YUMSEK, MELTEM, Ph.D. Enhancing Diagnostic Feedback in a K-12 Language Assessment: An Exploration of Diagnostic Classification Models for the Reading Domain. (2020)
Directed by Dr. Micheline Chalhoub-Deville, and Dr. Robert A. Henson. 233 pp.

Various stakeholders, such as educators, and policy makers, appeal for diagnostic feedback and actionable test results. Lack of diagnostic tests, or sufficient diagnosicity in reporting, has encouraged the use of diagnostic classification models (DCMs) for non-diagnostic tests. One particular context in which fine-grained test results are of utmost importance is English language proficiency (ELP) tests. ELP test results are used for critical decisions about English learners (ELs), such as classification, and placement in instructional programs.

This study implemented the DCM methodology to the reading domain of a K-12 ELP test that was taken by 23,942 ELs in grades 6-8, and pursued the viability of DCMs for low-stakes, diagnostic feedback. The study adopted a comprehensive methodology and elaborate research design by incorporating alternative Q-matrices, various diagnostic models, validation strategies for the Q-matrix, and model selection.

The results revealed that a Q-matrix created by experts was theoretically sound and more appropriate for diagnostic results from several aspects. Likewise, a saturated model, such as the log linear cognitive diagnostic model (LCDM), yielded a better fit at the test level. It was also deemed more suitable as the test items were either consistent with a compensatory or conjunctive model. The LCDM proved to be useful for exerting limited diagnostic information. Specifically, the mastery probabilities of individual attributes could be estimated accurately and consistently. Attributes could be separated to some extent, which supports the multidimensionality and makes the second language (L2) reading construct appropriate for the DCM analysis. Most items presented some diagnostic capacity, yet some items were more useful to differentiate masters and non-masters. The ability estimation was generally consistent across the LCDM and IRT models. However, some results, such as the variability of attribute classes, reflected the unidimensional structure of the test. Overall, this study contributes to the representation of L2 reading construct and has some implications for teaching ELs and test development.

ENHANCING DIAGNOSTIC FEEDBACK IN A K-12 LANGUAGE ASSESSMENT:

AN EXPLORATION OF DIAGNOSTIC CLASSIFICATION

MODELS FOR THE READING DOMAIN

by

Meltem Yumsek

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2020

Approved by

____
Committee Co-Chair

____
Committee Co-Chair

APPROVAL PAGE

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

Committee Co-Chair ______________________________

Committee Co-Chair ______________________________

Committee Members ______________________________

____

____
Date of Acceptance by Committee

____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

I made some high stakes decisions five years ago, and changed the course of my life and career to pursue my long-lasting dream of obtaining a Ph.D. degree. My path crossed with amazing ERM faculty, fellows, and professionals in the field who paved the way for my success and made this journey exceptional.

I owe deep and sincere gratitude to my co-chairs Dr. Micheline Chalhoub-Deville and Dr. Robert Henson. This project would not be possible without their unwavering good guidance and profound knowledge. They both worked diligently to help me accomplish my goals, and they were always generous with their time and encouragement. Their extra sympathy and flexibility during the hard times of the Covid-19 outbreak was not unnoticed. I would also like to thank Dr. Henson for introducing me the methodology I applied in my project and meeting me countless times to discuss my questions patiently. Your wisdom and kindness were much appreciated especially at the times when everything was not turning out as expected. You are a wonderful teacher and advisor. I am indebted to Dr. Chalhoub-Deville for the mentorship she provided me during my graduate studies. You opened up exciting opportunities for me to hone my professional skills. Your personal and professional advice brought me where I stand today. I was privileged to be your student and conduct research with you. You are my role model.

I am also grateful to the other members of my dissertation committee, Dr. John Willse and Dr. Mark Chapman who provided invaluable recommendations to improve my project. I am always astonished by the wealth of knowledge Dr. Willse has and his

practical and unique ideas. My methodology was strengthened and became more interesting thanks to you, Dr. Willse. I should also extend my appreciation to you for trusting me in other projects and promoting my operational experience. I cannot thank enough to Dr. Chapman for agreeing to serve in my committee and being an enthusiastic member to assist me in every way he can. It was an honor to collaborate with you in my internship project and I learned a great deal from you professionally. Your support to acquire data for my project is not unforgotten. Thank you for taking the lead and organizing the whole process for my behalf.

I am also thankful to Dr. Gary H. Cook for valuing my project and permitting access to the data for my research and to colleagues in WIDA who lent their support for accessing the data and providing input for my dissertation. I would like to note that the views expressed in this dissertation are those of the author. They do not purport to reflect the opinion or views of The University of Wisconsin or of WIDA. All ACCESS for ELLs test content is the property of the Board of Regents of the University of Wisconsin.

In addition, I am much obliged to the Ministry of National Education of the Republic of Turkey for sponsoring my graduate studies in the U.S. I was able to complete my studies and this project thanks to their financial support.

I would like to acknowledge my friends Dr. James Davis and soon to be Drs. Ramsey Cardwell and Julianne Zemaitis. We shared the joy, stress, success and sometimes failure in this journey. Collaborating with each of you on various projects was a wonderful experience. Thank you for your friendship and support. Julianne deserves an extra thank you for lending a hand whenever I need her and for being my family here.

Last but not least, words are not enough to express my gratitude to my parents and in-laws. They had to make so many sacrifices and endure intense feelings of longing. I would not be able to get through the whole process without their love, support, and prayers. I would also like to share the credit of my success with my husband, Turgay. He was the one inspiring me and boosting my courage to undertake the scholarship. Thank you for your unending genuine sympathy and trying all possible ways to uplift me in the roughest moments. We got this together!

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

"... there is no reason to limit the information provided to candidates to a cryptic message like, Sorry, you fail!" (Luecht, 2003, p. 24)

Large-scale proficiency testing is mandated by law in the U.S. (i.e., Every Student Succeeds Act [ESSA]) and used for accountability purposes as an instrument of educational reform (Chalhoub-Deville, 2016). Schools and teachers are held responsible for student success and they are given rewards and sanctions through accountability testing (Deville & Chalhoub-Deville, 2011). Although testing is primarily used for comparability and evaluation (Hartel, 1999; Hartel & Herman, 2005), the major drive behind it is to provide valuable information for teaching and student learning (Glaser & Nitko, 1970; Hartel, 1999; Shohamy, 1992). Testing mechanisms under accountability also require identifying low performing students and promoting effective supports for them (Hartel & Herman, 2005). Thus, the ultimate goal is to derive student learning, because success can occur as a result of learning.

Large-scale test results can be meaningful for instructional cycles and learning when the information produced is "detailed, relevant, diagnostic, and addresses a variety of dimensions rather than being collapsed to a general score" (Shohamy, 1992, p. 515). That is, diagnostic information is key for large-scale tests to add the desired value to learning and teaching. In this study diagnosis is operationalized based on Luecht's (2003) definitions as "useful feedback information for detecting and evaluating an examinee's strength and weakness" (p.6). Therefore, it has great potential to adapt instruction. The appeal for diagnostic information from large-scale testing has been heard for a long time. Nichols et al. articulated in 1995 that "Today, many want testing to be an integral part of instructional activity, helping to guide teachers and students to the eventual attainment of substantive educational goals" (p.1). This interest continues to increase today. Leighton and Gierl (2007) argue that educational accountability itself exerts pressure for diagnostic results as students are expected to be ready for ambitious knowledge-based workplace. It is of utmost importance to uncover students' strengths and weaknesses to create suitable learning opportunities. This argument is legitimate as such demands are manifested in recent educational laws. Large-scale tests should incorporate individualized information for teachers and principals so that student needs can be diagnosed and addressed (U.S. Department of Education, 2004). ESSA (U.S. Congress, 2015) (i.e., previously No Child Left Behind – NCLB as substantiated by Huff & Goodman, 2007) states that statewide assessments "produce individual student interpretive, descriptive, and diagnostic reports, consistent with clause (iii), regarding achievement on such assessments …"(Section 111 [b][2][B][x]).

In addition to policymakers, other stakeholders such as test developers (Huff & Goodman, 2007) and researchers advocate diagnostic information. For example, Ho (2014) as cited by Wolf et al. (2014) proposes to develop more detailed actionable reporting for English learners for better educational programming. Furthermore, test users demand more detailed test results (Huff & Goodman, 2007; Kunnan & Jang, 2009; Kim et al., 2016; Lopez, 2019). For instance, in their survey with K-12 teachers Huff and

Goodman (2007) have found out that most teachers (about 75%) value individualized information and more than half believe some diagnostic information can be gathered from large-scale assessments. It is apparent that educators look for diagnostic information to inform their instruction and decisions about their students. However, Huff and Goodman remark that teachers' use of large-scale tests for diagnostic purposes is limited due to a variety of reasons. For example, teachers report that test results are usually not comprehensive, clear, beneficial, or diagnostic to report students' strengths or weaknesses. Given the legal enforcement and interest from various stakeholders, more efforts to yield diagnostic and instructionally relevant results from large-scale tests is warranted.

Problem Statement

Predominant measurement models such as classical test theory or item response theory do no inherently provide the level of detail for teaching and learning (de la Torre, 2009). The ability (i.e., test construct) is assumed to be unidimensional and continuous in these models (de Ayala, 2009). Students are assigned a single score, and this score denotes their level of knowledge in relation to the construct (Leighton, 2009). These models only meet the accountability objectives (de la Torre & Minchen, 2014) by ordering and comparing students, but do not reveal information about students' knowledge or processes in specific areas (de la Torre, 2009; de la Torre et al., 2010; de la Torre & Minchen, 2014). When these models are employed, classification decisions or information beyond the total score entails separate *ad hoc* processes. In order to classify students, cut scores are needed. The cuts scores are determined through standard setting activities, which involves gathering educators together to define the characteristics of different proficiency levels and judging test items. However, proficiency level descriptions might also be broad. As the minimal form of diagnostic information, total score is split into subscores (de la Torre, 2009; Haberman, 2008; Kunnan & Jang, 2009). However, they are generally associated with general content areas or language domains as in language assessments (e.g., listening, reading etc.). The subscores might not present adequate detail. Hence, they are deemed superficial for a more complete understanding of student processes and learning (de la Torre, 2009). Furthermore, merely splitting the overall score does not ensure that subscores provide reliable additional information beyond the overall score (Haberman, 2008). Technical standards, policy mandates, and researchers urge to confirm the reliability and value of subscores (American Educational Research Association [AERA] et al., 2014, Standard 2.3; Haberman, 2008; Sinharay & Haberman, 2008; Puhan et al., 2010; Sinharay, 2014; U.S. Department of Education, 2018), which also requires a study of the subscores through different methodologies (e.g., dimensionality, factor analysis, and proportional reduction in mean squared error etc.)

An Alternative Approach: Diagnostic Classification Models

Diagnostic Classification Models (DCMs, Rupp et al., 2010) (a.k.a., Cognitive Diagnosis Models [CDMs]) have emerged as alternative measurement models to purvey finer level of information (de la Torre et al., 2010; de la Torre & Minchen, 2014) and structurally generate student classifications. They have gained popularity in the past decades in light of recent requests (Huff & Goodman, 2007; Liu et al., 2018). DCMs are psychometric models that produce classifications based on multiple constructs (Rupp et al., 2010), and they are “probabilistic, confirmatory, multidimensional, latent-variable models” (Rupp & Templin, 2011, p. 226). Because they are multidimensional models, they allow examining different skills and processes that students engage with. In addition to the pedagogical potential, DCMs have great virtue for understanding the test construct (Rupp et al., 2010). They also bridge different disciplines including cognitive psychology, content domain, psychometrics, and pedagogy (de la Torre et al., 2010; Jang et al., 2015; Leighton & Gierl, 2007; Rupp et al., 2010). For example, domain experts such as teachers, work with psychometricians and/or researchers to identify the underlying traits in a test. Teachers integrate diagnostic information yielded by the psychometric model to their instruction. Because DCMs are multidimensional and categorical, models (de la Torre, 2009; Rupp et al., 2010), they differ from the prevalent psychometric models.

DCMs are intended for exposing knowledge, abilities, processes, and/or strategies underlying the test performance (de la Torre, 2009; Yang & Embretson, 2007). Rather than a unidimensional construct, responses to items relate to multiple unobservable constructs in DCMs. These components are collectively referred to as attributes in DCM terminology (de la Torre et al., 2010; de la Torre & Chiu, 2016; Henson, 2009; Rupp et al., 2010). Because attributes are defined at a finer level, a continuous scale is not practical (de la Torre & Minchen, 2014), and the scales associated with each attribute are categorical. If attributes are assumed to be binary, the model will estimate whether students possess attributes or not. It is also possible to have non-binary attributes to estimate the degree to which a skill is possessed (Rupp et al., 2010).

Each test item measures one or more attributes. The attributes and their relation to each item is mapped in a Q-matrix. The model posits whether each item measures the specified attributes. Based on the status of each skill (e.g., mastery vs. non-mastery), a student is assigned to a class that is called a latent class (de la Torre et al., 2010) or the attribute profile (Rupp et al., 2010). Classification is a “multivariate profile” (Rupp & Templin, 2011, p. 225) and it indicates a student’s mastery status of each of the multiple measured attributes. This integral classification differentiates DCMs from other models.

DCMs generate more detailed information and have more utility, as a general ability is defined at finer levels (i.e., multiple attributes). Attribute level information hints areas an individual student needs to attend to, thus serves as a “roadmap” (Templin & Hoffman, 2013). Educators can plan instruction and develop materials to help individual students, or determine common problems among students at the group level and make more efficient use of instructional time and resources in addressing those problematic areas (de la Torre, 2009; de la Torre & Minchen, 2014; Liu et al., 2018; Sessoms & Henson, 2018). In this way, learning opportunities are maximized. DCMs also yield other useful information such as attribute hierarchies and relations which might be helpful to develop learning trajectories, and guide curriculum decisions (Templin & Bradshaw, 2014).

Early DCM research was concerned with the theoretical aspects and model building, while more recent efforts has focused on the accuracy of the models and promoting their application (Huebner, 2010). For example, DCM studies are undertaken to improve Q-matrices (e.g., Chen, 2017; de la Torre & Chiu, 2016). There have been also efforts to integrate DCMs with other modelling frameworks such as growth modelling (Madison, 2019). Given the amount of scholarship, DCMs have reached a level of sophistication and maturation and their implementation are considered feasible (Sessoms & Henson, 2018).

It must be noted that DCMs are developed to supplement diagnostic assessments, which are based on a cognitive or learning theory (de la Torre & Minchen, 2014). The attributes are rooted in these theories and are known *a priori*. DCMs are used to confirm these theories (i.e., confirmatory nature). Yet, given the lack of such assessments (de la Torre et al., 2010) and sufficient diagnostic information from currently used assessments, DCMs have been implemented for non-diagnostic assessments. Most of these applications pertain to constructs such as math and reading in local as well as international educational assessments like PISA, and TIMSS. (e.g. Chen & Chen, 2016; Lee et al., 2011; Yamaguchi & Okada, 2018). A common area of application for DCMs is the second language (L2) ability. Specifically, there has been considerable work applying DCMs to L2 admission tests (e.g., the Test of English as a Foreign Language Tests [TOEFL]) (e.g., Jang, 2009b; Kim, 2015; von Davier, 2008 etc.). The majority of these studies are replications of the same test data (e.g., TOEFL). Some of the admission tests used in these studies are not necessarily linked to instruction or curriculum. On the other hand, studies investigating DCMs for K-12 language assessments are much needed, yet they are scarce. Diagnostic information is relatively more pertinent in a K-12 context because there is a search for deriving student success and learning through policy initiatives at this level.

The Study Context

The context of the present study is an English Language Proficiency (ELP) assessment administered to K-12 English Learners (ELs) who are not native speakers of English. ELs are mandated to participate in state accountability tests by law. However, ELs' performance is typically below their monolingual English speaker peers (Deville & Chalhoub-Deville, 2011) which might partially stem from their linguistic limitations (Abedi, 2008; Menken, 2008). Therefore, it is of utmost importance to provide detailed feedback about ELs' language development and deficiencies to aid closing the achievement gap. It is worth pointing that ELs might catch up or even outperform their monolingual peers once they gain full language proficiency (e.g., Jang et al., 2013). The educational policy also projects to create equal learning opportunities for this subpopulation and attend to their needs (Deville & Chalhoub-Deville, 2011). Below, the general characteristic of the population and key challenges they face are fleshed out in order to better understand the justification for diagnostic feedback.

ELs, as designated by law, represent a growing subpopulation in the U.S. schooling context. According to 2016 school year data, they constitute roughly five million (10%) of the nationwide student population, and in some states (e.g., California, Nevada, Texas) they make up 17-20% of the students (NCES, 2019). ELs themselves are a heterogenous student body. They belong to various age groups (i.e., young vs. adult), have varying levels of L2 proficiency at every grade level, and come from diverse ethnic, and cultural backgrounds. ELs are plurilingual and they speak 400 different languages (U.S., Department of Education, n.d.). Some ELs might also need special education services (i.e., Individualized Education Plans) due to their disabilities, hence they have both EL and student with disabilities designations. In addition to learning subject content (e.g., math, science, history etc.) at school, they continue to acquire English.

Federal law, ESSA (U.S. Congress, 2015) and previously NCLB intends to maximize ELs' inclusion in statewide assessments. In this regard, once a marginalized group, ELs are included in tests for accountability purposes and have received more attention (Chalhoub-Deville, 2009; Deville & Chalhoub-Deville, 2011, Wolf et al., 2008). Educators are also urged to attend ELs and their needs in instruction to enhance learning opportunities and their success (Chalhoub-Deville, 2009; Deville & Chalhoub-Deville, 2011). One assessment ELs are required to participate in is ELP assessments. The law mandates states to administer ELP assessments annually (ESSA, 2015, Sec.1111(b)(2)(G)) to gauge the language development of ELs and report their progress towards the states' goals. These ELP assessments are required to be based on ELP standards addressing different proficiency levels in four language domains (i.e., reading, listening, speaking, and writing) which are also aligned with challenging academic content standards (ESSA, 2015, Sec.111(b)(1)(F)). Different language use domains such as general academic, discipline specific (e.g., math, social sciences), non-academic or social are addressed in these assessments (Lopez et al., 2016). The purpose of ELP assessments is to ensure that ELs attain appropriate language proficiency. By assuring adequate language proficiency, it is desired that ELs have the same opportunities and receive quality education as their native peers and therefore can show their true ability (Faulkner-Bond & Forte, 2016).

The two most ubiquitous uses of ELP assessments are to identify and monitor ELs' proficiency development (Bailey & Carroll, 2015; Wolf et al., 2008a). Scores obtained from these assessments are used to classify students as ELs or non-ELs (i.e., reclassification after initial administration) (See Wolf et al., 2008b). Results are also the indicator of a state's commitment to support ELs (Bailey & Carroll, 2015). ELP assessments serve other uses as well (Faulkner-Bond & Forte, 2016; Kim et al., 2016; Wolf et al., 2008a). One such use is the placement of ELs in language instruction programs. ELs are entitled to language support services (Faulkner Bond & Forte, 2016; Lopez et al., 2016), which aim to promote their language proficiency and facilitate transitioning to mainstream academic settings without/with minimal support[1]. Information obtained from ELP assessments also informs individual programming decisions such as composition of instructional time (i.e., time spent in mainstream class vs. language support programs), intensity of language program, and supports or accommodations ELs receive (Bailey & Carroll, 2015; Faulkner-Bond & Forte, 2016, Wolf et al., 2008a). For instance, ELs with a very low level of English proficiency might spend the majority of their instructional time in language services. Very few ELs can actually participate in academic school context without some type of language support. In the 2014-2015 school year, 97% of ELs participated in an English instructional program (U.S. Department of Education, n.d.). Therefore, language services have an important

[1] These programs might differ in content and scope. Such that, while some programs highlight only English and focus primarily on academic language, some programs encourage bilingual development of students. U.S. Department of Education and U.S. Department of Justice (2015) define four different programs (a) ESL/ELD, (b) structured English immersion, (c) transitional bilingual/early exit bilingual, (d) dual language/two-way immersion.

role to help students in addressing their weaknesses, reaching the desired levels of English proficiency, and getting fully prepared for the academic instruction.

Another frequent use of ELP assessments is reported to be *diagnosis* (Wolf et al., 2008a), which is also closely associated with language instruction. To date, there have been several efforts investigating diagnostic uses of ELP assessments in specific context of K-12 (Kim et al., 2016; Lopez, 2019; Wolf et al., 2008a). In a survey of state practices, Wolf et al. (2008a) have observed that although states claim to use ELP assessment for diagnosis, score reports lack sufficient diagnostic information that can be valuable for instruction. Similarly, in their survey of educators Kim et al. (2016), Lopez (2019) echo that educators require more fine-grained and actionable information such as strengths and weaknesses of the students to support learning and decision-making in their classrooms. However, they report very limited diagnostic usage of the assessments such as focusing on a language domain in which students are reported to be less competent. Therefore, there is still room to enhance information obtained from ELP assessments. If they allow sufficient detail, the ELP results can be actionable, as states and educators are eager to make use of such information.

