	Thermochemistry	Senior high school students	CTT; IRT (Rasch model) as a comparison with rule space model Rule Space Model;
Fumler et al. (2014)	Applying a force and motion learning progression over an extended time span using the force concept inventory	Grades 9–12	Rasch measurement model; Latent class analysis (LCA) (partial credit model)
Gao et al. (2018)	Buoyancy to model conceptual change	Grade 8	CTT; a latent class; rule space model analysis
Hokayem & Gotwals (2016)	Complex ecosystems	Elementary students	Rank correlation method; Interview
Jin et al. (2013)	Carbon-transforming processes in socio-ecological systems.	Grades 4 to 12	Interview

Appendix A Continued

Study	Learning progression domain/topic	Grade level	Methods/ Models
Kizil (2015)	Forces and motion	High school students	Partial Credit Model; Attribute Hierarchy Model (AHM); Generalized Diagnostic Model (GDM)
Merritt & Krajcik (2013)	Building a particle model of matter	Grade 6	Partial Credit Model
Osborne et al. (2016)	Argumentation in science	Grades 6–8	Partial credit model and its generalization the multidimensional random coefficients multinomial logic model (MRCMLM); Think-aloud Interview
Plummer & Krajck (2010)	Celestial motion	Grade 1-3 and grade 8 students	Interview
Plummer & Maynard (2014)	New or novel trajectory" to uncover new ways of describing an LP	Grade 8	Principal components analysis Partial credit model)
Smith et al. (2006)	Matter and the atomic-molecular theory	K through grade 2, 3 through 5, and 6 through 8.	
Songer et al. (2009)	Complex thinking about biodiversity	Grade 6	IRT analysis; Hierarchical Linear Modeling; Growth Model
Stevens et al. (2010)	The nature of matter; Atomic structure and the electrical forces; Multi-dimensional HLP	Middle school students	A systematic design approach, CCD, semi-structured interview
Steedle & Shavelson (2009)	Force and motion	Grade 7 to 12	Latent class analysis, A Bayesian approach to latent class analysis
Suzuki et al. (2015)	Students' Reasoning about Ecosystems (based on Hokayem & Gotwals (2016)'s framework)	Grade 6	Interview
Paik et al (2017)	Buoyancy	Grades 3–12	Partial credit model; interview

APPENDIX B. LIST OF STUDIES ON LEARNING PROGRESSIONS IN ENERGY

Study	Learning progression domain/topic and sequence	Grade level	Methods/ Models
Dawson-Tunik (2006)	Three levels: **Representational systems** ("At the representational systems level, children often provide elaborate observations of the movements of a bouncing ball, including the observation that a bouncing ball bounces lower and lower. ") **Single abstractions** ("At the single abstractions level, the term energy means something "behind" motion—a cause for motion." At this level, students may speak of energy transfer, explaining that the energy of a ball transfers to the floor during a bounce, much as a liquid flows from one location to another. They may also speak of gravity as a force that gradually takes away all of a bouncing ball's energy.") **Abstract mappings** ("At the abstract mappings level, kinetic and potential energy are finally understood as different energy states. Students can describe trans-formations from one energy state to the other, sometimes referring to types of potential energy.")	Grade 9	Interview; Rasch analysis

Appendix B Continued

Study	Learning progression domain/topic and sequence	Grade level	Methods/ Models
Herrmann-Abell et al.(2018)	Three levels (Basic, immediate, advanced), 14 energy ideas based on (6 ideas about the forms of energy, 6 ideas about energy transfer, and two other energy ideas about energy conservation, energy dissipation & degradation)	Grade 4 to 12 grade students 176 University students in Physics	Rasch model Kendall's tau correlation coefficients were calculated to assess the relationship between the difficulty of the items and the items' level on the learning progression (Herrmann-Abell & DeBoer, 2018). ANCOVA was used to perform a cross-sectional analysis of students' performance by grade controlling for gender, ethnicity, and whether or not English was their primary language
Jin & Anderson (2012)	Energy in Socio-Ecological Systems	4th Grade, 7th & 8th Grade, 9th, 10th, & 11th Grade	Interview
Jin & Wei (2014)	Energy in Socio-ecological Systems (sources of energy, nature of energy, and causal reasoning.)	NA	Interview
Lacy et al.(2014)	Energy (forms of energy, transfer and transformations; Dissipation and degradation; conservation)	Grades 3–5	Exploratory interviews
Lee & Liu (2010)	Energy concepts: the Knowledge Integration Perspective (Energy source, Energy transformation items, Energy conservation items)	Middle school students taught by 29 teachers in 12 schools.	Item response theory analysis based on the Rasch partial credit model to validate a learning progression of energy concepts on the knowledge integration construct.