The desire for diagnostic information in ELP specific context is also justified by research. Time spent in support programs influence reclassification or dropout rates (e.g., Faulkner-Bond, 2016; Kim, 2011; Menken, 2008; Slama, 2014). In addition to retaining EL status over longer periods of time, the assignment of unsuitable services causes similar problems. The content coverage or complexity in language services is not equal to mainstream classes (Walqui et al., 2010), which might also bring up the opportunity to learn issues. ELs might have limited opportunity to be exposed to high academic standards if they are misclassified (Lopez et al., 2016). By the same token, Lopez et al. argue that language services not addressing ELs' needs or not providing optimum linguistic and academic support might cause greater risk of disengagement and falling behind. The key to provide appropriate supports or place ELs in appropriate programs so that they can attain the language proficiency required to transition to mainstream classes is to first understand their needs (Abedi, 2010). Appropriate instruction matching to ELs' need will optimize academic success and can be possible with diagnostic feedback and fine-grained information. Diagnostic information can also support the design of the support programs. However, it must be noted the implication is not transitioning ELs out of support programs too quickly. Some students might continue to benefit from additional supports. Continuous monitoring of their language proficiency is also necessary because they might fall behind even after exiting EL status (Chalhoub-Deville, 2009). The main objective is to provide suitable services for ELs, and diagnostic feedback can enhance the quality of services.

Purpose and Research Questions

The focus of the present study is the application of DCM methodology to a K-12 ELP assessment, ACCESS for ELLs (Hereafter ACCESS). ACCESS is a large-scale, standardized, summative language assessment measuring the social and academic language development of over 2 million ELs across 40 states and districts in the U.S., every year. In addition to identifying and monitoring the language proficiency of ELs for accountability purposes, the test is also used for programming and instructional decisions such as placement to language support programs and creation of new curricula (CAL, 2017; Kim et al., 2016; WIDA Consortium, n.d.). The test developer also anticipates that scores can support decisions at classroom level (i.e., formative purposes). This study aims to investigate whether useful diagnostic information can be extracted from the assessment to provide more actionable results for educators, and students. It intends to research the feasibility of using an alternative modelling approach, DCMs, to demonstrate the extent to which nuanced information could be generated. The study is exploratory in nature and pursues to produce low stakes diagnostic feedback to aid stakeholders in decision-making related to teaching and learning. As the test was developed under a unidimensional paradigm (i.e., IRT), and not originally intended to be diagnostic, a reverse engineering, also called retrofitting approach (Gierl & Cui, 2008; Liu et al., 2018; Haberman & von Davier, 2007), is undertaken in this study. Specifically, the study proposes to identify the attributes underlying the assessment. It seeks to identify an appropriate model that could output diagnostic information and classifications. The study also intends to evaluate the viability of the DCM approach and comparability of DCM-based classification to ability under the original framework. The research is framed around the following questions:

(1) What are key underlying attributes represented in the ACCESS reading domain in middle grades for more advanced ELs?

(2) What DCM fits the data better?

    a. Does a general or specific restricted model better represent all items in the test?

b. Does a Standard-based or an Expert-defined Q-matrix show better fit?

(3) To what extent is it feasible to obtain diagnostic information using DCM?

a. What is the diagnostic capacity of the test items?

b. To what extent can students be appropriately classified using the model?

Significance

Alderson (2010) asserts that the diagnosing L2 learners' learning and proficiency related strengths and weaknesses is "under problematized and under researched" (p. 97). The present research means to contribute to this under-researched area by exploring the viability of diagnostic feedback through DCMs. Although quite a few examples of DCM implementations for language assessment can be found in the literature, employing DCMs for K-12 language assessments is scant. Jang et al. (2013; 2015) and Aryadoust (2018) have taken some initiative to investigate DCMs at K-12 level. However, Jang et al. consider a Canadian literacy test purported to be used for native and non-native students. Aryadoust, on the other hand, explores an ELP listening assessment for high school students in the Singapore context. To the knowledge of the author, there is a dearth of research at this level, specifically in the context of the U.S. The present study intends to address this gap by focusing on L2 reading domain within a K-12 English proficiency assessment. The diagnostic information also has a stronger motivation in the context of the study. The assessment considered is part of a language learning curriculum, thus everyday teaching. The assessment results can influence program, placement, and potentially the instructional decisions. Therefore, diagnostic feedback is instrumental. The study also contributes to the literature methodologically. The DCM methodology in the study presents the most complete approach by incorporating two competing Q-matrices, and an empirical Q-matrix validation method, and numerous models. The expert panel in the Q-matrix development process also differs from the previous research on language assessments by involving the test developer.

Furthermore, as pointed out by de la Torre (2009) and Yang and Embretson (2007) DCMs are mainly concerned with capturing knowledge, skills, or processes that learners demonstrate, as well as their patterns and relations. The application of DCMs thus requires an elaboration of knowledge, skills, or processes underlying the construct, and contributes to construct representation (Leighton & Gierl, 2007; Liu et al., 2018; Rupp et al., 2010; Yang & Embretson, 2007). To put it differently, the DCM methodology offers a new avenue to shed light on construct and supports construct validity. In fact, earlier conceptualizations of construct validity set the precursor for diagnostic models. Embretson (1983) reframes construct validation with two parts: (1) construct representation, and (2) nomothetic span. Construct representation embraces identification of the underlying theoretical mechanisms or attributes of item responses. It thus entails decomposing tasks into smaller attributes which might include declarative or procedural knowledge, response processes, strategies, standards. In a similar vein, Messick (1989) recognizes the substantive aspect that entails the analysis of response process to be a core component of construct validity. Despite limitations of retrofitting DCMs, item or task decomposition to identify underlying structure can still be valuable as it contributes to our understanding of the construct. In other words, the use of DCMs is "theoretically appealing" (Lee & Luna-Bazaldua, 2019, p. 529). In addition, DCMs can reveal information that conventional methods do not support. DCMs can answer questions like whether attributes characterizing the items show hierarchies or how they interact in relation to performance (i.e., compensatory, non-compensatory). Thus, the study has great potential to support construct representation of the K-12 ELP assessment.

Finally, exploring the validity in DCM applications is underexplored and is nascent in general. Only a few scholars (Borsboom & Mellenberg, 2007; Sessom & Henson, 2018; Yang & Embretson, 2007) discuss what validity and validation constitutes in the context of DCMs. Evidence related to external relations or utility of the DCM results are often overlooked. Only Jang (2009b), and Jang et al. (2015) explore external relations (e.g., self-assessments) and utility. This study explores the relationship between attribute profiles and ELs' proficiency level to understand whether the DCM methodology proves to be an acceptable approach.

Counterarguments and Justification

Retrofitting DCMs has been cautioned due to certain limitations (e.g., Gierl & Cui, 2008). Nevertheless, relatively less effort has been devoted to guiding the development of diagnostic tests (Henson, 2009) with some exceptions (e.g., Bradshaw et al., 2014; Henson & Douglas, 2005; Nichols, 1994). There is a dearth of diagnostic tests to support instruction and learning (de la Torre et al., 2010). Diagnostic tests are currently developed and used for identification and placement of students with cognitive disabilities (Haertel & Herman, 2005). Similarly, the language testing field lacks an applicable theory to build diagnostic assessments and thus they are often neglected (e.g., Alderson, 2005; 2007). Only a few general diagnostic language tests exist (Alderson, 1995, 2007; Alderson et al., 2015, Huhta, 2008; Liu, 2014). Other diagnostic notions in assessing language put teachers and classroom assessment at their center (e.g. dynamic assessment, learning-oriented assessment, diagnostic competence for teachers), and thus frame diagnosis as a teacher competence or responsibility.

It might be argued that teacher assessments (i.e., also sometimes referred to as classroom assessment) are more suitable to diagnose students' instructional needs. Although they are encouraged elsewhere (e.g., U.K., Australia), teacher assessments have received some criticism and are overlooked in the U.S. (Stiggins, 2001). The quality of teacher assessment and their capacity to yield proper diagnostic information have been challenged (Huff & Goodman, 2007). Teachers lack adequate education and assessment literacy to create and validate their own tests to diagnose their students' needs (Alderson, 2005; Stiggins, 2001; Leighton, 2009). Teachers are not allowed sufficient resources to improve their assessment and are rather excused from abiding by the standards applying to large-scale testing in the belief that they cannot achieve the same standards (Stiggins, 2001). For instance, the Standards for Educational and Psychological Testing (AERA et al., 2014) that are the most respected professional guidelines for testing practices, appeal more to measurement professionals but fall short of providing guidance for teachers (Camara & Lane, 2006). Realizing the full impact of teacher assessments, specifically for diagnosis, requires shifting and revisiting the current accountability system (Bennett, 2011).

Although large-scale tests are mainly used for summative purposes to document students' proficiency or achievement, there is some belief that diagnostic information can be gleaned from these assessments (Bennett, 2011; Kunnan & Jang, 2009). For example, large scale language tests have been used for diagnosis "albeit unsystematically" (Alderson et al., 1995, p. 12). Bennett (2011) asserts it might possible to exert limited diagnostic information from large-scale assessments. DCMs can provide a more systematic approach for diagnosis and provide detailed feedback. The dearth of well-designed reliable diagnostic and/or formative assessment justifies exploring DCMs for ELP assessments when coupled with increased interest and belief of the users for this purpose. It is worth stressing that although DCMs are promising with respect to provision of feedback, the use of such information cannot be promised (Templin & Hoffman, 2013). Teachers should also have a clear sense of the construct(s) in order to make suitable interpretations in relation to the performance (Wolf et al., 2016). Further professional training might be needed to support the use of diagnostic feedback.

In addition, designing diagnostic assessments from the ground up is an arduous task not only to due to effort, time, and resources but also the inevitability of abandoning the current practices or policies (Alderson, 2005; Bennett, 2011; Leighton & Gierl, 2007; Liu et al, 2018). Retrofitting can be acceptable for exploratory and research purposes to take advantage of the benefits of the DCM approach (de la Torre & Minchen, 2014). Exploring DCMs in a large-scale assessment context might be reasonable before allocating resources for true diagnostic assessments (Leighton, 2009; Liu et al., 2018).

Some theoretical and empirical evidence supports that language ability and L2 reading are multi-componential (e.g., Davis, 1968; Grabe, 1991; Hudson, 1996; Lumley, 1993; Koda, 2007; Munby, 1978; Weir et al., 1990 etc.). For example, several frameworks and taxonomies, which lay out the components of language ability or reading, exist in the literature (e.g., Bachman & Palmer, 1996; Davis, 1968). The multi-dimensionality of the language ability and L2 reading has motivated DCM applications for language assessments. It might be argued that unidimensional frameworks eliminate dimensionality in the test. However, dimensionality is a fuzzy concept (Sijtsma, 2008; Thissen, 2016). Instead, "... multidimensionality should be regarded as a continuum…" (Blais & Lauirer, 1995, p. 74). Henning (1992) suggests differentiating *psychometric* from *psychological* (i.e. theoretical) dimensionality. DCMs are based on psychological or substantive dimensions (Alderson et al., 2015; Li & Suen, 2013). Language constructs such as L2 reading entail complex cognitive processes (Alderson, 2005). Thus, as with other human traits, they can be theoretically decomposed into finer attributes (Liu, et al., 2018). Similarly, the assessment under scrutiny in the study defines different dimensions, which is elaborated in the methods section.

Organization of the Study

In Chapter 1, I have presented the reasons and sources (i.e., stakeholders) of interest in diagnostic information from large-scale assessments. I have also provided an overview of DCMs that can be used for exploratory purposes to glean diagnostic information from these assessments. In addition, the chapter has introduced the specific context of the study, ELs and ELP assessments and explained the purpose of the study, its rationale, and how it contributes to the current body of research in this area. In the following sections, I focus on DCM applications specifically to language constructs, the methodology that was implemented in the present study, and findings from the analysis. Chapter 2 begins with how diagnosis is perceived in relation to second language learning and assessment. It follows up with an overview of attribute specification, Q-matrix construction, the statistical aspect of various DCMs, and the retrofitting approach. I also focus on model evaluation in DCMs. This section includes a critical review of the suitability of L2 reading construct for DCM applications and research studies that analyze language assessment data with DCMs. Chapter 3 details the assessment, as well as data for the analysis. It also lays out the analysis for each research question by attending to the literature pertinent to these questions. It explains how the attributes and Q-matrix were identified, and how the model fit, and the feasibility of DCM application were evaluated in the study. Chapter 4 presents the salient findings for each research question. Chapter 5 synthesizes the results of the study and discusses them in relation to previous studies. Finally, this chapter explains the limitations of the present study and identifies some areas for future research.

CHAPTER II

REVIEW OF THE LITERATURE

Chapter 1 provided a broad overview of the study. I discussed the interest in diagnostic feedback and the grounds for it for a special population, ELs (e.g., to avoid disengagement issues). DCMs have been applied to large scale assessments in the absence of diagnostics tests for more detailed information. This study purports to implement a DCM methodology to a relatively understudied area, a *K-12* language assessment and is guided by the following research questions:

(1) What are key underlying attributes represented in the ACCESS reading domain in middle grades for more advanced ELs?

(2) What DCM fits the data better?

   a. Does a general or specific restricted model better represent all items in the test?

   b. Does a Standard-based or an Expert-defined Q-matrix show better fit?

(3) To what extent is it feasible to obtain diagnostic information using DCM?

   a. What is the diagnostic capacity of the test items?

   b. To what extent can students be appropriately classified using the model?

This chapter, Chapter 2, elaborates on the diagnostic methodology and presents the literature relevant to different aspects of the study. Specifically, the chapter starts with how diagnosis is conceptualized in language testing. It provides an overview of the attributes and Q-matrix concepts, and how they are derived. The chapter reviews the statistical aspects of DCMs and modelling considerations. Because multiple models were considered in the study, the chapter also presents model evaluation in the context of DCMs. The retrofitting framework guiding this research is presented next. The chapter concludes with a review of dimensionality arguments for L2 reading construct and DCM applications to language assessments.

Diagnosing Language Ability

Diagnosis has been considered an important aspect of language assessments (e.g., Shohamy, 1992). However, the theory of diagnosis for L2 learning and testing is insufficient (Alderson, 2005; Alderson, 2007; Alderson et al., 2015). As Alderson (2005) writes "language testing literature offers very little guidance on how diagnosis might appropriately be conducted, what content diagnostics tests might have, what theoretical basis they might rest on…" (p. 10). A well-established theory of diagnosis in L2, a precise definition of diagnosis, standard diagnostic procedures for L2 language assessments (Alderson et al., 2015), as well as example implementations are needed (Harding et al., 2015). Diagnosis is broadly referred to as identifying students' weaknesses (Alderson et al., 1995) or both strength and weaknesses to design relevant instructional activities (Alderson, 2005; Bachman, 1990; Bachman & Palmer, 2010).

Given this broad definition, English learners' strengths and weaknesses can be identified in multiple ways such as through the use of tests, teacher observations or assessment. Diagnosis in language assessment has been conceptualized in this direction. Some notions such as diagnostic competence for teachers (Edelenbos & Kubanek-German, 2004), dynamic assessment (Lantolf & Poehner, 2004), and learning oriented assessment (Turner & Purpura, 2016) emerged, which focus more on teachers as the provider of the diagnostic feedback. For example, in dynamic assessment instruction and assessment coexist in the same incident (Anton, 2018). The test/task performance is mediated by the teacher who asks questions or give hints (Harding et al., 2015). According Harding et al. there have been efforts to simulate the mediation in computer testing (e.g., Computerized Assessment of Language Proficiency) by providing predefined prompts when students respond incorrectly (e.g., read paragraph X again). Learning-oriented assessment is similar in that it focuses on classroom assessment to bridge learning and teaching. Teachers make inferences based on their assessments regarding skills, knowledge and processes of their learners, and feedback is highlighted in this model (Turner & Purpura, 2016). According to the authors the model intersects disciplines such as language acquisition, assessment, teacher education, pedagogy, and discourse.

These notions bear certain limitations. For example, there is not a standardized assessment design due to diversity of local classroom contexts (Turner & Purpura, 2016). The notions also work on the assumption that teachers are assessment literate. The success of the diagnostic event is very much dependent on the teacher's expertise and training in assessment. Teacher assessments have been criticized in the U.S. context (Stiggins, 2001). It is believed that they do not have adequate training or assessment literacy to develop their own tests and diagnose students' needs (Alderson, 2005; Huff & Goodman, 2007; Leighton, 2009; Stiggins, 2001). It is also difficult to draw comparisons across students in a single classroom setting; therefore, even as small-scale assessments they might not accommodate individual dynamic assessment for each learner (Poehner & Infante, 2016).

Standardized tests can overcome some of these limitations and they have been used for diagnosing language ability. However, there are few diagnostic tests including DIALANG, Diagnostic Language Needs Assessment (DELNA), Diagnostic English Tacking Assessment (DELTA) (Alderson, 1995; 2007; Alderson et al., 2015; Huhta, 2008; Liu, 2014). For example, DIALANG is an internet-based test assessing four domains of language in addition to grammar and vocabulary knowledge, and it is used generally in European context as it is based on the European framework for learning, teaching and assessing languages (i.e., CEFR) (Alderson, 2007). Similarly, DELNA and DELTA are developed for diagnosing the needs of English learners at the higher education context (Liu, 2014). Their use is also limited to their local contexts. These tests tap various subskills such as understanding main ideas, findings specific information, and making inferences (See Harding et al., 2015). According to Alderson (2005, 2007) the lack of diagnostic assessments stem from insufficient focus, theory or guidelines for L2 assessment. He also argues that the practical reasons such available resources and supports challenge devising such tests. Some researchers like Harding et al. (2015)

criticize diagnostic language tests as they tend to isolate skills. In other words, test items are designed to measure one specific subskill which they find inappropriate.

It is also claimed that "virtually any language test has some potential for providing diagnostic information" (Bachman, 1990, p. 60). Alderson et al. (1995) also support that it is possible to use achievement and proficiency tests for limited diagnosis. It has also been a common practice to report domain scores as diagnostic information in language assessments (Kunnan & Jang, 2009). Some researchers employed factor analytic approaches to evaluate whether it is reasonable to report diagnostic subscores from large-scale assessments (e.g., Ina'ami & Koizmi, 2012; Kuriakose, 2011; Sawaki, Stricker & Oranje, 2009). Reporting subscores is considered acceptable in these studies because the general language ability was subsuming the four domains (i.e., adequate fit for bi-factor models). Yet, these subscores are still very broad and inadequate (de la Torre, 2009). Diagnostic results should entail micro as well as macro skills (i.e., macro skills: domain scores such as reading, listening) (Shohamy, 1992). The rise of new methodologies, such as DCMs, have also received attention to extract diagnostic information from large-scale language tests (Alderson et al., 2005). Diagnostic feedback can be provided in a standardized manner to a larger number of examinees with this methodology.

This study operationalizes diagnosis as "useful feedback information for detecting and evaluating an examinee's strengths and weaknesses" (Luecht, 2003, p. 6) in L2 reading ability. It aims to derive such information through the use of a standardized test. In their outline of the features of diagnosis, Alderson (2005, p. 11) comment that diagnosis of language ability must be "low-stakes or no-stakes" and afford detailed feedback for remedial. The purpose of diagnostic information in this context is also to provide teachers more actionable results to tailor their instruction, and it is also low-stakes.

Attribute Specification, Q-Matrix Development, and Validation

Attributes. DCMs require defining multiple skills or knowledge components for a test rather than a broad construct for detailed feedback. These fine level elements represent different dimensions of the tests. These dimensions in the DCM context, refer to substantive or psychological dimensions (Alderson et al., 2015; Li & Suen, 2013). They are also labelled differently such as "latent characteristics, variables, traits, processes, skills, or attributes" (Rupp et al., 2010, p. 49) with attribute being considered a more inclusive term (de la Torre et al., 2010; de la Torre & Chiu, 2016; Henson, 2009). More formally, attributes "characterize test items and they may be interpreted as cognitive processes or skills that are required to perform correctly on test items" (Nichols et al., 1995, p. 11). Attributes can be specified through alternative methods or a combination of methods including expert input, verbal protocols with students (a.k.a., cognitive interviews), eye tracking studies, telemetry data including information about item response times, or test supports (Rupp et al., 2010). However, the use of these methods varies across studies (e.g., how the input from experts is collected), and there are not uniform procedures (Kim, 2015).

In practice, more general concepts or crude attributes are chosen (e.g. fraction, mathematic reasoning, vocabulary knowledge) (Nichols et al., 1995; Rupp et al., 2010).

On the other hand, Buck et al. (1998) contend that in some fields such as second language assessment, more “nuts and bolts” item/task characteristics might be chosen in explaining the test performance (p. 436). The specificity of the attributes is known as the grain size (Rupp et al., 2010). In fact, decisions about the grain size depend on different factors such as relevance, communicative and computational practicality (Nichols et al., 1995; Rupp et al., 2010; Gierl et al., 2009). Although it might possible to specify any relevant attributes depending the construct and its complexity, attributes should support meaningful and useful interpretations (de la Torre & Minchen, 2014; Gierl et al., 2009; Rupp et al., 2010). However, reliable estimation of a large number of attributes might be a challenge, specifically with respect to attribute profiles (Rupp et al., 2010). Thus, it is recommended no to have more than 10 attributes per test (de la Torre & Minchen 2014). Yet, DiBello et al. (2007) does not recommend reducing the number of attributes too much not to lose diagnostic utility. On the other hand, coarse representations might lead to underrepresentation of the construct (Aryadoust, 2018). It is also observed that most DCM studies including simulations specify 4 to 8 attributes, and a longer test length is required for more attributes (Rupp & Templin, 2008). 15-20 items represent the minimum test length for DCM application (de la Torre, 2009). Sessoms and Henson (2018) note the range of the number of the attributes is wider for applied studies. Their review of a DCM applications to real data reveals that an average of 8 attributes are specified with 4 being the minimum and 23 being the maximum number of attributes. The authors also indicate that most applied studies have used large sample sizes of 1,000-2,000 students. In simulation studies, 500 and 1,000 sample size conditions usually represent the lower bound. Madison and Bradshaw (2015) suggest 2,000 respondents are "large enough to avoid confounding results with a deficient sample size" (p. 499). Kunina-Habenicht et al. (2012) also show that the precision of the item parameters estimates in general diagnostic models, especially of the more complex parameters, increase with large samples (e.g., 10,000). However, sample size should also be investigated in relation to different grain size conditions. Skaggs et al. (2016) show the impact of the level of the grain size (e.g., 1, 4, 6, 8, 10 attributes) for varying sample sizes in a simulation study. Their findings demonstrate that 8 attributes at maximum can be estimated with success, but a long test (e.g., 115 items) and a large sample is a perquisite (>1,000). Having more than 8 attributes leads to convergence problems or an instability of the estimates. Bias in estimates is also positively correlated with the number of attributes. Alternatively, in the case of a large number of attributes, existing hierarchies between attributes might overcome these challenges by simplifying estimation (Rupp et al., 2010). Similarly, although there is not an upper limit to the number of attributes per item, the complexity of the item increases with the required number of attributes (i.e., multidimensionality within an item, Bradshaw et al., 2014). It is also not reasonable for an item to measure all attributes theoretically if there are many attributes. Most applied research specifies 1- 3 attributes per item (e.g., Kim, 2015; Ravand, 2016). In short, the desired grain size, instructional relevance, interpretability, and statistical considerations should guide attribute specification (Nichols et al., 1995; Rupp et al., 2010, Gierl et al., 2009).

The Q-Matrix. Each item's relation to each specified attribute is documented in a Q-matrix (Tatsuoka, 1983) to model item responses in a DCM (Rupp et al., 2010). A Q-matrix is an item by attribute matrix, and analogous to factor structure. Specifically, if an item is associated with an attribute, it is coded as 1 and 0 otherwise in the matrix. A Q-matrix where most items are associated with multiple skills shows a complex structure (Madison & Bradshaw, 2015). A Q-matrix is developed with the help of a subject matter experts. Experts can involve test developers, teachers, domain experts, researchers, psychometricians (Kunina-Habenicht et al., 2012; Madison & Bradshaw, 2015). Buck et al. (1998) lay out the steps in this process as: (1) forming an initial draft attribute list, (2) coding items for attributes and creating the incidence matrix, (3) employing the model and refining the coding (i.e., omitting attributes) iteratively based on results from estimation, (4) validating with another form.

The Q-matrix is central to DCMs as it formalizes the structure of attributes and allows confirming this structure (Rupp et al., 2010). In fact, the quality of diagnostic inferences is conditional on the quality of and attribute representation (Sawaki et al. 2009). The design of the Q-matrix ensures the fit and accuracy of the findings (e.g., Kunina-Habenicht et al., 2012; Madison & Bradshaw, 2015; Rupp & Templin, 2008). Madison and Bradshaw (2015) discuss the design consideration based on their evaluation of certain Q-matrix conditions and complexity. They recommend separate attributes at least one time, measuring each attribute with others if separation is not possible, and combining attributes if they are always jointly measured to increase the accuracy. They also caution that increased multidimensionality within an item (Bradshaw et al., 2014) negatively impact estimation capability. However, the authors also acknowledge that general models do not enforce isolation of the attributes for indefinability.

The composition of the expert panel (i.e., background, careers etc.) is also an important consideration in the Q-matrix development because it is a judgmental process. Experts indicate their subjective views about the attributes and their association with items; therefore, a Q-matrix is not always correctly specified (de la Torre & Minchen, 2019; Kang et al., 2019). Research supports that misspecification can pose serious issues and deteriorate the results. For example, Rupp and Templin (2008) evaluate misspecification conditions such as over-specification (replacing 0s to 1s), under-specification (replacing 1s to 0s), or omitting certain attributes for items with complex structure. The findings reveal biased item parameters and reduced classification accuracy especially for those with mis-specified attributes. Similarly, Kunina-Habenicht et al. (2012) obtain inaccurate classifications when misspecifications occur in the Q-matrix.

Validation. To minimize the impact of misspecification, some empirical methods are established to validate a Q-matrix and correct for misspecifications. When there are minor misspecifications, accuracy of attribute probabilities can be retained by correctly specified items and their parameters (de la Torre & Chiu, 2016). However, relying on a validation method can potentially yield more accurate estimates. Some example methods include discrimination index (de la Torre, 2008), an RMSEA based method (Kang et al. 2019), or a residual-based method (Chen, 2017). Despite the availability of various approaches, some of them are restricted to only a particular model (e.g., discrimination index and RMSEA method for the DINA model).

This study incorporated a method developed by de la Torre and Chiu (2016) as an extension of the discrimination index (de la Torre, 2008). The method, which is also known as the general discrimination index (the GDI; de la Torre & Minchen, 2019), computes the discrimination index (ς^2) for each item under the specified model, the GDINA or specific sub-models. Thus, unlike other validation methods, the method can be used to correct misspecifications with a range of models. The GDI denotes the difference in the correct response probabilities of masters and non-masters of the required attributes. In other words, it represents the variance of the probabilities between the two groups (de la Torre & Minchen, 2019). When the attributes specified for the item are correct, then the difference in the probabilities of the two groups should be the maximum. According to de la Torre and Chiu (2016), in this method, a GDI is computed for each possible attribute combinations for an item and then tested against a criterion value. The attribute combination that is larger than the criterion is considered to be the correct vector. When multiple correct q-vectors emerge, the simplest one with fewer attributes should be selected (de la Torre & Minchen, 2019). The authors have also developed a search algorithm to estimate and compare all possible GDIs for all possible q-vectors for an item and correct the misspecifications.

All empirical validation methods require further verification such as theoretical evidence. They are practical to correct misspecifications. Yet, the final decision about a Q-matrix must also be substantively supported (Jang, 2009a; Liu et al., 2018; Ravand, 2016). In addition, these methods should be assessed in light of other conditions, such as missingness. According to Dai, et al. (2018) imputing missing responses instead of treating them as incorrect result in a better performance for discrimination index. According to the authors, high misspecification rates, combined with a large number of attributes, might also lead to the poor performance of validation methods.

DCMs Employed in the Study

According to Rupp and Templin (2011) DCMs "are probabilistic, confirmatory multidimensional latent-variable models with a simple or complex structure" (p. 226). There are numerous models and estimation algorithms (e.g., rule space methodology, attribute hierarchy method). These models very from in several aspects including: attribute interactions (i.e., compensatory vs. non-compensatory), the level or information they yield (i.e., person, item, attribute), parametrization (i.e., general vs. constrained models), or estimation methods (i.e., MCMC, MML) (Rupp et al., 2010). Regarding attribute interactions, attributes can be combined in a compensatory and non-compensatory way to explain test performance. Henson, et al. (2009) elucidate that the conditional relationship between attributes and item responses is influenced by mastery status of the other attributes in non-compensatory models, but not in compensatory models. The authors further divide non-compensatory models as conjunctive and disjunctive models. In conjunctive models, the absence of a required attribute lowers the correct response probability and cannot be compensated by other attributes, while in disjunctive models the mastery of fewer attributes than required might result in high correct response probability. It must be noted that the model choices should incorporate theoretical considerations (von Davier, 2014) and should be systematic and theoretically reasonable. Model comparisons and empirical testing of the constraints can also guide model selection (Henson et al., 2009; Madison & Bradshaw, 2015), especially when model decisions cannot be based on theory.

Before introducing the models that were used for study, I describe the notation. Let $j = 1, \dots, J$ represent the items; $k = 1, \dots, K$ represent the attributes; $i = 1, \dots, I$ represent the individuals; $c = 1, \dots, C$ represent the latent classes. Each items relation to attributes is denoted by q_{jk} such that it equals to 1 if the attribute is measured by the item and 0 if it is not measured. Similarly, the status of attributes for an individual is represented by α_{ik} which equals to 1 for the mastery of an attribute but 0 for the non-mastery. α_c expresses the attribute profile, mastery status of attributes, for a class. There are 2^K classes, as the attributes are binary. For instance, when $K = 2$ there will be four different profiles: (1) having mastered the first attribute only, (2) having mastered the second attribute only, (3) having mastered both attributes, or (4) having mastered neither of the attributes. The length of q and α equals to K (de la Torre & Minchen, 2014). As shown below, the probability of a correct response in a DCM is yielded by q, α, and item parameters (e.g., intercept, main effect etc.). Correctly responding to an item is contingent on the attribute profile (i.e., conditional independence; Rupp & Templin, 2011). Any given model requires at least a moderate test length (e.g., 15-20 items) (de la Torre, 2009) and large sample sizes for reliable estimations (e.g. >1000).

The Log Linear Cognitive Diagnostic Model

The log linear cognitive diagnostic model (LCDM; Henson et al., 2009) is a general model for modelling item responses and attribute interactions. It does not require specifying attribute interactions *a priori* and offers a more flexible approach (Henson et al., 2009; Rupp et al., 2010). By using a logit link, the probability of correct response involves an intercept, main effects, and interaction parameters (Kunina-Habenicht et al., 2012; Rupp et al., 2010). This probability is mathematically:

$$(X_{ij} = 1 \mid a_c) = \frac{exp\ (\lambda_{0,j} + \lambda_j^T\ h\ (\alpha_c, q_j))}{1 + exp\ (\lambda_{0,j} + \lambda_j^T\ h\ (\alpha_c, q_j))} \tag{1}$$

As shown in equation 1, the correct response probability is conditional on class, and therefore is the same for respondents in the same class (Rupp et al., 2010). $\lambda_{0,j}$, the intercept is the probability of a correct response for examinees who have not mastered any of the attributes. λ_j^T represents the main effects and their interactions. The combinations of main effects and their interactions are expressed by $h\ (\alpha_c, q_j)$, which is a mapping function (Kunina-Habenicht et al., 2012). $\lambda_j^T\ h\ (\alpha_c, q_j)$ is:

$$\lambda_j^T\ h\ (\alpha_c, q_j)\ = \sum_{k=1}^{K} \lambda_{j,1,(k)}\, \alpha_{ck} q_{jk} + \sum_{k=1}^{K-1} \sum_{k'>1}^{K} \lambda_{j,2,(k,k')}\, \alpha_{ck}\ \alpha_{ck'} q_{jk}\ q_{jk'} + \cdots \tag{2}$$

where the first term is the main effect (i.e., represented by the subscript 1). For instance, there will be two $\lambda_{j,1,(k)}$ if the item measures two attributes. Main affects show the increase in the correct response probability for mastering the given attribute (Madison & Bradshaw, 2015). The second term is the two-way interaction (i.e., the subscript 2). k' denotes the second attribute in this case. Higher order interactions such as three-way interactions can also be specified (i.e., "…" in the equation). Similarly, the interaction term indicates the increase in the correct response probability for possessing all required attributes (Madison & Bradshaw, 2015). The LCDM is a saturated model as all possible effects and interactions are estimated. As a result, it has higher parametrization (Rupp et al., 2010). The LCDM also relates to two general models. It is a special case of general diagnostic model (GDM) (Rupp et al., 2010; von Davier, 2014b) and equivalent to the generalized DINA (G-DINA) with logistic link function, which allows other links (identity and log) (de la Torre, 2011; Ma, 2019).

The main advantage of the LCDM lies in the fact that core models can be reformulated with additional constraints on the intercepts and main effects (i.e. λs) (Henson et al., 2009; Rupp et al., 2010). In other words, it subsumes the core models. Because the same estimation procedure is applied to submodels (Rupp et al., 2010), model and item parameters become common across the models and can be compared. New models can also be defined by additional constraints or parameters (Henson et al., 2009; Rupp et al., 2010). The LCDM might also improve the fit in retrofitting studies as it allows the relationship between attributes and the observed outcome to be items specific (Rupp et al., 2010). Therefore, it is possible different submodels apply to different items rather than one specific model for all items (Rupp & Templin, 2011). However, more parsimonious specific models are also desired, and they also ensure easier interpretation (Lee & Luna-Bazaldua, 2019). General models, along with specific models, are also applied to language constructs possibly due to a lack of comprehensive theory about attribute relations, as discussed in the next section. Therefore, five restricted models that are also frequently used in retrofitting (i.e., the DINA, R-RUM, DINO, C-

RUM, HO-DINA) were implemented in the study. Next, I present these models and how they are formulated within the LCDM framework.

Common Restricted Models

The deterministic-input, noisy-and-gate ([DINA], Haertel, 1989; Junker, & Sijtsma, 2001) is a conjunctive model that requires the mastery of all attributes specified for an item in order for an individual to have a high correct response probability for that item. In the DINA, it is assumed that there are two groups of examinees: (1) those who have mastered all attributes measured by the item, and (2) those who have not mastered at least one of those attributes the item measures (Henson et al., 2009; Rupp e t al., 2010). Not mastering a required attribute is the same as missing all (de la Torre & Minchen, 2014). There are only two probabilities for each item. Given these two groups and the probabilistic quality (Rupp et al., 2010), the DINA includes a slip (s_j) and a guess (g_j) parameter. The slip parameter is the probability of an incorrect response when all required attributes are mastered, whereas the guess parameter represents the probability of a correct response despite non-mastery of at least one of the required attributes (Henson et al., 2009). The probability of correct response in the DINA model is mathematically defined as:

$$P(X_{ij} = 1|\xi_{ij} = (1 - s_j)^{\xi_{ij}} g_j{}^{(1-\xi_{ij})}) \tag{3}$$

where ξ_{ij} represents the mastery status of the attributes. It equals to 1 when all attributes are mastered, and the equation reduces to $1 - s_j$. In other words, for group 1 which are the masters of all attributes, probability of not slipping represent the correct response probability (Rupp et al., 2010). ξ_{ij} equals to 0 when one of the required attributes is not mastered, and the equation reduces to g_j. The probability of guessing represents the correct response probability for the second, missing at least one attribute, becomes (Rupp et al., 2010).

Because the LCDM defines correct response probability in terms of intercepts, main effects and interactions, the DINA model is obtained in the LCDM framework by constraining all main effects and lower order interactions to 0 (e.g., $\lambda_{j,1,(k)} = 0$). Only higher order interaction ($\lambda_{j,2,(k,k')}$if there are two attributes) is estimated, and it is positive. This higher order interaction is associated with $1 - s_j$ in the regular the DINA representation. The intercept λ_{0j} behaves like g_j and represents the probability of a correct response without knowing any of the attributes (i.e., as missing one attribute or all is the same).

Despite being the simplest model, the DINA is among the most restricted models (de la Torre, 2011; de la Torre & Minchen, 2014; Henson et al., 2009; Rupp et al., 2010). Splitting examinees into two groups might be problematic in some situations. First of all, the group in which examinees lack something clusters a lot of individuals together. The DINA pretends examinees are always in the non-mastery group. In addition, it might be useful to differentiate between the attributes that individuals lack.

A conjunctive model allowing attributes to contribute differently is the non-compensatory reparametrized unified model ([NC-RUM], Dibello et al., 1995; Hartz, 2002). This model is repeatedly applied to language assessment data (Stout et al., 2019), and is thus the most common model for language constructs. The model has a full and reduced version (aka., R-RUM). Because the reduced model is more prevalent in application due to its simplicity (Rupp et al., 2010), it was implemented in this study. Known also as the Fusion model, this model relaxes the equality for missing one or all attributes and addresses the limitation present in the DINA (Henson et al., 2009; Rupp et al., 2010). Because it is a conjunctive model, the correct response probability is lower when an attribute is not mastered, however, the probability is different in the absence of different attributes. Also, as more attributes are mastered, the correct response probability increases substantially. There are two parameters in the model. π^* is the correct response probability for examinees possessing all of the specified attributes. The other parameter r^*, decreases the probability for not possessing a required attribute and behaves as the penalty parameter (Rupp et al., 2010). The model includes one penalty parameter per measured attribute per item. Equation 4 shows how the correct response probability is represented as a function of these parameters:

$$P(X_{ij} = 1|\alpha_i) = \pi_j^* \prod_{k=1}^{K} {r^*}_{jk}^{q_{jk}(1-\alpha_{ik})} \tag{4}$$

Given an item measures an attribute ($q_{jk} = 1$) and the attribute is mastered in the class ($\alpha_{ik} = 1$), the probability of a correct response equals to π^*. However, if the required attribute is not mastered ($\alpha_{ik} = 0$), r^* decreases the probability of a correct response.

The NC-RUM is the most complex sub-model to represent within the LCDM (Henson et al., 2009). As the correct response probability changes with mastery of each attribute, Henson et al. have shown that the LCDM can be used to fit the NC-RUM by estimating intercepts and main effects. The main effects are positive, and all interactions are defined as a function of these main effects and intercepts. The correct response probability increases with the mastery of additional required attributes. However, the interaction is greater than just the sum of the main effects (Rupp et al., 2010). This increase in mastering additional attributes also captures the penalty concept in the regular representation. To put it another way, it is the inverse (Henson et al., 2009). The reader is recommended to see the original paper for a detailed description of the interaction term.

The deterministic-input, noisy-or-gate ([DINO], Templin & Henson, 2006) is similar to the DINA, yet it is a disjunctive model. Therefore, mastering any one of the required attributes is sufficient for correctly responding to an item. In other words, the probability of a correct response is the same for knowing one or more attributes. There are two groups of examines (i.e., those who mastered at least one attribute vs. those who mastered none) and two parameters (i.e., slip and guess) (Henson et al., 2009). However, they are defined differently. In the DINO, the slip is the *incorrect* response probability if at least one of the required attributes is mastered, whereas the guess is the *correct* probability if none of the required attributes is mastered (Rupp et al., 2010). Equation 5 shows the mathematical representation of the model:

$$(X_{ij} = 1|\omega_{ij}) = (1 - s_j)^{\omega_{ij}} g_j^{(1-\omega_{ij})} \tag{5}$$

As seen in the equation, the representation of the DINO is almost the same as the DINA except for ω_{ij} which shows the mastery status of an attribute. This change in the notation is due to change in the meaning of the parameters. If $\omega_{ij} = 1$ and any one of the attributes is mastered, the equation reduces to $1 - s_j$. Thus, not slipping denotes the correct response probability for the group which mastered one required attribute (Rupp et al., 2010). Similarly, if $\omega_{ij} = 0$ and all attributes are missed, the equation reduces to g_j.

In the LCDM, the DINO is modelled by setting the main effects and interactions to be equal with different signs. For instance, main effects are positive, and the two-way interaction is negative ($\lambda_{j,1,(k)} \ = -\lambda_{j,2,(k,k')}$). Thus, interaction does not increase the correct response probability. In other words, there is no additional advantage of knowing both attributes (Henson et al., 2009; Rupp et al., 2010). Interaction terms is associated with $1 - s_j$. Whereas, $\lambda_{j,0}$ is associated with guess parameter as it represents not mastering required attributes but correctly responding.