Study	Learning progression domain/topic and sequence	Grade level	Methods/ Models
Liu & McKeough (2005)	Energy sources and energy forms, energy transfer, energy degradation, energy conservation 5 levels: Activity/work, source/form, Transfer, degradation, conservation	Three kinds of different participants: students aged 9 years at the time of testing, typically grades 3 and 4; students enrolled in the two adjacent grades that contain the largest proportion of students of age 13 years at the time of testing, typically grades 7 and 8; students in the final year of their secondary grade, typically grade 12	Rasch partial credit model
Neumann et al. (2013)	Energy: Four hierarchical energy topics: forms, transfer, degradation, and conservation; and each topic has four hierarchical levels/complexity: facts, mappings, relations, and concept	Grades 6, 8, and 10 .	Rasch analysis Analysis of variances (ANOVA); Kendall's t correlation coefficient; student's t-test
Yao et al.(2017)	Ideas about energy (form, transform, Dissipation, Conservation) and four levels of conceptual development (Fact, mapping, relation and systematic) into account.	A total of 4550 students from Grades 8 to 12	Rasch analysis, more specifically, Partial credit Rasch model; ANOVA

REFERENCES

Alonzo, A. C., & Gotwals, A. W. (Eds.). (2012). *Learning progressions in science: Current challenges and future directions.* Springer Science & Business Media.

Alonzo, A. (2012). Eliciting students' responses relative to a learning progression. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science* (pp. 241–254). Rotterdam, The Netherlands: Sense Publishers.

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, *93*(3), 389-421. https://doi.org/10.1002/sce.20303

Aryadoust, V. (2018). A cognitive diagnostic assessment study of the listening test of the Singapore–Cambridge general certificate of education O-level: Application of DINA, DINO, G-DINA, HO-DINA, and RRUM. *International Journal of Listening,* 1-24. https://doi.org/10.1080/10904018.2018.1500915

Australian Curriculum, Assessment and Reporting Authority [ACARA]. (2009*). Shape of the Australian curriculum: Science.* from https://docs.acara.edu.au/resources/Australian_Curriculum_-_Science.pdf

Australian Curriculum, Assessment and Reporting Authority [ACARA] (2020a, June). *F-10 Science Curriculum.* from https://www.australiancurriculum.edu.au/f-10-curriculum/science/key-ideas/

Australian Curriculum, Assessment and Reporting Authority [ACARA] (2020b, June). *Science key ideas.* from https://www.australiancurriculum.edu.au/f-10-curriculum/science

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

Birenbaum, M., Tatsuoka, C., & Yamada, T. (2004). Diagnostic assessment in TIMSS-R: Between-countries and within-country comparisons of eighth graders' mathematics performance. *Studies in Educational Evaluation*, *30*(2), 151-173. https://doi.org/10.1016/j.stueduc.2004.06.

Black, P., Wilson, M. & Yao, S. (2011). Road Maps for Learning: A Guide to the Navigation of Learning Progressions. *Measurement: Interdisciplinary Research & Perspective, 9(*2-3), 71-123, https://doi.org/10.1080/15366367.2011.591654

Bliss, J. & Ogborn, J. (1985). Children's choices of uses of energy. *European Journal of Science Education, 7,* 195-203.

Boylan, C. (2017). Exploring elementary students' understanding of energy and climate change. *International Electronic Journal of Elementary Education, 1*(1), 1-15.

Brennan, R. L. (1972). A generalized upper-lower item discrimination index. *Educational and Psychological Measurement*, *32*(2), 289–303.

Brennan, R. L., & Stolurow, L. M. (1971). *An empirical decision process for formative evaluation: Research Memorandum No. 4*. Cambridge, MA: Harvard University.

Breslyn, W., McGinnis, J. R., McDonald, R. C., & Hestness, E. (2016). Developing a learning progression for sea level rise, a major impact of climate change. *Journal of Research in Science Teaching*, *53*(10), 1471-1499.

Briggs, D. C. & Alonzo, A. C. (2009, June). *The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression.* Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.

Briggs, D. C., & Alonzo, A. C. (2012). The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression. In *Learning progressions in science* (pp. 293-316). Brill Sense.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2*(2), 155– 192.

Brown, J. S., & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science*, *4*(4), 379-426.

Carragher N., Templin J., Jones P., Shulruf B., & Velan G. M. (2019). Diagnostic measurement: Modeling checklists for practitioners (Digital ITEMS Model 04). *Educational Measurement: Issue and Practice,38,* 89-90. Retrieved Dec 18, 2020, from https://ncme.elevate.commpartners.com/

Çepni, S., & Keleş, E. (2006). Turkish students' conceptions about the simple electric circuits. *International Journal of Science and Mathematics Education*, *4*(2), 269-291. https://doi.org/10.1007/s10763-005-9001-z

Chabalengula, V. M., Sanders, M., & Mumba, F. (2012). Diagnosing students' understanding of energy and its related concepts in the biological context. *International Journal of Science and Mathematics Education*, *10*(2), 241-266. https://doi.org/10.1007/s10763-011-9291-2

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29.