The DINO suffers from a similar limitation the DINA does. It does not differentiate individuals mastering different attributes endorsed by the item. To account for this restriction and allowing attributes to contribute differently (Rupp et al., 2010), the compensatory reparametrized unified model ([C-RUM], Hartz, 2002) is used. In the C-RUM, mastery of one of the required attributes is sufficient for a high correct response probability. The correct response probability changes depending on which attribute is mastered. However, the probability is not dependent on mastery or non-mastery of an additional required attributes (i.e., no equality constraint for main and interaction effects). The C-RUM includes an intercept ($\lambda_{j,0}$) and a slope ($\lambda_{j,1,(k)}$) parameters shown in equation 6. The intercept represents ($\lambda_{j,0}$) the correct response probability when no required attribute is mastered. The slope ($\lambda_{j,1,(k)}$) shows the increase in the correct response probability for mastery of an attribute. There is one slope per attribute per item.

$$P(X_{ij} = 1|\alpha_i) = \frac{exp(\lambda_{j,0} + \sum_{k=1}^{K} \lambda_{j,1,(k)} \alpha_{ik} q_{ik})}{1 + exp(\lambda_{j,0} + \sum_{k=1}^{K} \lambda_{j,1,(k)} \alpha_{ik} q_{ik})} \quad (6)$$

If an attribute is measured by an item ($q_{ik} = 1$), and the attribute is mastered ($\alpha_{ik} = 1$), the probability increases. However, if the attribute is not mastered ($\alpha_{ik} = 0$), there is neither contribution nor penalty (Rupp et al., 2010).

The C-RUM in the LCDM is obtained through constraining interaction terms to be 0 ($\lambda_{j,2,(k,k')} = 0$). Only the main effects are estimated. In this way, the correct response probability is a function of main effects. The C-RUM is conceptually very similar to the LCDM (Henson et al., 2009; Rupp et al., 2010).

The four models described above are used to model the relationship between attributes and responses to items. So, they model the probability of a correct response. However, it is also possible to model the attribute space. In modelling attribute space, a hierarchical attribute structure is assumed. One such model is the higher order DINA ([HO-DINA], de la Torre & Douglas, 2004). In fact, there are two parts to higher order models (Liu et al., 2018). More specifically, for language related constructs such as L2 reading, it is not uncommon to think a more general continuous trait such as a general language ability underlies the attributes (de la Torre & Douglas, 2004). This general ability can be thought as the θ in the IRT. It represents the higher order part in the model. This part assumes this common general ability influences the mastery of attributes thus models the relationship between them. The second part is actually a specific model. In the case of the HO-DINA example, the specific model is the DINA where the relationship between item responses and attributes are modelled using a conjunctive rule described above. The representation of the HO-DINA is shown in equation 7:

$$P(\alpha_{ik}|\theta_i) = \frac{exp(\lambda_{0k} + \lambda'_k\theta)}{1 + exp(\lambda_{0k} + \lambda'_k\theta)} \tag{7}$$

Note that the model is different from the representation of models for response and it includes intercept and slope parameters as well as a parameter for general ability. Although any specific model can be theoretically defined using the higher order structure, the software might be limited to estimate higher order models. In this study, only the HO-DINA was used.

The HO-DINA has the potential, especially in the retrofitting context because the assessments are designed under a unidimensional framework (de la Torre & Douglas, 2004; Liu et al., 2018). As mentioned, there might be a general ability subsuming the specific attributes. It also addresses the issue of large number of attributes and parameters (de la Torre & Douglas, 2004; Bolt, 2019). Bolt also advocates this model as it combines classification and ability estimation.

Model Evaluation in DCMs

Evaluation of the model fit is another crucial step in the DCM approach. However, current model fit research for DCMs is scarce (Chen et al., 2013; Hu et al., 2016; Lei & Li, 2016). Model fit and fit indices are still elusive and the search for the best evaluation methods is ongoing (Rupp et al., 2010).

Current model inspection approaches involve a comparison of a set of models that is relative fit, and comparison of the observed versus estimated values to determine the suitability of a specific model for the data, which is absolute fit. Relative fit might help eliminating models from further analysis when the correct model is not known based on theory, yet it is not adequate by itself as the chosen model might still fit poorly (Rupp et al., 2010). There are numerous relative and absolute fit indices, and several researchers have investigated the performance of these indices using simulations in the DCM context.

Akaike's information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC, Schwarz, 1976) are frequently used as relative fit indices in model evaluation in measurement. They are also applied to DCMs. The logic behind the two indices is the same as the number of parameters (p) functions as the correction term for Log likelihood (LL), which is obtained from model estimation. $AIC = -2LL + 2p$ and $BIC = -2LL + pln(N)$. As the number of parameters increases, the correction is bigger. BIC takes sample size into consideration in calculating the correction. The lower AIC and BIC are the better is the fit. Empirical research shows AIC to outperform and select more complex models while BIC picks a reduced model (Hu et al., 2016; Kunina-Habenicht et al., 2012; Lei & Li, 2016). However, Chen et al. (2013) mention a better performance for BIC and recommend it, although they agree that AIC chooses saturated models. Both indices are affected by misspecifications in the Q-matrix and sample size (Kunina-Habenicht et al., 2012; Lei & Li, 2016).

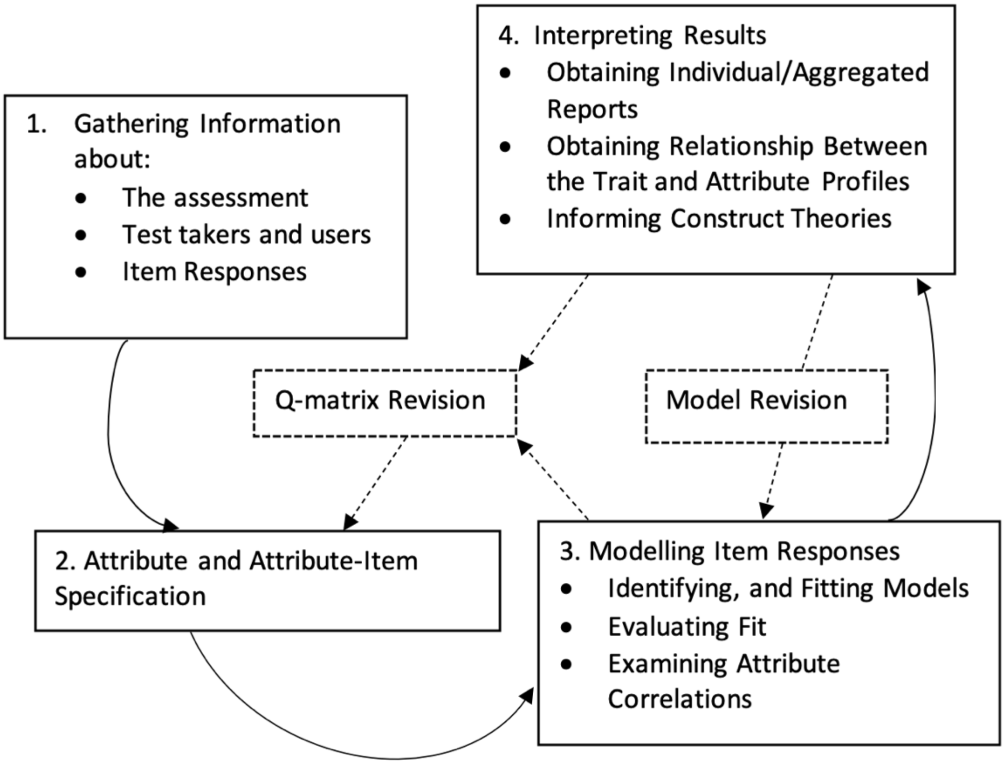
Compared to relative fit statistics, a wider variety of absolute fit indices are described in the literature. Absolute fit indices are at the item level and usually concerned with residuals (Lei & Li, 2016). Some of them entail full information statistics (e.g., χ^2 and G), or summary statistics such as averages, and are characterized as limited information fitness of good statistics (Rupp et al., 2010). Kunina-Habenicht et al. (2012) evaluate von Davier's root mean squared error of approximation (RMSEA) and mean absolute difference (MAD; Henson et al., 2008). According to the authors, RMSEA is the squared difference of observed and predicted response probabilities and it is weighted by the proportion of a class. MAD is similar and represents the average absolute difference of those probabilities, but it assumes equal weights. The fit gets better as both indices get closer to 0. The authors articulate sample sizes and the number of attributes significantly affect how these indices perform. RMSEA slightly outperforms MAD due to differential weights. Very large sample sizes (10,000 and above) increase the power of the indices. The indices are also in accordance with relative fit indices. The authors conclude that the indices are informative and help identify Q-matrix misspecifications, yet they are still limited to the choose the correct model and Q-matrix at the same time. Chen et al. (2013) define and compare three other indices: difference between observed and estimated "proportion of correct of items (p), Fisher transformed item correlations (r), and log-odds ratio (l) of item pairs" (p.126). They conclude the r and l are (i.e., bivariate) are similar and better than p (i.e., univariate). Similar to MAD and RMSEA, these indices show poor fit when both the model and Q-matrix are mis-specified. They favor saturated models. Hu et al. (2016) include both RMSEA and r used in Chen et al. (2013). The performance of indices is similarly in their study. They fail to identify a correct model when a Q-matrix is mis-specified, with the exception of a saturated model. The authors also warn that RMSEA can be misleading for an over-specified Q-matrix. Lei and Li (2016) inspect (1) average χ^2 (Chen & Thissen, 1997), (2) the mean absolute deviation of Q3 (MADQ3;

Yen, 1984), (3) the mean absolute deviation item residual covariances (MADres; MacDonald & Mok, 1995), and (4) mean absolute deviation of item correlations (MADcor; DiBello et al., 2007). They find average χ^2 as the most effective index for large samples (1,000). MADcor, MADres show good performance too. In summary, these studies show that limited information indices are trending and have some potential (Lei and Li, 2016; Rupp et al., 2010). Relying on multiple indices may result in more sound decisions, although absolute fit indices are more commonly used for Q-matrix misspecification. They are also parallel to relative fit indices and can be used jointly for more informed model selection (Chen et al., 2013; Kunina-Habenicht et al., 2012; Lei & Li, 2016). Among absolute fit indices, bivariate indices perform better (Chen et al., 2013; Lei & Li, 2016; Rupp et al., 2010).

The Retrofitting Framework

Retrofitting is a back-engineering approach. It entails reconsidering an extant test in a different paradigm and employ a different statistical tool, a diagnostic model to analyze responses (Gierl & Cui, 208; Lee & Luna-Bazaldua, 2019; Roussos et al., 2007). Stout et al. (2019) differentiate between retrofitting a test that is intended to be multidimensional but not based on a DCM versus one that is designed to be unidimensional. The latter describes the approach in this study. Retrofitting has been a common practice, to supplement simulation studies (Liu et al., 2018; Sessoms & Henson, 2018) or stand-alone applications, to explore different aspects of DCMs. Yet, implementations show some variations and not all of them adhere to a standard procedure (Lee & Luna-Bazaldua, 2019). Liu et al. (2018) describe the stages of the process to guide retrofitting research and increase the success of this “suboptimal” approach (p. 361). Their suggested framework (Figure 1) also inform this study. In this framework, the first step entails getting familiar with test, users, and respondents and obtaining information such as design structure, psychometric properties (e.g., dimensionality, item parameters, blueprints etc.). Such information appraises the legitimacy of the DCM, can inform attributes and model choice. The second step is the attribute specification and mapping. However, dimensions of the test might not be readily available, which requires a review and decomposition of items to generate them retrospectively (Gierl & Cui, 2008), and the Q-matrix represent the multiple dimensions (Haberman & von Davier, 2007). Empirical methods can be undertaken for the Q-matrix validation. Design recommendations (e.g., combining, retaining attributes, Liu et al, 2018; Madison & Bradshaw, 2015). can also help modifying the Q-matrix. Hartz et al. (2002) recommend retaining an attribute in the Q-matrix if it is measured at least by 3 items (in Kim, 2015). In terms of the model choice (the third step), a general model, as well as the specific models, are endorsed. The model suitable for the test is selected based on fit statistics. Finally, if adequate fit is observed then individual, aggregate, and attribute level results can be reported and interpreted (e.g., attribute distribution in the sample, correlation of the attributes, etc.). According to Liu et al. (2018), this process is iterative in the sense that attributes, their relations, and the model can be updated along the way.

Figure 1. The Iterative Retrofitting Framework (Liu et al., 2018, p. 362)


1. Gathering Information about:
• The assessment
• Test takers and users
• Item Responses
2. Attribute and Attribute-Item Specification
3. Modelling Item Responses
• Identifying, and Fitting Models
• Evaluating Fit
• Examining Attribute Correlations
4. Interpreting Results
• Obtaining Individual/Aggregated Reports
• Obtaining Relationship Between the Trait and Attribute Profiles
• Informing Construct Theories
Q-matrix Revision
Model Revision

Limitations of Retrofitting

A principled assessment design would streamline the DCM estimation, (Gierl & Cui, 2008; Rupp et al., 2010) and DCMs are actually intended to accompany diagnostic assessments. Therefore, retrofitting is considered to impose some limitations. Items that hang together are selected for educational tests. As a result, dimensionality would be low or absent in these assessments (Haberman & von Davier, 2007; Gierl & Cui, 2008). When a substantive theory is lacking, decisions about attributes and their relations might be arbitrary (Haberman & von Davier, 2007). Attributes might be too coarse, and the Q-matrix might show a low variability (Deonovic et al., 2019). On the other hand, they might be too sophisticated and cause identifiability or convergence problems (Deonovic et al., 2019; Gierl & Cui, 2008; Lee & Luna-Bazaldua, 2019). In addition, when attributes are highly correlated, it might not be useful to report them separately (Haberman & von Davier, 2007). Additional test items or variables might be required to ensure the accuracy of classifications (Deonovic et al., 2019). In short, the fit of the model or items is not promised, and can be weak (Gierl & Cui, 2008; Rupp et al., 2010). Some of these problems can be overcome by using empirical approaches, as in a Q-matrix validation. Retrofitting can be considered a proxy to obtain diagnostic information. When it is possible to use a diagnostic test, DCMs can produce low stakes feedback to support learning and construct validation (Liu et al., 2018). There are also example studies (e.g., Jang, 2009b; Jang et al., 2013; Kim, 2015 etc.) in which DCMs provided some useful results. These cases are detailed later in this chapter when DCM applications to language assessments are discussed.

Second Language Reading and Divisibility Arguments

Bachman (1990), and Bachman and Palmer (1996, 2010), define language ability to consist of language knowledge and strategic competence (i.e. metacognitive strategies). Language use occurs as a dynamic interplay between the two components (Phakiti, 2003), and is influenced by other factors (e.g., personal, affective, topical etc.). In addition, language knowledge is represented by different components: organizational (e.g., knowledge of vocabulary, syntax, phonology, cohesion, rhetoric etc.) and pragmatic knowledge (e.g., functional and sociolinguistics which are further broken to smaller elements) in their framework. An assessment can incorporate specific elements or a combination of them (Bachman & Palmer, 2010). This prevalent model supports that language construct can be broken down to components. Bachman and Palmer (2010) also perceive four language skills (i.e., reading, writing, listening, speaking) as "the contextualized realization" of language ability (p. 56). Therefore, the model is regarded suitable to explain the L2 reading process especially due to the inclusion of knowledge components as well as strategies (Kim, 2015; Phakiti, 2003; Weir, 2005). However, skills-based conceptualization of language ability is also predominant (Buck, 2001), which encouraged investigation of specific language skills including L2 reading (e.g., speaking in Sawaki, 2007; reading and listening in Song, 2008)

Nevertheless, L2 reading research is not an easy task (Hudson, 1996; Koda, 2012). As Hudson (1996) argues, the difficulty partially stems from the feature of the construct, which entails latent and inferred processes (Hudson, 1996). There are opposing views regarding L2 reading, its divisibility and components (Alderson, 2000). The debate around its attributes is ongoing (Weir, 2005). Some scholars postulate specific attributes cannot be specified, or separated (e.g., Alderson 1990a, Alderson, 1990b; Alderson & Lukami, 1989; Rost, 1993). For instance, Rost (1993), by using a factor analysis approach, proposes a general reading factor to account for most of the variance in the test performance. Similarly, relying on expert opinion, Alderson and Lukami (1989) and Alderson (1990a, 1990b) conclude specific attributes that reading items measure, or a hierarchy for skills (i.e., low vs. high order), cannot be fully determined. However, these studies have been subject to methodological criticism (e.g., composition of panel, operationalization of the hierarchy, factor analysis etc.) (e.g., Mathews, 1990; Weir et al., 1990). The unitary perspective of the reading construct is also found improper. It would suggest how the reading is assessed is insignificant (e.g., focus on grammar vs. inferencing) (Urquhart & Weir, 1998), or identifying reading related problems is not possible (Koda, 2007). Alderson (2000) explains that this standpoint might also lead to the underrepresentation of the construct because there is a risk that some components are ignored.

Reading as a multicomponent construct is also well received by researchers (e.g., Davis, 1968; Grabe, 1991; 2009; Hudson, 1996; Koda, 2007, 2012; Lumley, 1993; Munby, 1978; Urquart & Weir, 1998; Weir 2005). Reading includes several subskills (Koda, 2007; 2012) as well as strategies (Weir, 2005) and thus it is a multidimensional construct. Similarly, Hudson (1996) describes reading as the interplay between processes, knowledge, and abilities such as lexical, syntactical knowledge, as well as a higher order of processes and strategies. Several skill lists and taxonomies have emerged (Alderson & Lukami, 1989), and some like Munby's taxonomy has been influential in language testing (Alderson & Lukami, 1989; Mathews, 1990). For example, Davis (1968) specifies eight reading skills (e.g., recalling/inferring word meaning, making inferences, understanding explicit content, weaving ideas, recognizing purpose etc.). He also develops a test using these skills and concludes reading is not indivisible based on his uniqueness analysis. Munby (1978) proposes eighteen skills including understanding explicit/implicit information, skimming, scanning, understanding cohesion, etc. (as cited in Alderson & Lukami, 1989; Alderson, 2000). Similarly, Grabe (1991) asserts researchers associate L2 reading with six different components at minimum (i.e., recognition, vocabulary and structural knowledge, formal discourse knowledge, content knowledge, synthesis and evaluation skills, and metacognitive knowledge). However, taxonomies are also criticized for not being inclusive (Hedgcock & Ferris, 2009), including overlapping skills (Alderson, 2000; Hedgcock & Ferris, 2009; Macmillan, 2016; Weir et al., 1990), being arbitrary (Mathews, 1990), or divergent from each other (Hedgcock & Ferris, 2009). In addition to the acceptance of the divisibility at the theoretical level, some researchers have undertaken empirical analyses to uncover reading skills or confirm their separation. Lumley (1993) reports a successful attribute specification and ordering by domain experts, which converged with the statistical difficulty of the items. Some studies do confirm some distinction is possible between comprehension of explicit and implicit meanings through factor analysis (Kim, 2009; Song, 2008). Rule space methodology have also been applied to L2 reading, and a large number of attributes are identified in large-scale tests (e.g., Buck et al., 1997). Finally, verbal protocols, eye tracking studies, and self-reports are employed to uncover strategies in reading processes (e.g., Anderson, et al., 1991; Brunfaut & McCray, 2015; Cohen & Upton, 2006).

Despite the lack of consensus on the divisibility or the specific attributes that characterize L2 reading, the divisibility notion is powerful (Alderson, 2000; Lumley, 1993). It is also a sensible approach (Harding et al., 2015) for several reasons. According to Grabe (1991) “reading components perspective is an appropriate research direction to the extent that such an approach leads to important insights into the reading process” (p. 382). Similarly, Weir et al. (1990) argue that efforts to justify skill identifiability are critical and contribute to the understanding of the construct and test development. Despite the uncertainty of the components, they are deemed useful and practical for teachers and testers (Lumley, 1993). In fact, they are acknowledged in teaching and can be possibly incorporated into testing (Weir, 2005). L2 reading needs to be decomposed to distinguish it from other skills, especially if it is reported separately (Urquhart & Weir, 1998; Weir, 2005). Measuring skills and strategies is unavoidable (Weir, 2005). Profiling attributes to uncover weakness, rather than overlooking them, is seen as a reasonable resolution until a consensus is reached about the divisibility (Urquhart & Weir, 1998). However, establishing more descriptive and comprehensive diagnosis theories or models for L2 is necessary to elucidate the complex and multicomponent reading construct (Alderson et al., 2015). In addition, insights about the interactions between the attributes and hierarchical structure is critical (Hedgcock & Ferris, 2009; Urquhart & Weir, 1998; Weir, 2005). Specifically, rather than presuming hierarchies they should be explored empirically (Hedgcock & Ferris, 2009).

With respect to the interaction between skills, some researchers (e.g., Bernhardt, 2005) suggest a more compensatory relationship. However, Bernhardt identifies three dimensions for L2 reading: L1 literacy, L2 ability, and unexplained variance including cognitive skills, strategies, and motivational factors. As seen, an important aspect such as cognitive skills are lumped together. There is still much to be discovered about the relationships of reading attributes including hierarchy and skill dependencies (Weir, 2005). DCMs not only enable identifying attribute interactions and profiling but also overcome the limitations of the earlier methodologies. For instance, expert view can be confirmed by the statistical evidence. According to Jang (2017) “Traditional psychometric approaches often fail to identify multiple skills separately. Common factor analytic approaches are inappropriate for identifying highly correlated skills… (p. 10)". Traditional dimensionality analyses might also fail to support substantive dimensions that might be evident in content analysis or expert judgement (Li & Suen, 2013).