Chen., H & Chen, J. (2016) Retrofitting Non-cognitive-diagnostic Reading Assessment Under the Generalized DINA Model Framework, *Language Assessment Quarterly, 13*(3), 218-230. https://doi.org/10.1080/15434303.2016.1210610

Chen, R. F., Eisenkraft, A., Fortus, D., Krajcik, J., Neumann, K., Nordine, J., & Scheff, A. (Eds.). (2014). *Teaching and learning of energy in K-12 education.* New York: Springer.

Chen, F., Zhang, S., Guo, Y., & Xin, T. (2017). Applying the Rule Space Model to develop a learning progression for thermochemistry. *Research in Science Education, 47*, 1357-1378. https://doi.org/10.1007/s11165-016-9553-7

Chen, J. (2012). *Applying item response theory methods to design a learning progression-based science assessment* (Doctoral dissertation). Retrieved from ProQuest LLC. (Accession No. 1-267-18583-X)

Choi, K. M., Lee, Y. S., & Park, Y. S. (2015). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science and Technology Education, 11*(6), 1563-1577. https://doi.org/10.12973/eurasia.2015.1421a

Cox, R. C., & Vargas, J. S. (1966, April). *A comparison of item-selection techniques for norm referenced and criterion referenced tests*. Paper presented at the Annual Conference of the National Council on Measurement in Education, Chicago, IL.

Curriculum Development Council [CDC]. (Hong Kong, China). (2017a). *General Studies for Primary Schools Curriculum Guide:(primary 1-primary 6).* Government Logistics Department.

Curriculum Development Council [CDC]. (Hong Kong, China). (2017b). Science Education Key Learning Area Curriculum Guide (Primary 1– Secondary 6). Retrieved from https://www.edb.gov.hk/attachment/en/curriculumdevelopment/renewal/SE/SE_KLACG_eng_draft_2017_05.pdf

D'Agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional sensitivity of a state's standards-based assessment. *Educational Assessment, 12*(1), 1-22. https://doi.org/10.1080/10627190709336945

Dawson-Tunik, T. L. (2006). Stage-like patterns in the development of conceptions of energy. In X. Liu & W. J. Boone (Eds.), *Applications of Rasch measurement in science education* (pp. 111–136). Maple Grove, USA: JAM Press.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*(4), 595. https://doi.org/10.1007/S11336-008-9063-2

de la Torre, J. (2009). Dina model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130. Retrieved from https://www.jstor.org/stable/40263519

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

Dogan, E., & Tatsuoka, K. (2008). An international comparison using a diagnostic testing model: Turkish students' profile of mathematical skills on TIMSS-R. *Educational Studies in Mathematics*, *68*(3), 263-272. https://doi.org/10.1007/s10649-007-9099-8

Duit, R. (1984). Learning the energy concept in school–empirical results from the Philippines and West Germany. *Physics Education, 19* (2), 59-66.

Duit R. (1986). Der Energiebegriff im Physikunterricht. Kiel: IPN.

Duit, R. (2014). Teaching and learning the physics energy concept. In *Teaching and learning of energy in K–12 education* (pp. 67-85). Springer, Cham

Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education*, *47*(2), 123-182. https://doi.org/10.1080/03057267.2011.604476

Draney, K. (2009, June). *Designing learning progressions with the BEAR assessment system.* In Learning Progressions in Science (LeaPS) Conference, Lowa City, IA, USA.

Effatpanah, F. (2019). Application of cognitive diagnostic models to the listening section of the International English Language Testing System (IELTS). *International Journal of Language Testing*, *9*, 1-28.

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

Erickson, G., & Tiberghien, A. (1985). Heat and temperature. *Children's ideas in science,* 52-84.

Fortus D., & Krajcik J. (2012) Curriculum Coherence and Learning Progressions. In Fraser B., Tobin K., McRobbie C. (eds) *Second International Handbook of Science Education. Springer International Handbooks of Education*, vol 24. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-9041-7_52

Foy, P., Arora, A., & Stanco, G. M. (2013). TIMSS 2011 User Guide for the International Database. *International Association for the Evaluation of Educational Achievement*.

Frankenberg, E., Garces, L. M., & Hopkins, M. (Eds.). (2016). *School integration matters: Research- based strategies to advance equity.* New York, NY: Teachers College Press.