Applications of DCMs to Language Assessments

The earliest DCM application to language assessments can be traced back to the implementation of rule space methodology (i.e., antecedent of DCMs) to large-scale English proficiency tests such as the TOEFL or the Test of English for International Communication (TOEIC) (e.g., Buck et al., 1997; Buck & Tatsuoka, 1998). The motivation for all of the very early applications has been to gain more insights into the language constructs, specifically listening and reading via a new methodology, due to inconclusive findings and limitations presented by earlier traditional, factor analytic approaches (Buck et al., 1997). However, large number of attributes (e.g. 16 attributes, 8 interaction terms) have emerged in these studies. Yet, the early attempts are claimed to be successful (i.e., attributes explained above 95% of variance), and support their capacity for deconstructing L2 domains into smaller attributes in order to understand their impact on learner performance for diagnostic purposes (Lee & Sawaki, 2009). More recent DCM applications for language construct fall into three broad classes: (1) the Q-matrix development and validation, (2) exploration of DCMs and its feasibility for providing diagnostic feedback, and (3) model comparison. Key variables of these studies are provided in Appendix A.

Q-Matrix Studies

There are few studies detailing the specific procedures (i.e., statistical and qualitative) for the Q-matrix development for language assessments (Jang 2009a; Li & Suen, 2013; Sawaki et al., 2009). While Jang (2009a) focuses on the TOEFL iBT reading, Sawaki et al. (2009) add the TOEFL iBT listening to their Fusion model analysis. All of these studies incorporate Q-matrix validation just based on the Fusion model parameters (e.g., combining skills, fixing item parameters). Sawaki et al. (2009) present a rudimentary approach for the Q-matrix development and included broader attributes (i.e., 4 for each domain). In their initial attribute specification, they reviewed test specifications, test development frameworks, the relevant literature (i.e., on subskills of reading and listening constructs), and completed task analyses. The authors claim an acceptable fit for the final Q-matrix, based on consistency of classifications and item parameters.

Jang (2009a) develops a more inclusive approach and draws from multiple sources to identify the initial set of attributes such as a review of the literature for reading taxonomies, test blueprints and frameworks, classical item and dimensionality analysis, as well as textual analysis (e.g., word frequency, text length, rhetorical structure, etc.). Moreover, she incorporates think-aloud protocols to capture the actual skills and processes used by the test takers, and seeks to confirm the attributes emerged with subject experts. Jang specifies nine attributes and asserts most examinees (90%) have been correctly classified by the model. However, the Fusion analysis also show some items are not discriminating ($r^* > 0.8$). She concludes attributes might be highly correlated for these items. Based on the verbal protocols, she mentions that L2 academic reading construct incorporates both compensatory and some non-compensatory attribute associations and draws attention to a more flexible modelling approach.

Li and Suen (2013) study underlying subskills of the reading domain of Michigan English Language Assessment Battery (MELAB). They have also conducted verbal protocols with students to verify the initial set of skills and included expert opinion to create the Q-matrix. In cases of discrepancies between the two sources, the authors rely on the student data, as it is a more reliable indicator of the actual reading processes. Li and Suen also compare the initial and revised Q-matrices and conclude that they yield the same findings (i.e., acceptable fit). However, like Jang Li and Suen report poor diagnostic capacity for certain items on the test.

Clearly, these studies differ to some extent with respect to the scope of empirical evidence involved in constructing the Q-matrix or attributes. Initial skill identification is grounded in literature, reading taxonomies, and an expert view in all studies. Jang (2009a) and Li and Suen (2013) incorporate different sources (e.g., verbal protocols) to arrive at a more reliable Q-matrix. The composition of the panel is worth mentioning. Involving qualified experts about skill processes and their development is a critical factor for successful implementation (Rupp et al., 2010). The expert panels in these studies consist of graduate students, except for Sawaki et al., which can be attributed to the ease of access to this population. However, involving various, more experienced experts, specifically the test/content developer as in Sawaki et al. or teachers, is important due to their increased familiarity with the items or the test taker population. Another important characteristic of the Q-matrix is the specificity of the attributes identified. Although all of the studies focus on similar constructs and Sawaki et al. (2009) and Jang (2009a) use exactly the same test, the number and scope of attributes varied. For instance, while word knowledge is common to both Jang and Sawaki et al., Jang differentiates between context dependent and independent word knowledge. Likewise, Jang considers syntax knowledge to be a part of the reading comprehension, whereas Sawaki et al. do not represent a similar attribute in their Q-matrix. This variation might be expected, as reading comprehension is complex, yet different skills associated with the same test imply the importance of multiple pieces of evidence. However, it is apparent that there is uncertainty with respect to granularity of attributes. As stressed by Rupp et al. (2010) and others, appropriate grain size should be aligned with the purposes of diagnostic information, theory, and estimation requirements. These studies show there are variations how specific procedures are applied. As substantiated by Jang, it is an arduous and complex task to identify meaningful attributes and develop the Q-matrix.

Example Methodology and Feasibility Studies

The next line of studies demonstrates the application of a specific DCM for language assessments and explore the feasibility. Therefore, they entail the evaluation of the model and some of them also involve Q-matrix construction. von Davier (2008) illustrates the use of the general diagnostic model (GDM) for the TOEFL iBT reading and listening domains. He also compares the GDM to a unidimensional and two dimensional IRT models to evaluate the suitability of the GDM. The Q-matrices have been created by consulting experts and no further validation is undertaken. In order to experiment whether reading and listening can be treated as a single comprehension domain, the author employs joint calibrations of these skills. Joint calibrations include a unidimensional model, a two-dimensional model, and the GDM. Overall, two dimensional IRT is observed to perform better than the GDM based on relative fit (i.e., -2log likelihood), yet when skill mastery probabilities are compared to IRT θ estimates, three of the four skills are highly correlated with θ (0.85-0.95). One skill which deviates from the trend is a prerequisite for another skill, meaning a hierarchical relationship exists.

Jang (2009b) applies the Fusion model to LanguEdge (i.e. TOEFL preparation courseware). The justification for DCM has a sound basis in this study as LanguEdge serves as a teaching and learning tool. Jang also uses self-assessment surveys to probe into the relationship between reading profiles and an examinee's self-efficacy with their reading ability. She fits the model twice, and between two applications the instruction continues. What makes this study exceptional is the quest for the utility evidence. Jang creates diagnostic report cards for examinees and their instructors. With both groups, she conducts pre- and post- surveys/interviews inquiring about their perceptions of the reports (i.e., elements in the reports) and/or the extent to which these reports are useful for learning or teaching. In assessing the model, Jang examines parameter estimates (i.e., difficulty, discrimination), MADcor, correct classification rates and concludes the model fit overall is reasonable (e.g., MADcor < 0.05). The mastery status of most students is determined. Mastery probabilities and observed total scores are also positively correlated (0.94), and most items but not all show reasonable diagnostic capacity. In the second application Jang notes positive changes in attribute status for almost half of the students. However, students' self-assessments are moderately or weakly correlated with the attribute mastery probabilities. Most students have a positive attitude towards their report cards and believe they are helpful and informative, to some extent. However, the reports have created some confusion, such as the meaning of being a master of a skill but not getting a perfect score. Jang also mentions the utility of information hinges upon congruity between teachers' opinion about student performance and statistical results.

Kim (2015) also undertakes a Fusion analysis for a local college level L2 reading test. Unlike other studies, the Q-matrix for this study identifies a combination of cognitive strategies (e.g., skimming) and linguistic components (e.g., lexical and cohesive meaning). Kim includes more attributes than other studies (i.e., 10), which rely on language ability models (i.e., Bachman & Palmer, 1996; Purpura, 2004) and literature on L2 reading construct. Similar to Jang, Kim evaluates the item parameters, classification consistency, and observed and predicted estimates. She also analyzes skill mastery profiles at different proficiency levels (e.g., beginner, intermediate), which yields important findings. According to Kim, there is variation with respect to mastery status at different levels. Overall, Kim's study reveals an acceptable fit (MADcor < 0.05). Yet, about half of the items poorly differentiate between masters and non-masters (< 0.40). Conversely, examinees classifications are consistent 88% of the time. Despite Kim's motivation to investigate pedagogical usefulness for different stakeholders, she does not dig into the utility aspect. However, specifically, the analysis of attribute mastery across different proficiency levels support DCMs can offer valuable information with respect to learning paths for different subgroups.

In a similar local, high-stakes, L2 proficiency test context, Ravand (2016) showcases an application of the G-DINA. In specifying the attributes, the author works with domain experts and students taking the exam previously to extract the skills. The Q-matrix is also revised with the GDI method. In judging model fit, Ravand uses several fit indices (i.e., MADcor, RMSEA, MADres, $M\chi^2$) and classification consistency and accuracy. Therefore, the evaluation of the model fit is more extensive than the other methodological examples. The author finds model fit and classifications acceptable (e.g., fit indices < 0.05, classification accuracy and consistency are 0.81 and 0.73 respectively). Yet, the variability of classes is limited, and two flat classes are apparent. The correlation between attributes are also high for some attributes (i.e., 0.78-0.95). Ravand evaluates the models at the item level and finds both compensatory and non-compensatory relationships apply to items. He attributes the altering skill relationships to different factors such as attribute difficulty, content area, and cognitive intensity of the skills.

Kim (2011) and Xie (2017) diverge from other studies as they successfully initiate the R-RUM for an L2 writing domain. Rather than using the writing prompts as the starting point for the skill specification, Kim (2011) has developed a writing rubric and identified the subskills by focusing on the raters' cognitive processes while rating the writing prompts. For the R-RUM application, 2 essays of 120 students graded dichotomously based on the checklist are used. This study is a relatively small-scale research compared to others. Overall, the author mentions good fit. However, 34% of the descriptors (i.e., with the majority being related to grammar) lack diagnostic power which might have stemmed from misspecifications in the Q-matrix. Nevertheless, the author notes variation in the attribute profiles and evidence for test-retest reliability (i.e., similar proportion of masters for all skills across two forms except for mechanics attribute). Xie (2017) uses the same checklist and the R-RUM for another small scale (N = 472) local writing test to replicate the results. However, unlike Kim, Xie employs three-facet Rasch for rater reliability and finds significant rater severity. Rater severity is not unexpected, because some raters have graded over 70 essays. The Q-matrix has been slightly revised based on statistical evidence and the results between initial and revised matrix are compared. The revised Q-matrix outperforms the original matrix (i.e., greater discrimination for items). The fit based on observed and estimated estimates is reported reasonable.

Methodological examples for L2 proficiency at K-12 context is limited. Jang et al. (2013) is one of the few examples. The authors explore reading development of more than 10,000 students using a literacy test (i.e., reading, writing and, grammar). The authors specifically focus on immigrant and Canadian born L2 speakers and scrutinize the impact of the length of residence and home language on reading skill mastery, which is estimated employing the R-RUM. An expert team of graduate students have created the Q-matrix by analyzing items, which is refined with statistical findings from the R-RUM. The closeness of observed and estimated parameters and high correct classification rate (0.83- 0.98) are considered evidence for model accuracy. The results reveal that length of residence and reading achievement are related. The authors discuss, despite that native students outperform their peers, multilingual students reach similar performance levels within 5 years. However, they acknowledge time itself cannot be a sufficient to explain performance as students vary with respect skill development patterns. This study also shows given good fit; DCM can be insightful to explain learning trajectories for various subgroups of L2 speakers.

In another study Jang, et al. (2015) explore the utility of the diagnostic information in the K-12 context. The authors consider factors such as reading profiles, self-efficacy, and goal orientation, and whether these factors agree with attribute profiles. The authors also collect feedback about the diagnostic reports. The study involves parents in the evaluation of the diagnostic information. The authors make use of the same assessment described above with a very small sample (N = 44) and obtain the profiles from the R-RUM. The result indicate with some scaffolding students are able to understand reports. Also, students who are oriented to master reading skills are more interested in diagnostic information. The authors find, the relationship between self-assessments and skill mastery was conflicting such that students with low mastery overrate ability. The pattern is switched for students with high mastery probabilities. The parents find the reports informative and even some have initiated a discussion with their child and sought further action to help them. Some parents also desire to get guidance about how to help their kids.

The major limitation in these methodological studies is the choice of a model. Six out of eight studies employ the non-compensatory RUM models, but not justify is their choice enough. Only Kim (2011) and Xie (2017) mention that the R-RUM is suitable given its prevalent use. However, the justification is not comprehensive enough. The choice of the RUM models might be closely associated with the accessibility of computer programs. These studies also verify the validity of DCM applications is nascent in general. Utility, relations to external variables, are overlooked. Only Jang (2009b), Jang et al. (2015) explore external relations (e.g., self-assessments) and utility. However, exploration of the use of diagnostic information might reveal intriguing findings. For instance, it is shown that diagnostic information can be confusing for students or conflicting for teachers.

Model Comparison Studies

Emergence of computer programs and frameworks allowing to fit multiple models (e.g. the CDM package and the LCDM/GDINA framework) have led to several model comparison studies. In addition to being concerned with the best model to represent language assessments, the studies contribute to insights about the attribute associations (i.e., compensatory or non-compensatory). These studies also merit attention for incorporating and exemplifying several model fit indices.

Lee and Sawaki (2009) evaluate the performance of the GDM, latent class analysis (LCA), and the Fusion model using the same Q-matrix and assessment from Sawaki et al. (2009). They inspect RMSEA, distribution of masters and non-masters for separability, examinee proportions in each profile under each model, and classification consistency across two forms (i.e., test-retest reliability). In the study Fusion model yields better fit based on RMSEA. Three models perform similarly with respect to other criteria. For instance, except for one skill in both reading and listening domains, distribution of mastery probabilities is similar. However, the variability of profiles is low across all three models with the majority of the examinees being assigned to two profiles. The authors also find the classification under GDM to be more consistent and moderate relationship between the forms except for organizing and synthesis reading skills. The general ability is also moderately correlated with attributes under models. They suggest that, despite being designed to be parallel, form might not necessarily tap the same subskills in the Q-matrix. In conclusion, all models yield similar results in the study.

Yi (2017) also fits four models (i.e., the DINA, DINO, NIDO, C-RUM) within the LCDM framework using the same data and Q-matrix as Lee and Sawaki. The model evaluation includes analysis of RMSEA, AIC and BIC. All models show good fit based on RMSEA (< 0.05) yet, the NIDO produces the highest RMSEA. The C-RUM exhibits the best fit based on AIC and BIC. Again, the NIDO is the worst fitting model based on relative fit. Yi concludes L2 comprehension skills to involve compensatory relationship. He points out that the relationship between subskills is another important consideration in test construction and validation. Therefore, he emphasizes that test developers should consider skill relationship during item design.

Li et al. (2016) implement the GDINA framework to select an appropriate model for MELAB. Four specific models, the DINA, DINO, ACDM, and R-RUM are also compared. The study is more comprehensive and involves comparison of classification results, relative fit, as well as absolute fit. More specifically, Li et al. (2016) analyze the examinees proportions across skill profiles, like Lee and Sawaki (2009), and report -2LL, AIC, BIC, MADcor, MADRES, and $M\chi^2$. Overall, the ACDM and R-RUM show comparable fit with the LCDM based on the relative and absolute fit indices. The DINA and DINO underperform. Skill profiles are also comparable among the LCDM, ACDM, and R-RUM while DINA and DINO show some divergence. The authors conclude ACDM as a more parsimonious model to be more appropriate and claim evidence for compensatory nature of reading attributes.

Another model comparison study of L2 reading within GDINA framework is Ravand and Robitzsch (2018). The authors investigate the performance of six models (i.e., the G-DINA, DINA, DINO, R-RUM, and ACDM) for a local high-stakes reading test. The Q-matrix is adopted from Ravand (2016). Overall, the G-DINA outperforms the other models, and the C-RUM is the second-best fitting model. Yet, the authors express the R-RUM and ACDM show comparable fit to the G-DINA with respect to absolute and relative fit and proportion of examinees across skill profiles. The DINA and DINO show more divergence. In conclusion, the authors suggest that compensatory and non-compensatory models perform equally given a large sample size (N = 21,642). Therefore, the authors advocate an item-level model specification.

Aryadoust (2018) also works with a local high-stakes test, yet at K-12 level. Diagnostic information carries more significance as the author mentisons mock tests are part of the teaching curriculum. The study is oriented to understand the nature of listening comprehension comparing five models: the DINA, DINO, HO-DINA, G-DINA, and R-RUM. It is also the smallest scale study (N = 205) for model comparisons. The Q-matrix for the study is developed by reviewing test frameworks, conducting verbal protocols with examinees, and innovatively incorporating an eye-tracking study. The subskills include test-taking strategies, in addition to listening skills, which distinguishes it from previous studies. The Q-matrix is not revised iteratively in this study. This study also relies on comparison of absolute and relative fit indices. The best fit is displayed by the R-RUM. The G-DINA is also a comparable model. The DINO underperforms in comparison to other models. However, based on tetrachoric correlations, some attributes seem to correlate above 0.80 (i.e., high correlation). Also, 67% of the examinees are assigned into the profile where all skills are mastered.

The model comparison studies show that the LCDM/GDINA ascend to be trending frameworks. Mixed conclusions with respect to model choice corroborate the utility of more generalized frameworks. Specifically, for a complex construct such as language, a general framework allows item specific models (e.g., Ravand, 2016; Ravand-Robitzsch, 2018). Yet, it is important to indicate the rationale for the selection of the constrained models is not laid out well in some studies. Moreover, analog models have not been employed in some instances.

The findings of model comparisons are not conclusive in these studies, either. In general, RUM models have been found to perform similar to general models. Also, the DINA and DINO fit worse, which supports their restricted nature (e.g., Henson et al., 2009). Except for Li et al. (2016), researcher advocate both compensatory and non-compensatory models for reading comprehension. However, these conclusions might be ambitious and should be verified in future studies.

Summary

The studies discussed here provide insights about the language construct and illustrate the feasibility of new modelling avenues. They bear similarities and differences and present important implications. All of these applications are retrofitting. Except for two, the studies reviewed deal with receptive language constructs, such as grammar, listening, and reading. Sessoms and Henson (2018) also note that 39% of DCM implementations after 2009 are concerned with L1 or L2 reading constructs. It can thus be inferred that receptive skills are easier to deconstruct into subskills. Despite some variation, there are also common skills across studies for the same construct (e.g., main idea, inference, vocabulary, and syntax for reading construct). The studies also show variation with respect to the number of attributes (i.e., range of 3-10), hinting at the uncertainty of the attributes raised by some researchers (e.g., Alderson, 2000; Lumley, 1993). Although Jang (2009b) stresses the importance of blending different sources in Q-matrix construction, working with domain experts and task analysis is the most common approach. One limitation voiced by Alderson (2010) is the lack of documentation for the specific procedures in deconstructing items. The empirical Q-matrix validation methods have not been adopted in these studies with the exception of Ravand (2016), which might have a high potential to overcome mis-specification issues. Additionally, the authors claim mixed findings with respect to the nature of attribute relationships, signifying the need for more research to shed light on L2 reading. As echoed by Stout et al. (2019) Fusion model has been used in study after study yet, general models are preferred in more recent studies. Conclusions about model choices should be treated cautiously and should not be oversimplified. As substantiated by Ravand (2016) and Ravand and Robitzsch (2018), the correct model might interact with item features. The studies also show the validity of the DCM methodology is an underexplored area. Evidence for relations to external variables and utility are scarce. Utility aspects can clarify some issues like grain size. How diagnostic information is perceived and used by different stakeholders are thus important questions to address. Finally, except for a few studies (Aryadoust, 2018; Jang et al., 2013; 2015), the context of the studies is for standardized assessments in higher education. However, none of the studies are concerned with English Language Proficiency (ELP) assessments for young learners.

CHAPTER III

METHODOLOGY

The previous two chapters provided the background information for the study. Chapter 1 described the appeal for diagnostic information from large-scale tests and the use of DCMs with these tests due to a lack of diagnostic assessments and skepticism about other methods (e.g., teacher assessments). Chapter 1 also introduced the study context, ELs and ELP assessments. Diagnostic feedback might better allow attending to ELs' needs and enhance learning opportunities for this population who show performance gaps. Chapter 2 reviewed the notion of diagnosis in language assessments and different arguments about L2 reading construct. Despite the interest, the field lacks a theory or systematic methods for diagnostic language testing. Chapter 2 also detailed different aspects of a DCM methodology. In a DCM study, first, attributes are specified and coded for test items which are then modelled along with item responses by using a general (e.g., LCDM) or a constrained model (e.g., DINA, C-RUM etc.). If multiple models are used, they are compared using relative and absolute fit indices. Chapter 2 also delved into the research studies using language assessments, most which pertained to comprehension constructs and college level proficiency tests. This chapter, Chapter 3, focuses on how the DCM methodology was implemented in the present study. The chapter is structured in four parts. First, the purpose of the study along with the research questions are reviewed. Subsequently, a detailed account of the data is presented. This includes information about the assessment, the test form that was used in the study, and the participants that the form was administered to. In the third part, specific procedures and analyses to address each of the research questions are detailed.