Friedman, J. (2013). *Tools of the trade: when to use those sample weights.* from https://blogs.worldbank.org/impactevaluations/tools-of-the-trade-when-to-use-those-sample-weights

French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement, 47*, 299–317. https://doi.org/10.1111/j.1745-3984.2010.00115.x

Fulmer, G. W., Liang, L. L., & Liu, X. (2014). Applying a force and motion learning progression over an extended time span using the Force Concept Inventory. *International Journal of Science Education, 36*(17), 2918-2936. https://doi.org/10.1080/09500693.2014.939120

Furtak, E. M. (2012). Linking a learning progression for natural selection to teachers' enactment of formative assessment. *Journal of Research in Science Teaching, 49,* 1181–1210. https://doi.org/10.1002/tea.21054

Gao, Y., Zhai, X., Andersson, B., Zeng, P., & Xin, T. (2018). Developing a Learning Progression of Buoyancy to Model Conceptual Change: A Latent Class and Rule Space Model Analysis. *Research in Science Education*, 1-20. https://link.springer.com/article/10.1007/s11165-018-9736-5

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*(2), 1-24. https://doi.org/10.18637/jss.v074.i02

Gilbert, J., & Pope, M. (1986). Small group discussions about conception in science: A case study. *Research in Science and Technological Education. 4*(1), 61-76. https://doi.org/10.1080/0263514860040107

Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement: Interdisciplinary Research and Perspectives*. *6*(4), 263-268. https://doi.org/10.1080/15366360802497762

Gierl, M. J., Wang, C., & Zhou, J. (2008). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT. *Journal of Technology, Learning, and Assessment, 6*(6), n6. Retrieved Dec 18, 2020, from https://files.eric.ed.gov/fulltext/EJ838616.pdf

Gunckel, K. L., Covitt, B. A., Salinas, I., & Anderson, C. W. (2012). A learning progression for water in socio-ecological systems. *Journal of Research in Science Teaching, 49*(7), 843-868. https://doi.org/10.1002/tea.21024

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

Haeusler, C. (2013). Examining the curriculum and assessment framework of the Australian Curriculum: Science. *Curriculum Perspectives*, *33*(1), 15-30. Retrieved Dec 18, 2020, from https://www.researchgate.net/profile/Carole_Haeusler/publication/260165258_Examining_the_curriculum_and_assessment_framework_of_the_Australian_curriculum_Science/links/54adcf300cf24aca1c6f6d53.pdf

Haladyna, T. M. (1974). Effects of different samples on item and test characteristics of criterion-referenced tests. *Journal of Educational Measurement*, *11*(2), 93–99.

Haladyna, T., & Roid, G. (1981). The role of instructional sensitivity in the empirical review of criterion-referenced test items. *Journal of Educational Measurement, 18*(1), 39-53.

Hanson, B. A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics, 23*(3), 244-253. https://doi.org/10.3102/10769986023003244

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128.

Herrmann-Abell, C. F., Hardcastle, J., & DeBoer, G. E. (2018). *Using Rasch to Develop and Validate an Assessment of Students' Progress on the Energy Concept*. Paper presented at the annual meeting of the American Educational Research Association. New York, NY.

Hokayem, H., & Gotwals, A. W. (2016). Early elementary students' understanding of complex ecosystems: A learning progression approach. *Journal of Research in Science Teaching, 53*(10), 1524-1545. https://doi.org/10.1002/tea.21336

Hsu, T. (1971, April). *Empirical data on criterion-referenced tests*. Paper presented at the Annual Conference of the American Educational Research Association, New York.

Hsu, C. L., & Wang, W. C. (2015). Variable-length computerized adaptive testing using the higher order DINA model. *Journal of Educational Measurement*, *52*(2), 125-143. https://doi.org/10.1111/jedm.12069

Husen, T. (Ed.). (1967a). *International Study of Achievement in Mathematics: A comparison of twelve countries* (Vol. I). New York: John Wiley & Sons.

International Association for the Evaluation of Educational Achievement (IEA). (2013). *TIMSS 2011 International Database* [Data file]. https://timssandpirls.bc.edu/timss2011/international-database.html

Im, S., & Park, H. J. (2010). A comparison of US and Korean students' mathematics skills using a cognitive diagnostic testing method: linkage to instruction. *Educational Research and Evaluation*, *16*(3), 287-301. https://doi.org/10.1080/13803611.2010.523294

Ing, M. (2018). What about the "instruction" in instructional sensitivity? Raising a validity issue in research on instructional sensitivity. *Educational and Psychological Measurement, 78*(4), 635–652. https://doi.org/10.1177/0013164417714846

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), 031-73. https://doi.org/10.1177/0265532208097336

Javidanmehr, Z., & Anani Sarab, M. R. (2019). Retrofitting Non-diagnostic Reading Comprehension Assessment: Application of the G-DINA Model to a High Stakes Reading Comprehension Test. *Language Assessment Quarterly, 16*(3), 294-311. https://doi.org/10.1080/15434303.2019.1654479

Jin, H., & Anderson, C. W. (2012). A learning progression for energy in socio-ecological systems. *Journal of Research in Science Teaching*, *49*(9), 1149-1180. https://doi.org/10.1002/tea.21051

Jin, H., Zhan, L., & Anderson, C.W. (2013). Developing a Fine-Grained Learning Progression Framework for Carbon-Transforming Processes. *International Journal of Science Education, 35*(10), 1663-1697, https://doi.org/10.1080/09500693.2013.782453

Jin, H., & Wei, X. (2014). Using ideas from the history of science and linguistics to develop a learning progression for energy in socio-ecological systems. In *Teaching and learning of energy in K–12 education* (pp. 157-173). Springer, Cham. https://doi.org/10.1007/978-3-319-05017-1_9

Johnson, P., & Tymss, P. (2011). The emergence of a learning progression in middle school chemistry. *Journal of Research in Science Teaching*, *48*, 849–877. https://doi.org/10.1002/tea.20433

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272. https://doi.org/10.1177/01466210122032064

Kabiri, M., Ghazi-Tabatabaei, M., Bazargan, A., Shokoohi-Yekta, M., & Kharrazi, K. (2017). 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

Kasai, M. (1997). *Application of the rule space model to the reading comprehension section of the TOEFL* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, Urbana, IL.

Krajcik, J. S., McNeill, K. L., & Reiser, B. J. (2008). Learning-goals-driven design model: Developing curriculum materials that align with national standards and incorporate project-based pedagogy. *Science Education*, *92*, 1–32. https://doi.org/10.1002/sce.20240

Kennedy, C. A., Wison, M., & Draney, K. (2005). Construct map*. Computer program*. Berkeley: Berkeley Evaluations and Assessment Research Center, University of California

Kirkwood, V. & Carr, M. (1988). *Learning in Science Project (Energy) Final Report.* Science Education Research Unit, University of Waikato – Hamilton Teachers' College Hamilton, NZ

Kizil, R. C. (2015). *The marginal edge of learning progressions and modeling: Investigating diagnostic inferences from learning progressions assessment* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global (Accession No. 3743727).

Kosecoff, J. B., & Klein, S. P. (1974, April). *Instructional sensitivity statistics appropriate for objectives-based test items*. Paper presented at the Annual Conference of the National Council on Measurement in Education, Chicago, IL.

Köhn, H. F., & Chiu, C. Y. (2016). A proof of the duality of the DINA model and the DINO model. *Journal of Classification, 33*(2), 171-184. https://doi.org/10.1007/s00357-016-9202-x

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement*, *65*(6), 935–953. https://doi.org/10.1177/0013164405275668

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

Lacy, S., Tobin, R. G., Wiser, M., & Crissman, S. (2014). Looking through the energy lens: a proposed learning progression for energy in grades 3–5. In *Teaching and learning of energy in K–12 education* (pp. 241-265). Springer, Cham.

Lee, Y. S., de la Torre, J., & Park, Y. S. (2012). Relationships between cognitive diagnosis, CTT, and IRT indices: An empirical investigation. *Asia Pacific Education Review, 13*(2), 333-345. https://doi.org/10.1007/s12564-011-9196-3

Lee, H. S., & Liu, O. L. (2010). Assessing learning progression of energy concepts across middle school grades: The knowledge integration perspective. *Science Education*, *94*(4), 665-688. https://doi.org/10.1002/sce.20382

Lee, Y., 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,* 172189, https://doi.org/10.1080/15434300902985108

Lehrer, R., & Schauble, L. (2015). Learning progression: The whole world is not a stage. *Science Education, 99*, 432–437. https://doi.org/10.1002/sce.21168

Leighton, J. P., & Gierl, M. J. (Eds.). (2007). Cognitive diagnostic assessment for education: Theory and applications. New York: Cambridge University Press.

Li, H., Qin, Q., & Lei, P. W. (2017). An examination of the instructional sensitivity of the TIMSS math items: A hierarchical differential item functioning approach. *Educational Assessment, 22*(1), 1–17. https://doi.org/10.1080/10627197.2016.1271702

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

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

Liu, X., & McKeough, A. (2005). Developmental growth in students' concept of energy: Analysis of selected items from the TIMSS database. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *42*(5), 493-517. https://doi.org/10.1002/tea.20060

Liu, X., & Tang, L. (2004). The progression of students' conceptions of energy: A cross-grade, cross-cultural study. *Canadian Journal of Math, Science & Technology Education*, *4*(1), 43-57. https://doi.org/10.1080/14926150409556596