Purpose and Research Questions

The purpose of this study is to explore the utility of the DCM methodology to obtain diagnostic information from a large-scale, K-12, ELP assessment. For this purpose, the diagnostic methodology was applied to the reading domain of Assessing Comprehension and Communication in English State-to-State for ELs (ACCESS) in the study. By implementing an alternative measurement framework, the study intends to illustrate the extent to which the DCM framework can offer more detailed, actionable results to test users including ELs and their teachers for ACCESS reading domain. It must be acknowledged that a different measurement framework, IRT, informed the development of the test. Therefore, the study employed a retrofitting approach for exploratory reasons. It adopted the steps in the retrofitting framework presented in Chapter 2. The study does not intend to suggest replacing the current measurement framework for the test, but rather pursue low-stakes feedback with the DCM methodology that can aid everyday instructional settings. Informed by this purpose, the study proposes to identify the attributes a suitable diagnostic model for the reading domain. The study also aims to evaluate the selected model for the viability of the methodology by focusing on item, person, and attribute estimates, as well as comparisons between the DCM based classifications and ELs' proficiency estimated under the original framework (i.e., IRT). The following research questions guide the study:

(1) What are key underlying attributes represented in the ACCESS reading domain in middle grades for more advanced ELs?

(2) What DCM fits the data better?

   a. Does a general or specific restricted model better represent all items in the test?

   b. Does a Standard-based or an Expert-defined Q-matrix show better fit?

(3) To what extent is it feasible to obtain diagnostic information using DCM?

   a. What is the diagnostic capacity of the test items?

   b. To what extent can students be appropriately classified using the model?

Assessment: ACCESS

ACCESS is an annual, large-scale, standardized language proficiency assessment for K-12 ELs. It is a widely used instrument in the U.S., given that it is administered to over 2 million ELs across 39 states every year. The test has paper and online versions. While some states only offer paper ACCESS (e.g., Florida), others administer the online test, which is multi-stage adaptive. States can also choose to offer both. For the purposes of this study, paper ACCESS was used. The test is a product of the World Class Instructional Design and Instruction (WIDA), a multi-state consortium based in the University of Wisconsin-Madison, and the Center for Applied Linguistics (CAL). The two organizations are responsible for the design, development, technical quality, and validation of the assessment.

The stakes attached to ACCESS are fairly high as the test is primarily used for decisions related to proficiency development and planning (Fox & Fairbairn, 2011). As noted by Fox and Fairbairn, each EL student continues to take ACCESS until they become proficient depending on their states' criteria. The test can also be used for different purposes as claimed by the test developer, such as; program and curricular decisions (e.g., placement in support programs, curriculum development), instructional planning and classroom assessment (e.g., scaffolding students, determining domains to focus on) (CAL, 2017; Kim et al., 2016; WIDA Consortium, 2019; WIDA, n.d.). Specially, to foster the link between the test and instruction, the test is linked to the instructional resources (Fox & Fairbairn, 2011) (e.g., standards, can-do descriptors). Therefore, the results of ACCESS potentially influence teaching/learning decisions, and diagnostic feedback can be deemed valuable. In addition, diagnostic information can support the interpretation of scores and contribute to the validity of inferences made about an EL's proficiency. However, validity of the diagnostic information itself should be ensured before such endeavors are undertaken.

ACCESS is designed to measure social, instructional, and academic English proficiency (Bauman et al., 2007; CAL, 2017; Wolf et al., 2008a). It aims evaluate ELs' proficiency, and their language competence in the classroom and school context, especially in their interactions with content, peers, and teachers (WIDA Consortium, 2019). Although it incorporates the social dimension, it is within the context of the school environment. It incorporates all four language domains: listening, reading, speaking and writing. Each domain is tested separately. This study focused solely on the reading domain. The reading construct is operationalized as "process, understand, interpret, and evaluate written language, symbols, and text with understanding and fluency" (WIDA Consortium, 2007, p.11). It thus entails comprehension of texts related to academic content or classroom/school setting. I chose to delve into the reading domain because as a receptive domain, reading is given more weight to estimate the overall score (i.e., reading and writing account for 35% each, speaking and listening 15%) by the test developer (CAL, 2017; WIDA Consortium, 2020.). Thus, between the two comprehension skills, it is given slightly more importance. Reading is also acknowledged to be crucial skill for academic development of L2 learners (Grabe, 1991). DCMs are also frequently applied to L2 reading construct given the belief for its divisibility (e.g., Davis, 1968; Grabe, 1991; Weir et al., 1990 etc.). For this reason, it is possible to draw comparisons across the studies.

ACCESS is also a standards-based assessment, and each domain in ACCESS is linked to the WIDA standards: (1) *social and instructional language*, (2) *the language of language arts*, (3) *the language of mathematics*, (4) *the language of science*, and (5) *the language of social studies*. While the first standard expresses ELs' ability to "communicate for social and instructional purposes" in the academic environment, the remaining standards are related to ELs' competence to "communicate information, ideas, and concepts necessary for academic success" in each of the corresponding subject area (WIDA Consortium, 2012, p. 4). Each ACCESS test item measures ELs' ability to use English for communication in relation one of the standards. As can be realized from the standards, the test underscores academic English and communication (Bauman et al., 2007; WIDA Consortium, 2014).

With the belief that authentic contexts increase learning opportunities (WIDA Consortium, 2012)., a thematic approach is undertaken when writing test items (Fox & Fairbairn, 2011). In other words, items and tasks are written around different themes in relation to each standard, which are also aligned with standards of other content areas (WIDA, 2012). Some example themes that are illustrated by WIDA Consortium (2007, 2012) include classroom activities, assignments, instructions, research, school life and behavior *for social instructional language*; narration, main ideas, literature, peer editing, biographie*s for language of arts*; decimals, lines, fractions, algebra, data interpretation *for math*; ecosystems, climate, solar system, life cycle*s for science*; globalization, maps, agriculture, democracy, maps, government *for social studies*. It must be highlighted that the purpose of ACCESS is not to measure the ELs' knowledge with respect to these topics (Bauman et al., 2007). Themes just provide a context to realize language objectives. The focus is on an EL's knowledge of language and their language use.

In the K-12 context, research indicates that English language learning is assumed to be a process that spans over multiple years and language use to be affected by age and maturity (WIDA Consortium, 2014). ELs enter the school system with varying levels of English proficiency. Grade level does not correspond to English proficiency. Put differently, an EL in higher grade levels can still have lower English proficiency. Therefore, for every grade cluster, tests have to cover a wide range of proficiency levels. The same test cannot be used for different grade levels because of varying degrees of maturity, and content requisites. Given this assumption, separate test forms are developed for specific grade clusters. There are six different clusters (i.e., vertical dimension): Kindergarten, Grade 1, Grades 2-3, Grades 4-5, Grades 6-8, and Grades 9-12. In addition, forms in each grade cluster are broken down into tiers which differ with respect to their difficulty (i.e., beginning: A, intermediate: B, advanced: C) (WIDA Consortium, 2019). The purpose of the tiered system is to give ELs a better testing experience by delivering them a form that is more appropriate for their level. Tier decisions for the paper tests are made by educators and benchmark tests if available (e.g., WIDA Model).

In order to describe an EL's performance on the test, a scale score and a proficiency level is reported for each domain. These domain scores are combined with different weights to report several composite scores (e.g., overall, literacy, oral, comprehension scores). Item difficulty is taken into consideration in estimating scale scores and scales differ across domains. Proficiency levels, whether domain or composite, are estimated using scale scores. There are 6 proficiency levels (i.e., Entering, Beginning, Developing, Expanding, Bridging and Reaching) associated with different descriptors at each level. The descriptions relate to three dimensions: vocabulary use as the word dimension, language forms and convention as the sentence dimension, and linguistic complexity as the discourse dimension (WIDA Consortium, 2012). Proficiency level scale score correspondence varies across domains as well as grades because the language ELs are exposed to and are required to produce are not the same across grades. Proficiency levels allow connecting test performance with performance definition of the standards and can-do descriptors. For more information about the scores, the reader is recommended to see WIDA score guide (WIDA Consortium, 2020).

ACCESS shows complex design features (Fox & Fairbairn, 2011). It embeds multiple levels (e.g., grades, clusters) and involves various dimensions (e.g., domains, standards, language functions). Overall, the ACCESS system has been documented to have good psychometric properties (Bauman, 2007; Bunch, 2011; Fox & Fairbairn, 2011). In addition, the quality of the system is assured through rigorous validity studies, and continuous test reviews (Wolf et al., 2008).

Test Form Used in the Study

As stated above, this study focused on ACCESS reading domain. Because test items vary in each grade cluster and tier (i.e., form), one cluster and one tier were chosen for the manageability of the study because even test forms targeting the same level might differ with respect to the underlying skills they measure (Lee & Sawaki 2009).

This study utilized Grade 6-8, tier C form. A middle grade cluster was chosen as ELs are exposed to more elaborate language that might present richer attributes and attribute interactions at this level. For instance, test items in Grade 1 tend to measure more picture-word associations or would demand processing a few sentences as students are just developing their literacy skills. Furthermore, in middle grades, diagnostic feedback has more utility. ELs in the middle grades still have several more years of education in front of them and can benefit from diagnostic information unlike high school ELs who are closer to exiting the school system sooner. In addition, upper elementary grades have received less attention in reading and literacy research (Koda, 2007). A more advanced tier was preferred for the study for similar reasons. According to WIDA Consortium (2019), tier A forms are designed for beginner ELs who have just joined the U.S. school system and have zero or very limited English background. Thus, the range of proficiency was potentially very narrow among these students. On the other hand, tiers B and C target ELs with developing or expanding proficiency. These ELs have been gaining language proficiency for some time yet, they might still lack some necessary skills to be deemed fully proficient. Therefore, proficiency variance was expected to be greater for the B and C forms, which is desired for a test to show better diagnostic capacity.

The test form that was used for the study included 27 items. The test length was appropriate given de la Torre's (2009) recommendation of minimum 15-20 items for the DCM analysis. There were also practice items preceding the actual questions. Practice items served to warm up students for the domain. They were not part of the scoring. Full practice sets for the reading or other domains are not available. The test developer would like to guard against teaching to the test by not releasing a full form (Fox & Fairbairn, 2011). Given the confidentiality of the items and copyright issues, I was not able to present the actual items here. Sample reading items can be found on test developer's website.

A total of nine themes were included in the form. Three items were associated with each theme or topic. The length of the text ELs need to process differed for each item. Some items related to the same topic were based on a common longer reading text, commonly referred as testlets (Wainer & Kiely, 1987). These reading testlets were roughly 210-270 words. Other items were tied to shorter, individual reading paragraphs ranging between 50 and 130 words. Paragraphs within the same theme were related but each item required processing the individual paragraph independently. With respect to the structure, texts included in this form were descriptive. They also included descriptions of a process or cycle or problem-solution relations. Graphical elements such as shapes, pictures accompanied the items. All items were in multiple choice format with four options. They required a single correct response and are scored dichotomously. Missing responses were coded as incorrect.

Table 1. Types and Length of the Texts Used in the Study

Item number	Text type	Length
1-3	Instructional text	272 words
4-6	Newspaper story	250 words
7, 8, 9,10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 25, 26, 27	Short content-related paragraphs	50 – 130 words
16-18	Manual	209 words
22-24	Textbook excerpt	219 words

Test-taker Sample

The data for the study came from 2017-2018 administration of ACCESS Paper. A total of 23,944 Grade 6-8 ELs responded to the reading form C that was used for the analysis in the study. 2 students were removed from the analysis because they were reported to be in grades 4-5. Item responses of 23,942 ELs were used in the study. 42% of the ELs taking the test were enrolled in grade 6, 31% were enrolled in grade 7, and 27% were enrolled in grade 8 by the time they took the test. 976 unique schools used the specific form in the study. It was surmised that not all of the schools using the paper form were low socio-economic schools and delivered the paper format for this reason. Although state information was not included in the data set for confidentiality, it is known that some states like Florida administer only paper ACCESS to their ELs. So, the ELs were not necessarily low-performing students and variability was expected with respect to performance.

Table 2. Demographic Characteristics of the Test-takers

	Grade 6	Grade 7	Grade 8	Overall
N	10,099 (42%)	7,425 (31%)	6,418 (27%)	23,942
Gender				
Female	4,652 (19.4%)	3,383 (14.1%)	2,995 (12.5%)	11,031 (46%)
Male	5,413 (22.6%)	4,038 (16.8%)	3,417 (14.3%)	12,869 (53.7%)
Ethnicity				
Hispanic	8,074 (33.7%)	5,879 (24.6%)	5,127 (21.4%)	19,082 (79.7%)
African-American	1,261 (5.3%)	916 (3.8%)	762 (3.2%)	2,939 (12.3%)
Asian	394 (1.6%)	300 (1.3%)	240 (1%)	934 (3.9%)
Native American	482 (2%)	452 (1.9%)	451 (1.9%)	1,385 (5.8%)

Table 2 above summarizes the demographic characteristics of the ELs in each grade. It must be noted that gender percentages do not add up to 100% because this information was not reported for some students. In addition, some students might have belonged to multi-ethnic families and they selected more than one ethnicity characterization to describe themselves. Among all ELs taking the form C, 46% were female and 53.7% were male students. The ethnicity distribution shows that majority of the ELs taking the form were Hispanic (i.e., 79.7%). African American ELs constituted the second biggest sub-group with 12.3%. All of these ELs were also participating in EL support programs. Except for newcomer students, all students were in these programs for more than a year, with an average time of 4 years. Few students reported that they were benefitting from these services more than 10 years which might be attributed to a miscoding issue.

Procedures and Analyses

Research Question 1

The initial step in a DCM study is the specification of the attributes and developing a Q-matrix. The first research question thus pertained to attributes in the ACCESS reading domain, and the specific procedures to identify and code the attributes that are necessary to respond to each test item. As presented in Chapter 2 and mentioned by Rupp et al. (2010), various methods from task decomposition to eye-tracking are used when defining the attributes in an assessment. In this process, different sets of attributes or Q-matrices might emerge (Sawaki et al., 2009), especially in a retrofitting context. In this study, two sets of attributes and two Q-matrices were developed using different approaches. Because this is a retrofitting study and actual attributes are not known, it is judicious to define attributes in alternative ways in search of the most feasible solution to explain performance. The two matrices to be described next were compared based on the model fit indices (i.e., refer to procedures for the second research question). Because multiple models (i.e., 6) were considered in the study, the Q-matrix and model combination that showed a better fit was adopted for the remaining analyses. This involved 12 comparisons (6 models for each Q-matrix).

Development of the Standards-based Q-matrix

Buck and Tatsuoka (1998) show that a variety of components including task features, knowledge components, skills, strategies, processes, or anything that might influence performance, can be represented as attributes. According to Li and Suen (2013) test specifications, if available, can be used as a starting point for the Q-matrix development, as a practical and reasonable strategy (e.g., Chen & Chen, 2016; Xu & von Davier, 2008). In this respect, test standards can function in the same way. Test items were designed to measure the five standards described in the preceding section. These standards represent ability to use English in five areas (i.e., social-instructional, math, language of arts, science, and social sciences) and represented a dimension, or in other words an attribute in the study. However, standards are coarsely described, and they do not indicate what specific knowledge components or processes students apply when responding to an item. In addition to the standards, each reading item was developed to measure one of the three "key uses of language": recount, explain, or argue. These key uses depict the processes ELs engage with in responding to items. For instance, an "argue" function is associated with identifying evidence or differentiating facts from opinions. Along with the standards, key uses of language were specified as attributes for a more nuanced explanation of test performance. Students are not likely to engage with an item by isolating use from the targeted standard (i.e., content area language). It is exactly the interaction of the standard and key language use that students have to consider when attempting to respond to questions. Another reason to combine the standards and key uses was the structure of the Q-matrix. Each item was associated with one standard and one key use. To put it differently, using only standards or key uses would result in a simple structure. Although it is possible to implement DCMs to data with simple structure, it causes information loss (Rupp & Templin, 2011). According to Rupp and Templin, the model resembles a multidimensional factor analysis or IRT as "multidimensional continua are merely discretized" (p. 230-231). More complex structure also yields intriguing information such as interdependence of attributes given item responses. However, the authors argue that DCMs with simple structure may still be desired due to the classification mechanism. A simple structure may also be more appropriate (i.e., what seems to fit the data best) especially given how the test is constructed.

The standard-item associations were readily available in the blueprints for the test form. Unfortunately, the key use-item relationships were not included in these blueprints or presented elsewhere for the form used in the study. For this reason, the researcher mapped each key use to the items. The grade and domain-specific definitions and examples of key uses (Table 3), as well as the item specifications were used in this process. More specifically, item specifications (i.e., blueprints) laid out model performance indicators, which describe the item/task features and language functions needed for the item. The researcher matched key uses to the items based on these descriptions. Item content was also reviewed to confirm the decisions, because in some cases the item itself did not align with the blueprint specification. The researcher also documented the rationale for the coding of each item in detail. Because this was a subjective process to some degree, the researcher worked with two experts from the test developer team who were familiar with the standards and key uses to verify the matching. Researcher's coding along with the rationale for key use choices, and detailed definition and examples of key uses were provided to the experts. They each indicated their agreement with key use-item associations, and provided short descriptions when they disagreed along with their choice. The researcher then reviewed all input and coded a key use for an item if at least two experts agreed on it.

Table 3. Definitions of the Key Uses of Language (WIDA Consortium, 2016, pp. 2-9)

Key Use	Definition	Reading Processes include:
Recount	To display knowledge or narrate experiences or event	Identifying/sequencing topic sentences, main ideas, details, conclusions, summaries; matching details of content-related topics to main ideas, summarizing text absent judgements, evaluating how a central event is introduced/elaborated
Explain	To clarify the "why or the "how" of ideas, actions, or phenomena	Sequencing events based on cause and effect, highlighting evidence that points how systems function, sorting elements of genre, comparing and contrasting information, identifying factors that contribute to phenomena
Argue	To persuade by making claims supported by evidence	Identifying/evaluating evidence to support analysis, classifying pros and cons of claims, developing a stance, distinguishing among facts, judgements, speculation

In addition to reflecting what the test developer intends to measure, standards and key uses could be more easily interpreted by teachers due to their familiarity with these dimensions (i.e., tied to instructional design and materials). However, it was anticipated this Q-matrix might pose certain estimation challenges (i.e., time intensive) due to having a large number of attributes (i.e., 5 standards + 3 key uses and a total of 8). Example processes in the key uses also demonstrated that it was possible to define more traditional reading attributes for the items (i.e., identifying main idea, inference).

Development of the Expert-defined[2] Q-matrix

Exploring several competing Q-matrices can be informative although they represent different theoretical assumptions (Kunina-Habenicht et al., 2012). Jang (2009a) also recommends including different Q-matrices to evaluate the completeness of the Q-matrix. Similarly, Kang et al. (2010) suggest comparing Q-matrices developed with different approaches. Reid et al. (2018) provide a standards-defined and expert-defined approaches to Q-matrix creation. An examination of the performance descriptions also supported the plausibility of creating another Q-matrix, by breaking down reading using conventional reading processes or subskills (see examples below). Yet, these skills were not prespecified or there was a list of attributes to work with.

- Locate main ideas about behaviors (WIDA Consortium, 2012, p. 80)
- Infer results of adhering or not adhering to behavioral expectations (WIDA Consortium, 2012, p. 80)
- Draw conclusions about resources or agricultural products (WIDA Consortium, 2012, p. 34)

Postulating the attributes and identifying the Q-matrix demands input from domain experts that might include, teachers, researchers, test developers, or measurement

[2] The term is first used by Reid et al. (2018).

experts (Kunina-Habenicht et al., 2012; Madison & Bradshaw, 2015). For the sake of this study, a group of domain and measurement experts including experts from the test developer were convened.

Subject Matter Experts. A team of seven experts contributed to the development of the second Q-matrix. Unlike the common trend of inclusion of mostly graduate students for accessibility reasons, this research also brought the test developer on board. Test developer involvement was particularly insightful due to their close acquaintance with the content of the test. Before describing the process, a brief background of each subject matter expert (SME) is provided.