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

Malley, L., Neidorf, T., Arora, A. & Kroeger, T. (2020, June). The Science Curriculum in Primary and Lower Secondary Grades. from http://timssandpirls.bc.edu/timss2015/encyclopedia/countries/united-states/the-science-curriculum-in-primary-and-lower-secondary-grades/

Martin, M., & Mullis, I. (2011). TIMSS and PIRLS achievement scaling methodology. *Methods and procedures in TIMMS and PIRLS. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College*, 1-11. from https://timssandpirls.bc.edu/methods/pdf/TP11_Scaling_Methodology.pdf

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). *Methods and Procedures in TIMSS 2015*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/publications/timss/2015-methods.html

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. Retrieved from https://link.springer.com/content/pdf/10.1007/BF02296272.pdf

Masters, G.N., Adams, R.J., & Wilson, M. (1990). Charting of student progress. In R. Husen & T.N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies* (Vol. 2, suppl., pp. 628–634). Oxford: Pergamon Press.

Masters, G.N., & Wright, B.D. (1997) The Partial Credit Model. In: van der Linden W.J., Hambleton R.K. (Eds.), *Handbook of Modern Item Response Theory.* Springer, New York, NY.

McDonnell, L. M. (1995). Opportunity to learn as a research concept and policy instrument. *Educational Evaluation and Policy Analysis, 17,* 305–322.

Merritt, J., & Krajcik, J. (2013). Learning progression developed to support students in building a particle model of matter. In *Concepts of Matter in Science Education* (pp. 11-45). Springer, Dordrecht.

Mirzaei, A., Vincheh, M. H., & Hashemian, M. (2020). Retrofitting the IELTS reading section with a general cognitive diagnostic model in an Iranian EAP context. *Studies in Educational Evaluation, 64*, 100817. https://doi.org/10.1016/j.stueduc.2019.100817

Mo, Y., Singh, K., & Chang, M. (2013). Opportunity to learn and student engagement: A HLM study on eighth grade science achievement. *Educational Research for Policy and Practice*, *12*(1), 3-19. https://doi.org/10.1007/s10671-011-9126-5

Mohan, L., Chen, J., & Anderson, C. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching, 46*, 675–698.

Mohan, L., & Plummer, J. (2012). Exploring challenges to defining learning progressions. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science* (pp. 139–147). Rotterdam, The Netherlands: Sense Publishers.

Moodley, K., & Gaigher, E. (2019). Teaching electric circuits: Teachers' perceptions and learners' misconceptions. *Research in Science Education*, *49*(1), 73-89. https://doi.org/10.1007/s11165-017-9615-5

Mullis, I., Martin, M., Ruddock, G., Sullivan, C. & Preuschoff, C (2009.) *TIMSS 2011 Assessment Frameworks.* TIMSS & PIRLS International Study Center Lynch School of Education, Boston College. Retrieved from https://timssandpirls.bc.edu/timss2011/downloads/TIMSS2011_Frameworks.pdf

Next Generation Science Standards [NGSS]. (2013). Executive summary. from http://www.nextgenscience.org/sites/ngss/files/Final%20Release%20NGSS%20Front%20Matter%20-%206.17.13%20Update_0.pdf

National Research Council [NRC]. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: The National Academies Press.

National Research Council [NRC]. (2012). *A framework for K-12 science education: practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13165.

Neumann, K., Viering, T., Boone, W. J., & Fischer, H. E. (2013). Towards a learning progression of energy. *Journal of research in science teaching*, *50*(2), 162-188. https://doi.org/10.1002/tea.21061

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods*, *6*(3), 328-362. https://doi.org/10.1177/1094428103254673

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states.* Washington, DC: The National Academies Press.

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

Nordine, J., Krajcik, J., & Fortus, D. (2011). Transforming energy instruction in middle school to support integrated understanding and future learning. *Science Education*, *95*(4), 670-699. https://doi.org/10.1002/sce.20423

Ontario Ministry of Education [OME] (2007). *The Ontario Curriculum, Grades 1–8: Science and Technology,2007.* from http://www.edu.gov.on.ca/eng/Curriculum/elementary/scientec18currb.pdf

Opitz, S. T., Harms, U., Neumann, K., Kowalzik, K., & Frank, A. (2015). Students' energy concepts at the transition between primary and secondary school. *Research in Science Education*, *45*(5), 691-715. https://doi.org/10.1007/s11165-014-9444-8

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S. Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*, *53*(6), 821-846. https://doi.org/10.1002/tea.21316

Paik, S. H., Cho, B. K., & Go, Y. M. (2007). Korean 4-to 11-year-old student conceptions of heat and temperature. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 44*(2), 284-302. https://doi.org/10.1002/tea.20174

Paik, S., Song, G., Kim, S., & Ha, M. (2017). Developing a four-level learning progression and assessment for the concept of buoyancy. *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(8), 4965-4986. https://doi.org/10.12973/eurasia.2017.00976a

Peak, H. (1953). Problems of observation. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences* (pp.243-299). New York: Dryden Press.

Peşman, H., & Eryılmaz, A. (2010). Development of a three-tier test to assess misconceptions about simple electric circuits. *The Journal of educational research*, *103*(3), 208-222. https://doi.org/10.1080/00220670903383002

Plummer, J. D., & Maynard, L. (2014), Building a learning progression for celestial motion: An exploration of students' reasoning about the seasons. *Journal of Research in Science Teaching, 51, 9*02–929. https://doi.org/10.1002/tea.2115

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

Popham, J. W. (1971). Indices of adequacy for criterion-reference test items. In J. W. Popham (Ed.), *Criterion-referenced measurement: An introduction* (pp. 79–98). Englewood Cliffs, NJ: Educational Technology Publications.

Popham, W. J. (2006). *Determining the instructional sensitivity of accountability tests.* Paper presented at the Large-Scale Assessment Conference, San Francisco, California

Plummer, J. D., & Krajcik, J. (2010). Building a learning progression for celestial motion: Elementary levels from an earth-based perspective. *Journal of Research in Science Teaching, 47*(7), 768–787. https://doi.org/10.1002/tea.20355

Plummer, J. D., & Maynard, L. (2014). Building a learning progression for celestial motion: An exploration of students' reasoning about the seasons. *Journal of Research in Science Teaching*, *51*(7), 902-929. https://doi.org/10.1177/0013164410382250

Quebec Education Program (2009). Progression of Learning Science and Technology. Retrieved from https://stpaulelementary.files.wordpress.com/2019/02/5.4.4_scitech_en_progressions-of-learning.pdf

Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modelling using R. *Practical assessment. Research Evaluation, 20*(11), 1–12. from https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1278&context=pare

Ravand, H. & Robitzsch, A. (2018): Cognitive diagnostic model of best choice: a study of reading comprehension, *Educational Psychology*, *38*(10), 1255-1277. https://doi.org/10.1080/01443410.2018.1489524

Ramírez, M. J. (2006). Understanding the low mathematics achievement of Chilean students: A cross-national analysis using TIMSS data. *International Journal of Educational Research*, *45*(3), 102-116. https://doi.org/10.1016/j.ijer.2006.11.005

Robitzsch, A., Kiefer, T., George, A. C., Uenlue, A., & Robitzsch, M. A. (2020). Package 'CDM'. *Handbook of diagnostic classification models.* New York: Springer.

Roseman, J. E., Linn, M. C., & Koppal, M. (2008). Characterizing curriculum coherence. In Y. Kali, J. E. Roseman, M. C. Linn, & M. Koppal (Eds.), *Designing coherent science education: Implications for curriculum, instruction, and policy* New York: Teachers College Press.

Roudabush, G. E. (1974, April). *Item selection for criterion-referenced tests*. Paper presented at the Annual Conference of the American Educational Research Association, New Orleans, LA

Rupp, A., Choi, Y., Gushta, M., Mislevy, R., Thies, M.C., & Bagley, E. (2009, June). *Modeling learning progressions in epistemic games with epistemic network analysis: Principles for data analysis and generation.* Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.

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

Saglam-Arslan, A. & Kurnaz, M. A. (2009). Prospective physics teachers' level of understanding energy, power and force concepts. *Asia-Pacific Forum on Science Learning and Teaching, 10*(1), Article 6.

Schmidt, W. H., C. C. McKnight, R. T. Houang, H. C. Wang, D. E. Wiley, L. S. Cogan, and R. G. Wolfe. (2001). *Why schools matter: A cross-national comparison of curriculum and learning*. San Francisco: Jossey-Bass.

Schmidt, W. H., Cogan, L. S., Houang, R. T., & McKnight, C. C. (2011). Content coverage differences across districts/states: A persisting challenge for US education policy. *American Journal of Education, 117*(3), 399-427. https://doi.org/10.1086/659213

Schmidt, W. H., Wang, H. C., & McKnight, C. C. (2005). Curriculum coherence: An examination of US mathematics and science content standards from an international perspective. *Journal of curriculum studies, 37*(5), 525-559. https://doi.org/10.1080/0022027042000294682

Schwartz, D. L., Bransford, J. D., Sears, D., & Mestre, J. P. (2005). *Efficiency and innovation in transfer* (pp. 1– 51). Greenwich, CT: Information Age.

Shin, N., Stevens, S.Y., Short, H., & Krajcik, J. (2009, June). *Learning progressions to support coherence curricula in instructional material, instruction, and assessment design*. Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). FOCUS ARTICLE: implications of research on children's learning for standards and assessment: a proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research & Perspective, 4*(1-2), 1-98. https://doi.org/10.1080/15366367.2006.9678570

Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *46*(6), 610-631. https://doi.org/10.1002/tea.20313

Stapleton, L. M. (2006). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling*, *13*(1), 28-58. https://doi.org/10.1207/s15328007sem1301_2

Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching, 46*(6)*,* 699-715.