SME 1 is a professor of educational measurement. Serving on advisory boards of various testing programs, the SME has considerable experience in test design and development. The SME has also managed the development of an ELP test for accountability purposes and is quite knowledgeable about the context of the study and population. SME 2 and 3 are graduate students specializing in language testing. Both SMEs have undergraduate and/or graduate degrees in linguistics/applied linguistics. The two SMEs are also quite familiar with the DCM methodology and have experienced teaching English as a foreign language (< 4 years).

The rest of the SMEs (SME 4 - 7) are from the test developer team who are responsible for design, development, review (including content), and validation of the ACCESS system. They have been previously engaged in the development of other high stakes ELP tests as well. All of the SMEs from the test developer also have a graduate degree in a related field such as English, applied linguistics, and language testing. They have had several years of experience in teaching English as a foreign language. One of them also has a sound grasp of DCMs and the Q-matrix development.

Table 4. Background Characteristics of the SMEs

	SME 1	SME 2	SME 3	SME 4	SME 5	SME 6	SME 7
Native speaker		√		√		√	√
Degree in Applied Linguistics or Language Testing	√	√	√	√	√	√	
Teaching Experience	√	√	√	√	√	√	√
Test Development Experience	√			√		√	√
Familiarity with DCMs		√	√		√		
Familiarity with test content				√	√	√	√

The Process. Content analysis can be used to draft the initial list of attributes (e.g., Sawaki et al., 2009; Jang et al., 2013), and it requires an analysis of the content of the items and decomposing them in order to extract the skills and process. Specifically, experts solve items and specify the skills each item is associated with by answering "what skills and/or processes are required in order for a learner to answer this question correctly?" (Sawaki et al., 2009, p. 196).

Three of the SMEs (i.e., SME 1, 2, and 3) were engaged with the task analysis to establish attributes for the second Q-matrix. This process involved both individual and group work. In this process SME 3, who was the researcher in this study, conducted a literature review of common L2 reading attributes. Several theoretical reading taxonomies, empirical studies about reading skills, including DCM studies, were explored to exemplify common reading skills (See Appendix A, C). Th example list served as a helpful document in the item content analysis process. In an introductory meeting, SME 3 explained the purpose of the study, introduced the attribute and Q-matrix concepts, as well as the assessment to the other two SMEs. They reviewed the example skills SME 3 obtained from the literature and test materials. SMEs were reminded that, the example skills were meant to give an idea about the possible reading attributes; however, other attributes might emerge based on the task analysis. SMEs also discussed the steps in the development process. Then, three SMEs worked on example items together to start deriving the attributes defining items. The purpose here was to ensure everybody shares a common understanding of the task. SMEs completed the analysis of the actual test items individually. The group met again to review the attributes each SME elicited. They went over each item to discuss attributes they defined and shared their interpretations of the attributes. A consensus among SMEs was sought to finalize the attribute list. Following this meeting SME 3 wrote definitions of each attribute and coding considerations based on their discussion and shared it with other SMEs. Each SME separately coded each item again for each of the attributes. A final meeting was held to discuss if revisions for the attributes, definitions, or coding considerations were necessary. SMEs compared the Q-matrices and discussed any discrepancies, along with their rationale. It was decided that one of the attributes could be broken down to ensure more specificity. Other attributes were considered sufficient. There was general consensus on most item-attribute associations. Because a new attribute was added, the group agreed to match the items with the attributes for one last time.

After the initial conceptualization of attributes were finalized, SMEs from the test developer team were involved in the process to code the Q-matrix. There was a total of 7 SMEs for the Q-matrix coding. All SMEs were provided with descriptions, coding examples, and other materials (e.g., test items, an attribute list, item and distractor analysis). SMEs were also asked to indicate the attributes that they believe were missing in the list that might be associated with items. They were also requested to provide their rationale for the choice of specific attributes for each item. This information was sought to understand their matching and resolve potential disagreement among experts. In addition, all SMEs rated their confidence for their attribute coding for each item on a scale from 1 to 5 (i.e., 1 indicating not confident and 5 indicating very confident). The researcher verified with each SME that the task was clear. Communication with SMEs was conducted via email.

Once the researcher received all the Q-matrices, they were compared to build the final Q-matrix. An attribute was coded for the item if it was selected by 4 or more SMEs (i.e., among 7 SMEs). There was not substantial disagreement and experts' and descriptions for the coding was adequate to make corrections.

Empirical Validation. The expert developed Q-matrix was empirically validated using de la Torre and Chiu's (2016) the general discrimination index (GDI). To reiterate, in this method the attribute combination that yields the highest difference in the correct response probabilities of masters and non-masters is defined to be the correct q-vector for a specific item. The proportion of variance is estimated for each combination of the attributes and compared against a criterion value. When multiple attribute combinations yield high variance, the simplest q-vector is chosen. Ma & de la Torre (2020) incorporated the method to the GDINA package which was used for the purposes of the validation in this study. Pure statistical revision is not encouraged because changes might be due to chance (Ravand, 2016). Thorough understanding of attributes and items is necessary (Jang, 2009a). Substantive knowledge should be incorporated to determine if the recommendations suggested by the empirical method are sound. Experts' coding, their written rationale and item specifications (i.e., blueprint) were relied on to make the final decisions about the attributes in this study. Furthermore, in order to explore whether the modifications are due to chance, the sample was divided into two as training sample and validation sample. The empirical validation was conducted with the training sample and it was explored whether the modifications hold for the validation sample (refer to the procedures for research question 2).

Research Question 2

The second research question was concerned with the identification of the best fitting DCM (a general vs. specific model) and Q-matrix for the data. Because an established theory for model selection (von Davier, 2014) is currently missing for language constructs (e.g., Alderson, 2005, 2007; Alderson et al., 2015), a set of different models were included in the study. The review of DCM comparison studies for L2 reading supported this decision. The evaluation of models presented mixed findings (i.e., Chapter 2, comparable performance for compensatory and non-compensatory models). The approach to incorporate various models is also a common practice and recommended in the retrofitting framework. A general model, the LCDM, along with five restricted models (i.e., the DINA, DINO, R-RUM[3], C-RUM, HO-DINA) were fit to the data using the CDM package (George et al., 2016) in R software (R Core Team, 2014). Especially the LCDM might be appropriate given inconclusive findings about how attributes affect performance as general models do not impose *a priori* rules for attribute dependencies with relationship to the probability of correct response. Response processes are complicated, more so for complex constructs such as reading (Alderson, 2000). Two attributes might contribute differently to performance for two different items (Ma et al., 2016). Yet, considering the parsimony and simplicity of interpretations (Lee & Luna-Bazaldua, 2019; Ma et al., 2016), the restricted models were involved. These models were selected from the most commonly implemented DCMs in general and for language assessments.

Because multiple models and samples (i.e., training and validation) were considered, relative and absolute fit indices were combined to evaluate model fit and to identify a suitable DCM. The combination of absolute and relative fit indices was also congruent with the literature (e.g., Kunina-Habenicht et al., 2012; Li et al., 2015; Liu et al., 2018 etc.). The previous research (e.g., Chen et al., 2013; Lei & Li, 2016) showed their performance might vary across studies. Hence, it was more reasonable to include multiple indices from both groups of indices rather than relying on one index. These included the most widely used indices in previous studies.

With respect to relative fit AIC and BIC were reported for each model. A smaller value indicates the best fitting model for both AIC and BIC. Additionally, because the

[3] It is the reduced non-compensatory RUM.

restricted models are nested models within the LCDM, likelihood ratio tests were used to evaluate whether more parsimonious restricted models fit significantly better than the saturated LCDM (e.g., Liu et al., 2018). The likelihood ratio test can be expressed as the comparison between the likelihood (-2LL) of the two models. Because it has an approximate χ^2 distribution with k degrees of freedom (i.e., difference in the number of model parameters), the difference was tested for significance. A significant test result shows a better fit for the saturated model.

Although absolute fit indices are recommended for misidentification in Q-matrices comparing them across models can contribute to model selection decisions (e.g., Kunina-Habenicht et al., 2012; Li et al., 2016). They were also needed for the study to compare fit of the training and validation samples. Six absolute fit indices presented in Table 5 were used in the study. In addition to being common indices, previous studies reported an acceptable to good performance for the selected indices for the study (Lei & Li, 2016; Li et al., 2016). Generally, as these indices approach zero, the fit improves (Kunina-Habenicht et al., 2012; Ravand, 2016). Some researchers also report or use cut-offs (Table 5) for acceptable fit that were also adopted for the study.

In addition, the following criteria (Rupp et al., 2010, p. 165) were also referred to for the item-level model selection because a general model was included in the study.

- DINA: zero main effects, positive interactions
- NC-RUM: positive main effects, positive interactions
- DINO: positive main effects, negative interactions
- C-RUM: positive main effects, zero interactions

Table 5. Absolute Fit Indices and the Criteria

Index	Explanation	Cut-off
Average RMSEA (von Davier, 2005)	The root mean square error of approximation that is the comparison of observed and estimated response probabilities (Kunina-Habenicht et al., 2012). The average RMSEA for each model is reported for model evaluation purposes.	< 0.05 (Kunina-Habenicht et al., 2012; Lei & Li, 2016; Ravand, 2016)
MADcor (Dibello et al., 2007; Henson et al., 2009)	The mean absolute difference of correlations that is the comparison of observed and estimated item correlations across items pairs	< 0.05 (Dibello et al., 2007; Jang, 2009b; Li et al., 2016; Liu et al., 2018, Ravand, 2016; Ravand & Robitzsch, 2018) =<0.06 Henson et al., 2009; Lei & Li, 2016)
MADres (McDonald & Mok, 1995)	The mean absolute difference of item residual covariances	<0.05 (Ravand, 2016; Lei & Li, 2016)
SRMSR, (Maydeu-Olivares, 2013, Maydeu-Olivares & Joe, 2014)	The standardized root mean square residuals that represents the comparison between observed and estimated correlations. More precisely it is the squared root of the mean of the squared difference between observed and expected correlations across item pairs	< 0.05 (Maydeu-Olivares, 2013; Maydeu-Olivares & Joe, 2014; Liu et al., 2018; Ravand & Robitzsch, 2018).
MADQ3 (Yen, 1984)	Mean absolute difference of Q3 values. Q3 shows the Pearson correlation of the item residuals (Christensen et al., 2017)	<0.05 (Liu et al., 2018; Lei & Li, 2016; Ravand, 2016; Ravand & Robitzsch, 2018)
$M\chi^2$ (Chen & Thissen, 1997)	A χ^2 test based on the comparison between the observed and expected response frequencies across item pairs. The maximum χ^2 of all item pairs is reported. If the maximum difference is large and significant, dependency is present.	Non-significant (Rupp et al., 2010)

Note. Some authors (Li et al. (2016); Lei & Li, 2016) suggest definite criteria are not available. The smaller the difference between what is observed in the data and what is estimated by the model, the better the fit is. The cut-off values denote what the studies used, suggested or found based on their analysis of data.

For the selection process, after the Expert-defined Q-matrix was validated with the calibration sample using the GDI, six models were fit and compared using both group of indices. Treating the best fitting model and the validated Q-matrix as the final model and Q-matrix, they were applied to the validation sample. The fit between the training and validation samples were compared using absolute fit indices. The validation sample was expected to fit reasonably well as the training sample to ensure the modifications based on the empirical validation were not capitalizing on chance, or the model was not overfitting the data. Because the empirical validation was not undertaken for the Standards-based Q-matrix, six models were fit only to the validation sample. The best fitting model for the Standards-based Q-matrix was also selected based on absolute and relative fit indices. As a final step, the final model for the Standards-based Q-matrix was compared against the final model for the Expert-defined Q-matrix using all indices to select the Q-matrix for the study.

Research Question 3

The final research question was concerned with the viability of diagnostic information after the best fitting DCM was identified. Under this research question, the characteristics of item, person, and attribute estimates were scrutinized. The goal here was to understand whether the items presented diagnostic information, and whether student classifications and attribute patterns were acceptable.

With respect to the items, the probabilities of correctly responding to each item were examined closely. Specifically, it was expected that the correct response probability is low for students who do not possess any of the attributes specified for an item. In other words, it indicates that students do not guess but master attributes to answer a specific question. For high quality items, main effects and/or interactions should be high as they indicate how much they relate to the items, or indicate potential problems such as misfit or misspecification (Templin & Hoffman, 2013). Another important measure for high diagnostic capacity items that was used in the study was discrimination power. Rupp et al. (2010) define discrimination in the DCM context as the degree to which items distinguish between two examinees groups (1) those who know more attributes and (2) those who know fewer attributes. They broadly represent discrimination index as $d_i = p_{ah} - p_{al}$ where the p_{ah} is the correct response probability for the first group who has higher ability, and p_{al} is the correct response probability the second group who has lower ability. The authors suggest that this global index can be applied by taking the difference in the correct response probabilities of examinees who mastered all required attributes (p_{ah}) and who mastered none of the required attributes (p_{al}). Thus, the index simply denotes the deviance between correct responsibility of the two groups. The authors note d_i is large for items with high discriminating power. However, this is a crude measure. The authors also acknowledge that the index overlooks other categories and only considers two groups. However, it can still be useful and provides "a general benchmark" for item quality (Rupp et al., 2010, p. 282). Therefore, it was adopted for the study. In addition to the correct response probabilities and discrimination, item fit was evaluated based on RMSEA which is an item specific fit measure. To reiterate, RMSEA values provide a comparison of the predicted and observed probabilities (Kunina-Habenicht et al., 2012). Generally, items with an RMSEA value lower than 0.05 is assumed to show a good fit (Table 5). The RMSEA for each item were presented for fit information. Using this criterion, it was explored whether or not there were particular misfitting items.

In addition, diagnostic capacity is desired, person estimates such as classifications (i.e., latent classes) and attribute estimates should demonstrate certain characteristics to determine interpretations based on DCMs are acceptable. Several estimates or measures were included for this purpose. Class probabilities which denote that the probability of observing a specific class in the population were reported. In addition to this, the proportion of ELs in each latent class were estimated. Class probability and proportions should show some variability for a successful DCM application. In other words, students should not be clumped together in certain classes. For instance, if a majority of the students are assigned to two classes where none of the attributes vs. all of the attributes are mastered respectively, it can be inferred that attributes are correlated, and the test follows a unidimensional structure (Lee & Sawaki, 2009). Therefore, rather than decomposing reading to smaller attributes for diagnostic feedback, it would be more reasonable to represent it as a broad construct. Moreover, student classifications were explored for consistency and accuracy to provide useful and effective provisions to the students. Cui et al. (2012) and Wang et al. (2015) developed consistency and accuracy indices for the DCM context that were used in the study. Classification consistency index (P_c) denotes the probability of consistent classification to a latent class for a student from a random draw if the same or parallel test is administered again. Classification accuracy index (P_a), on the other hand, expresses the probability of accurate classification to the correct latent class for a student from a random draw. Wang et al. applied these notions to attribute level. These indices were helpful when understanding whether classifications were acceptable or not. Cui et al. recommend evaluating the fit of the model before using the measures. Additionally, based on simulations they observed that indices get higher with items with a high discrimination power, attributes with more dependencies, and a smaller number of attributes. Cui et al. also suggest 0.7-0.8 for acceptable consistency and accuracy (as cited in Ravand, 2016; Ravand & Robitzsch, 2018). Both of these measures were obtained from CDM package.

In relation to the attributes, average probabilities of mastering each attribute (i.e., difficulty), the proportions of students mastering each attribute, and correlations between attributes were estimated and reported. Attribute difficulties reveal important information about the characteristics of the population. Attribute mastery probabilities were scrutinized whether they were reasonable (Dibello et al., 2007). For instance, the probability of attaining inference type skills was expected to be lower than the probability of understanding the main idea due to the complexity of the former. In order to estimate the proportion of students who mastered an attribute, certain cuts are needed. Rupp et al. (2010) suggest probabilities around 0.5 are not certain, meaning information might not be adequate. A probability higher than 0.5, on the other hand, will be an indication of the mastery. Jang (2009b) applied 0.4 as the upper bound for non-mastery and 0.6 as the lower bound for the mastery. Thus, for students with a probability range of 0.4-0.6 (i.e., uncertainty region), attribute status cannot be determined. This criterion was applied in the study to explore whether the proportion of students in the uncertainty region was small, which might hint that the test was useful to determine mastery status of the majority of the students (e.g., Jang, 2009b). The distribution of individual attributes was also checked to confirm whether the majority of the students could be successfully classified (e.g., Lee & Sawaki, 2009). Furthermore, correlations between attributes were analyzed to understand whether it was reasonable to break down the reading construct to the attributes (Templin & Hoffman, 2013). Although some level of correlation was expected between the attributes, if the attribute correlations are fairly high, it would not be practical to represent reading as a multidimensional construct. According to Kunina-Habenicht et al. (2012), correlations between subscores and subdomains in educational tests range between 0.5 and 0.8. Sessoms and Henson (2018) imply tetrachoric correlations among attributes exceeding 0.9 are high and an indication of non-distinctive attributes.

Finally, the viability of DCM was revealed by comparing how results obtained from DCM related to results obtained under the test's original measurement framework (i.e., IRT) (e.g., Liu et al., 2018). It was expected that student classifications from both methodologies were congruent with each other. If DCM classifications diverge to a great degree, it would not be meaningful to provide information obtained from DCM for instructional use. In particular, the dataset included ELs' proficiency levels estimated under Rasch model. The distribution of proficiency levels across masters and non-masters of each attribute, as well as for each latent class, were reported for this purpose. Also, ELs' ability under the original framework (θ) was estimated. For each latent class and individual attributes, θ distribution was reported for the relationship between classes/mastery and ability under the unidimensional model. It was hypothesized that an EL's proficiency level and θ was higher for masters than non-masters and their ability increased as students master more attributes.

CHAPTER IV

RESULTS

Given the continuous interest in diagnostic information and the use of DCMs for large-scale assessments to address these demands (Chapter 1), this study undertook a DCM methodology for a K-12 language assessment for low-stakes diagnostic feedback. The study aimed to respond to the following research questions:

(1) What are key underlying attributes represented in the ACCESS reading domain in middle grades for more advanced ELs?

(2) What DCM fits the data better?

   a. Does a general or specific restricted model better represent all items in the test?

   b. Does a Standard-based or an Expert-defined Q-matrix show better fit?

(3) To what extent is it feasible to obtain diagnostic information using DCM?

   a. What is the diagnostic capacity of the test items?

   b. To what extent can students be appropriately classified using the model?

In Chapter 2 different phases of DCM methodology from, Q-matrix development, model selection and evaluation were reviewed. The study incorporated two Q-matrices (i.e., based on the standards vs. expert input), several models (e.g., LCDM, DINA, DINO, R-RUM, C-RUM, HO-DINA), and validation strategies for Q-matrix and model selection. The feasibility of the application was evaluated form several aspects such as accuracy and consistency of classes and attribute mastery as well as properties of the attributes and the items. This chapter focuses on the findings from the DCM implementation. It is divided into four main sections. It starts with a summary of classical item and test analysis. The second section elaborates on the two Q-matrices employed in the study. The third section presents comparisons among several models and the two Q-matrices for the final model and Q-matrix selection. The fourth section evaluates the final model for the viability of using the DCM by describing item parameter estimates, item discrimination index, class probabilities and proportions, attribute correlations, accuracy and consistency of the profile, and individual attributes. The chapter ends by discussing the relationship of these findings to the test's original framework.

Classical Item and Test Statistics

In order to establish a basic understanding of the test and the items, overall performance based on raw scores, as well as CTT statistics, were examined first. The mean reading score was 13.9 out of 27 points with a standard deviation of 4.83. Raw scores were normally distributed and the majority of the ELs showed an average performance (Histogram in Appendix F, Figure 1). Cronbach's alpha that is the internal consistency index for the test (Crocker & Algina, 2008) was 0.77. Alpha shows the degree to which items are interrelated, which is required for unidimensionality (Schmitt, 1996). The fact that items were not perfectly interrelated was desired for the study as the objective was to divide the reading construct into several attributes.