Stevens, F. I. (1993). Applying an opportunity-to-learn conceptual framework to the investigation of the effects of teaching practices via secondary analyses of multiple-case-study summary data. *Journal of Negro Education, 62*, 232–248.

Stevens, S. Y., Delgado, C., & Krajcik, J. S. (2010). Developing a hypothetical multi-dimensional learning progression for the nature of matter. *Journal of Research in Science Teaching, 47*(6), 687–715. https://doi.org/10.1002/tea.20324 Osborne

Stevens, S. Y., Shin, N., & Krajcik, J. S. (2009, June). *Towards a model for the development of an empirically tested learning progression*. In learning progressions in science (LeaPS) conference, Iowa City, IA.

Suzuki, K., Yamaguchi, E., & Hokayem, H. (2015). Learning Progression for Japanese elementary students' reasoning about ecosystems. *Procedia-social and behavioral sciences*, *167*, 79-84.

Swaminathan, H. (1994). Differential item functioning: A discussion. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories of measurement: Problems and Issues* (pp. 171–180). Ottawa, Canada: University of Ottawa.

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

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

Toker, T., & Green, K. (2012). An Application of Cognitive Diagnostic Assessment on TIMMS-2007 8th Grade Mathematics Items. *Online Submission*. from https://files.eric.ed.gov/fulltext/ED543803.pdf

Törnroos, J. (2005). Mathematics textbooks, opportunity to learn and student achievement. *Studies in Educational Evaluation*, *31*(4), 315-327. https://doi.org/10.1016/j.stueduc.2005.11.005

Trumper, R. (1990). Being constructive: an alternative approach to the teaching of the energy concept-part one. *International Journal of Science Education, 12*(4), 343-354, https://doi.org/10.1080/0950069900120402

Trumper, R. (1991). Being constructive: an alternative approach to the teaching of the energy concept-part two. *International Journal of Science Education, 13*(1), 1-10 https://doi.org/10.1080/0950069910130101

Trumper, R. & Gorsky, P. (1993). Learning about energy: The influence of alternative frameworks, cognitive levels, and closed-mindedness. *Journal of Research in Science Teaching, 30*(7), 637-648. https://doi.org/10.1002/tea.3660300704

Wang, J. (1998). Opportunity to learn: The impacts and policy implications. *Educational Evaluation and Policy Analysis, 20*, 137–156.

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

Watts, D. Michael (1983). Some alternative views of energy. *Physics Education, 18* (5), 213-217.

West, P., Rutstein, D. W., Mislevy, R. J., Liu, J., Levy, R., Dicerbo, K. E., ... & Behrens, J. T. (2012). A Bayesian network approach to modeling learning progressions. In *Learning progressions in science* (pp. 255-292). Brill Sense.

Wilson, M. (1990). Measurement of developmental levels. In T. Husen & T.N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies. Supplementary* (Vol. 2, pp. 152–158). Oxford: Pergamon Press.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M. (2008). Cognitive diagnosis using item response models. *Zeitschrift für Psychologie/Journal of Psychology*, *216*(2), 74-88.

Wilson, M. (2009). Measuring Progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*, 716–730. https://doi.org/10.1002/tea.20318

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. Applied *Measurement in Education, 13*(2), 181–208.

Wu, X., Wu, R., Chang, H. H., Kong, Q., & Zhang, Y. (2020). International comparative study on PISA mathematics achievement test based on cognitive diagnostic models. *Frontiers in psychology, 11*, 2230. https://doi.org/10.3389/fpsyg.2020.02230

Yamaguchi, K., & Okada, K. (2018). Comparison among cognitive diagnostic models for the TIMSS 2007 fourth grade mathematics assessment. *PloS one*, *13*(2), e0188691. https://doi.org/10.1371/journal.pone.0188691

Yao, J. X., Guo, Y. Y., & Neumann, K. (2017). Refining a learning progression of energy. *International Journal of Science Education*, *39*(17), 2361-2381. https://doi.org/10.1080/09500693.2017.1381356

Yi, Y. (2012). *Implementing a cognitive diagnostic assessment in an institutional test: a new net-working model in language testing and experiment with a new psychometric model and task type*. (Unpublished doctoral dissertation). University of Illinois at Urbana Champaign, Urbana- Champaign, IL.

Yoon, B., & Resnick, L. B. (1998). *Instructional validity, opportunity to learn, and equity: New standards examinations for the California mathematics renaissance*. Los Angeles, CA: Center for the Study of Evaluation.