Table 6. Test Descriptive Statistics and Reliability

N	Mean	sd	Min.	Max.	Skew.	Kurt.	Alpha	SEM
23, 942	13.854	4.834	0	27	0.169	2.491	0.768	2.33

Note. sd = standard deviation, Skew = Skewness, Kurt = Kurtosis, Alpha = Cronbach's Alpha, SEM = Standard Error of Measurement

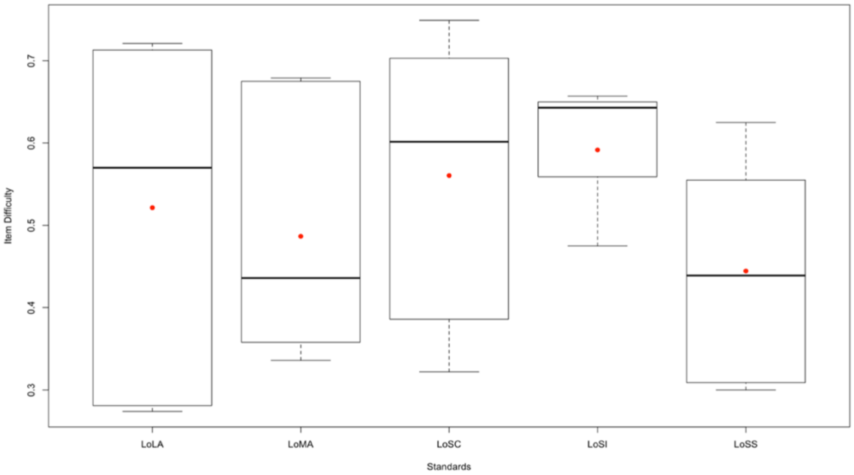
Table 7. Summary of the CTT Item Statistics

	Mean	sd	Median	Minimum	Maximum
p-value	0.513	0.160	0.529	0.274	0.749
point biseral	0.289	0.082	0.274	0.111	0.427

With respect to the item statistics, p-value (item difficulty index) and point biserial (item discrimination index) were computed and a distractor analysis was conducted using the CTT package (Willse, 2018). The summary of the item statistics is presented in Table 7 and the full statistics can be found in Appendix F (Table 1). The average difficulty of the items was 0.51. None of the items were too easy as the maximum p-value was 0.75 (Item 12). Seven items (9, 15, 17, 18, 23, 26, and 27) were relatively hard as 35% or fewer ELs correctly responded to these items. Items 17 and 18 were the hardest items of the test with a p-value of 0.27 and 0.28, respectively. The point biserial value of the items ranged from 0.11 to 0.43, with an average of 0.29. Bachman (2004) considers point biserial of 0.30 and above, while Henning (1987) is less conservative and recommends at least 0.25 for well discriminating items on a language test. In this respect, there were several poor discriminating items such as 9, 10, and 27 with point biserial values less than 0.20.

The difficulty of the items was also examined across the five standards as shown in Figure 2. To reiterate, standards represent the ability to use English in social instructional settings and in relation four content areas: language of arts, math, science and social studies. Items related with social instructional language were easier than others. The spread of the items was wider for the other four subject areas, and they included both easy and hard items. Items covering social studies and math were relatively harder, on average, compared to other areas.

Figure 2. Item Difficulty across the Five Standards



Note. LoLA= Language related with language arts, LoMA= Language related with math, LoSC= Language related with science, LoSI= Social instructional language, LoSS= Language related with social studies. Red dots represent the means.

Attributes and Q-matrices

The Standards-based Q-Matrix

Two alternative Q-matrices were developed for this study in search of an optimal Q-matrix fitting the data. The first Q-matrix, referred to as the Standard-based Q-matrix, consisted of two broad dimensions. The standards and key uses of language that the test was built to measure were treated as the attributes. The standard for each question that pertains one of the five areas listed above were extracted from the test blueprints and coded in the Q-matrix. Because such linking was missing for the key uses, the researcher (SME 3) mapped the key uses to items which was then reviewed by the two SMEs (SME 4 & 6) from the test developer team. In brief, the key uses were operationalized as:

- Recount: identifying, retelling, or summarizing details
- Explain: understanding processes/cycles, relationship between concepts/ideas, consequences
- Argue: understanding judgements, hypotheses, claims, or evidence

The ratings of all three SMEs are presented in Table 8. As evident from the table, there was almost perfect agreement among SMEs. There were only differences with respect to four items. SME 4 and SME 3 associated item 7 and item 16 with Explain. Because the other two raters agreed that the item was related with Recount, the final coding for these items was Recount. SME 6 specified additional attributes for two items. The rater mentioned some students may also benefit from Recount and Argue in answering items 10 and 26, respectively. Other SMEs mapped a single key use to items, which was maintained for this study. The two SMEs from the test developer team also commented that the researcher's key use-item mapping was reasonable and justifiable. Therefore, the Standard-based Q-matrix was finalized as shown in Table 9.

In the final matrix, except for the Social Instructional Language, all other attributes in the standards dimension were associated with six items. Only the three items were related with the Social Instructional Language in the form. For the key use dimension, Recount was measured 10 times, Explain 13 times, and Argue only 4 times. Each attribute from the standards dimension were measured together with an attribute from the key use dimension except for Recount-Science, Explain-Social and Instructional Language, Argue-Math. This was expected as Argument is about opinions, while math items were concerned with reading passages about math concepts. Similarly, Science items were about reading passages related to cycles or hypotheses that aligned better with either the Explain or Argue attributes.

Table 8. The Mapping of the Key Uses to Test Items

Items	SME 3	SME 4	SME 6
1	Recount	Recount	Recount
2	Recount	Recount	Recount
3	Argue	Argue	Argue
4	Recount	Recount	Recount
5	Recount	Recount	Recount
6	Argue	Argue	Argue
7	Recount	**Explain**	Recount
8	Explain	Explain	Explain
9	Recount	Recount	Recount
10	Explain	Explain	Explain, **Recount**
11	Explain	Explain	Explain
12	Explain	Explain	Explain
13	Recount	Recount	Recount
14	Explain	Explain	Explain
15	Argue	Argue	Argue
16	**Explain**	Recount	Recount
17	Explain	Explain	Explain
18	Explain	Explain	Explain
19	Recount	Recount	Recount
20	Explain	Explain	Explain
21	Explain	Explain	Explain
22	Explain	Explain	Explain
23	Explain	Explain	Explain
24	Argue	Argue	Argue
25	Recount	Recount	Recount
26	Explain	Explain	Explain, **Argue**
27	Explain	Explain	Explain

Note. Bold indicates the rating was different than other SMEs. An example for the complete mapping for the Standards-based matrix could not be provided in the study because it included actual descriptions from the blueprints, which are confidential information.

Table 9. The Final Q-matrix Based on the Standards and Key Uses

	Attributes							
Items	LoSI	LoMA	LoLA	LoSC	LoSS	Recount	Explain	Argue
1	1	0	0	0	0	1	0	0
2	1	0	0	0	0	1	0	0
3	1	0	0	0	0	0	0	1
4	0	0	1	0	0	1	0	0
5	0	0	1	0	0	1	0	0
6	0	0	1	0	0	0	0	1
7	0	1	0	0	0	1	0	0
8	0	1	0	0	0	0	1	0
9	0	1	0	0	0	1	0	0
10	0	0	0	1	0	0	1	0
11	0	0	0	1	0	0	1	0
12	0	0	0	1	0	0	1	0
13	0	0	0	0	1	1	0	0
14	0	0	0	0	1	0	1	0
15	0	0	0	0	1	0	0	1
16	0	0	1	0	0	1	0	0
17	0	0	1	0	0	0	1	0
18	0	0	1	0	0	0	1	0
19	0	1	0	0	0	1	0	0
20	0	1	0	0	0	0	1	0
21	0	1	0	0	0	0	1	0
22	0	0	0	1	0	0	1	0
23	0	0	0	1	0	0	1	0
24	0	0	0	1	0	0	0	1
25	0	0	0	0	1	1	0	0
26	0	0	0	0	1	0	1	0
27	0	0	0	0	1	0	1	0

Note. LoSI= Social instructional language, LoMA= Language related with math, LoLA= Language related with language arts, LoSC= Language related with science, LoSS= Language related with social studies

Table 10. The Number of Items per Attribute in the Standards-based Q-matrix

		Standards Dimension					
		LoSI	LoMA	LoLA	LoSC	LoSS	Total
Key Use Dimension	Recount	2	3	3	-	2	*10*
	Explain	-	3	2	5	3	*13*
	Argue	1	-	1	1	1	*4*
	Total	*3*	*6*	*6*	*6*	*6*	

The Expert-defined Q-Matrix

The Expert-defined Q-matrix was developed by two groups of SMEs. Three SMEs (SME 1, 2, & 3) established the attributes necessary to complete the items on the test, after which all SMEs, including the four working for the test developer, matched attributes to items. After reviewing the list of common L2 reading attributes and content of the items, there was a consensus that the items on the test aligned with the attributes in other DCM studies for reading construct. A total of seven attributes emerged (Table 11).

Table 11. Attributes in the Expert-defined Q-matrix

Attribute	**Explanation**[4]
Vocabulary (VOC)	Understanding the key words/phrases in the text dependent or independent of the context. The attribute also entails recognition and knowledge of synonyms, antonyms, and the association between similar words in the text and answer choices (i.e., paraphrase).
Cultural and Conceptual References (CUL)	Understanding the idea of concept. The attribute is closely related with the vocabulary attribute, but it requires knowledge of the "extended meanings". Just knowing the meaning of the words might not be adequate and it might require understanding at a conceptual level. Some concepts might be rooted in the culture and may be unfamiliar to a student from a different culture (e.g., community service, leadership training) (Bachman, 1990, p.97).
Grammar (GRM)	Understanding and processing complex sentences (e.g., relative clauses), and compound clauses including numerous grammatical and cohesive devices such as conjunctions. The attribute involves recognizing pronoun references.
Explicit Information and Details (EXP)	Deriving and comprehending explicit important information and details from the text. The attribute involves scanning the text and finding the details, and/or matching (i.e., answer choice and sentence in the text).
Inference (INF)	Comprehending information by making inferences. The information is implicit or overtly stated in the text. For example, the attribute requires connecting information in the text with an example situation.

[4] When describing the attributes, the definitions in Jang (2009a), Li and Suen (2013), and Sawaki et al. (2009) were benefited due to the similarity of the attributes.

Summary and Synthesis (SUM)	Connecting and integrating information across adjacent sentences or parts of the text (e.g., paragraphs, charts). The attribute entails summarizing, understanding the gist of the paragraphs, or interpreting rhetorical relations (e.g., problem-solution).
Sequences and Processes (SEQ)	Understanding sequential language, steps or order in a process or cycle. Information presented includes description of a sequence/steps and/or sequential language (e.g., first, second, eventually) that needs to be processed for a correct response.

Attribute l, Vocabulary, is related with understanding the key words and phrases, as well as recognizing synonyms and paraphrase. Because vocabulary knowledge might superficially apply to most items on a language test, EL's grade and proficiency level would determine selecting the attribute for an item. The attribute was intended specifically for items requiring knowledge of difficult, content-specific, technical vocabulary. Attribute 2, Cultural and Conceptual References, is tightly connected with Vocabulary. However, it involves an understanding of the extended meaning of some concepts which might be culture specific. This attribute relates to Bachman's (1990) sociolinguistic competence (i.e., cultural references and figures of speech). For example, a student might understand the individual words in "community service", which might not be adequate. Being familiar with the phrase at the conceptual level is essential. Also, community service is commonly practiced in the U.S., however, it might be unfamiliar to some newcomer ELs in whose country this practice is not widespread. Attribute 3, Grammar, entails processing compound sentences. This includes understanding pronoun references, conjunctions and other cohesive elements. Like the Vocabulary attribute, Grammar might be applicable to most items because baseline Grammar knowledge is necessary to comprehend texts. This attribute is considered when extracting meaning from sentence structure is deemed necessary. Attribute 4, Explicit Information and Details involves deriving specific details from the text and comprehending explicit information. It requires scanning the text for transparent details and matching it with the correct answer. Attribute 5, Inference, requires understanding implicit information and making inferences. Given the grade level of ELs, inferences can be low level. For example, students might need to associate the information with an example situation. What distinguishes this attribute from the previous one is the transparency of the information. Attribute 6, Summary, represents integrating information from adjacent sentences or different parts of the text (i.e., across paragraphs or cells of a chart), to make meaning. In some situations, this attribute requires understanding the gist, summary, or rhetorical relations. Finally, Attribute 7, Sequences and Processes, is related to understanding the description of processes, cycles, and sequential language.

The seven attributes previously stated were deemed sufficient to represent the items on the test, as SMEs did not raise concerns for the attributes presented to them or suggest any additional attributes. One exception was for Cultural and Conceptual References. Two of the SMEs pointed that this attribute would apply to very few items and the test did not rely on culture-specific background knowledge. SME 2 also raised the point about its similarity to idiomatic language that is part of vocabulary knowledge. As a native speaker, they also mentioned judging whether concepts are culture-specific was arduous. The remaining attributes concurred with the skill and task descriptions in the blueprints. In the blueprints, vocabulary requirement ranged from general to specialized, technical vocabulary for items. In a similar vein, some items and their reading stimuli were prescribed to include simple sentences and modifiers, while others contain complex sentences with multiple clauses, and a variety of modifiers. Other attributes also appeared directly or indirectly in the item specifications as also aligned with performance descriptions (WIDA Consortium, 2012). For example, items were related with identifying, inferring, interpreting, predicting, summarizing, sequencing characteristics, or details from the reading texts. It is worth noting that three SMEs did not see the blueprint descriptions while specifying the attributes and they were used by the researcher to confirm the attributes *ad hoc*.

After attributes were fleshed out, all SMEs reviewed items and coded attributes for items to come up with a Q-matrix. They also provided their rationale and confidence in their rating of each item. An example Q-matrix can be found in the Appendix E. To begin with, SMEs were confident in their item-attribute rating (4 out of 5 on average). This can hint at the robustness, as well as clarity of attributes and the overall task. It was observed that SMEs were less certain about their coding when an additional attribute needed to be defined for an item. Variability in their rating confidence was greater for items measuring language of math. SMEs were also less assured when coding items related to language of social studies and math (Appendix F, Figure 2).

Table. 12. Confidence Ratings for Attribute-Item Mapping

	SME 1	SME 2	SME 3	SME 4	SME 5	SME 6	SME 7	Items
Mean	4.22	4.04	3.48	4.22	3.56	4.00	4.63	4.02
sd	0.89	0.85	1.25	0.80	0.93	0.48	0.56	0.35

Table 13 shows the mapping of attributes among 7 SMEs. Each cell indicates the number of SMEs selecting an attribute for an item. As the table illustrates, 5 or more SMEs selected the same attribute for an item in most cases. When an additional attribute is needed, four of the SMES also agreed on the additional attributes. Some variability occurred, as anticipated, due to the complexity of the reading construct. Attributes being specified by four or more raters were included in the initial Q-matrix (Table 14). It must be noted that SMEs' ratings had been modified when necessary, based on the descriptions for their coding. Mainly, if an SME explicitly stated they hesitated to select the attribute, their rating was updated to 0. If the description of a rationale for selecting attributes contradicted with the selection (i.e., or the attribute was missing despite being mentioned), the rating was corrected as well. For instance, SME 4 expressed that the stem of Item 12 was worded as synthesis, yet it was actually testing the understanding of a specific sentence. Nevertheless, the SME still marked summary for the item, which was corrected to reflect Explicit Information. These modifications did not change the initial Q-matrix, except for 3 items. Vocabulary was added to Item 7 (i.e., originally marked by 3 SMEs) because SME 6 and 7 mentioned the attribute in their description. Being less conservative about attributes was preferred, at this stage, as the Q-matrix would be validated. On the other hand, this attribute was deleted from Item 21 as one of the SMEs aroused doubt selecting it. Finally, for Item 25, SME 4's selection was updated to Summary from Explicit Information because the rater commented that the stem requires understanding the whole paragraph. This change was also supported by the blueprint, as the item was depicted as the summary of a situation. Furthermore, Conceptual and

Cultural references were excluded from the matrix. It was associated with only two items. SME's rationale for selecting this attribute also varied. Hartz et al. (2002) recommends keeping an attribute associated with at least 3 items (in Kim, 2015, p. 237). Even if it was related with more items, it could be merged with Vocabulary, because they always occurred together.

Table 13. Attribute-Item Mapping for the Expert-defined Q-matrix

Items	VOC	CUL	GRM	EXP	INF	SUM	SEQ
1	2	0	1	**5**	1	2	1
2	0	0	1	**6**	**4**	0	0
3	**7**	0	1	**6**	1	1	0
4	0	0	3	**7**	0	0	1
5	**4**	4	0	2	0	**7**	0
6	**6**	3	0	**6**	1	0	1
7	**5***	1	1	**7**	1	0	0
8	0	0	0	**5**	3	1	**5**
9	1	0	**4**	**5**	1	3	**5**
10	1	0	2	3	0	0	**7**
11	3	0	0	2	**4**	**5**	0
12	**4**	0	0	**7**	0	2	0
13	**6**	0	**4**	2	2	1	0
14	0	0	0	1	0	**6**	1
15	**4**	0	**4**	2	**7**	0	0
16	1	1	0	**7**	0	0	**7**
17	2	2	0	2	0	**5**	**6**
18	1	0	0	**4**	**6**	**4**	0
19	0	1	0	**6**	0	1	0
20	0	0	2	**5**	1	1	**6**
21	3*	0	0	3	**4**	**4**	0
22	0	0	0	**7**	1	1	**4**
23	**4**	0	0	**5**	0	**4**	0
24	1	1	0	2	**7**	2	0
25	**6**	4	0	2	1	**4***	0
26	**6**	3	2	3	2	2	0
27	**6**	2	0	2	**5**	2	0

Note. * denotes the refined items based on SME explanations. Attributes selected by 4+ raters are bolded. VOC= Vocabulary, CUL= Cultural/Conceptual References, GRM= Grammar, EXP= Explicit Information and Details, INF= Inference, SUM= Summary, SEQ= Sequences and Process.

Table 14. The Initial Expert-defined Q-matrix

Items	VOC	GRM	EXP	INF	SUM	SEQ
1	0	0	1	0	0	0
2	0	0	1	1	0	0
3	1	0	1	0	0	0
4	0	0	1	0	0	0
5	1	0	0	0	1	0
6	1	0	1	0	0	0
7	1	0	1	0	0	0
8	0	0	1	0	0	1
9	0	1	1	0	0	1
10	0	0	0	0	0	1
11	0	0	0	1	1	0
12	1	0	1	0	0	0
13	1	1	0	0	0	0
14	0	0	0	0	1	0
15	1	1	0	1	0	0
16	0	0	1	0	0	1
17	0	0	0	0	1	1
18	0	0	1	1	1	0
19	0	0	1	0	0	0
20	0	0	1	0	0	1
21	0	0	0	1	1	0
22	0	0	1	0	0	1
23	1	0	1	0	1	0
24	0	0	0	1	0	0
25	1	0	0	0	1	0
26	1	0	0	0	0	0
27	1	0	0	1	0	0
Total	*11*	*3*	*15*	*7*	*8*	*7*

Note. VOC= Vocabulary, GRM= Grammar, EXP= Explicit Information and Details, INF= Inference, SUM= Summary, SEQ= Sequences and Process.

The initial Q-matrix was finalized as shown in Table 14. It comprised of six attributes. Grammar was the least frequent attribute (i.e., 3 items), while Vocabulary and Explicit Information were the most frequent attributes. 7 items had a simple structure, meaning they were associated with a single attribute. Except for Grammar, all attributes were measured alone at least one time. The remaining twenty items were complex. They were related with two attributes except for Items 9, 15, 18, and 23, which were associated with 3 attributes. The attributes were coupled with each other at least one time, except Vocabulary-Sequences and Inference-Sequences. In summary, the structure of the initial Q-matrix was fair despite that attributes were retrofitted.

Agreement Rate among SMEs for the Expert-defined Q-matrix Coding

Due to the participation of multiple raters in the coding of the Expert-defined Q-matrix, the variability and similarity of their ratings was examined. Fleiss et al. (2003) suggest that when there is sufficient agreement among raters, their ratings reflect the actual dimensions. On the other hand, substantial disagreement would render findings undependable. Fleiss Kappa, which is an interrater agreement index for categorical ratings among multiple raters, is used for this purpose. It is analogous to interclass correlation coefficient, which is used for the ratings on the continuous scale. Landis and Koch (1977) provide six categories for agreement rate shown in Table 15.

The agreement rate for the individual attributes ranged between 0.22 and 0.66, meaning there was fair to substantial agreement. There was more variability with respect to the selection of Grammar and Extracting Explicit Information, yet raters substantially agreed on Sequencing. The variability was projected as the SME group was large and no overall group discussion was held. Kappa was also computed for the SME 1, 2, 3 who established the attributes and discussed their mapping twice. There was substantial agreement among the three of them for all attributes except for Extracting Explicit Information and Summary, for which they moderately agreed on. These findings demonstrate that they most likely benefited from their small group discussion. On the other hand, the agreement rate among the test developer group was much lower (Appendix F, Table 2).

Table 15. Attribute-level Agreement Rate among SMEs

	All SMEs (N= 7)			SME 1, 2, 3		
	Fleiss Kappa	z statistic	p-value	Fleiss Kappa	z statistic	p-value
VOC	0.381	9.074	0	0.604	5.44	0
GRM	0.216	5.151	0	0.777	6.99	0
EXP	0.234	5.564	0	0.54	4.858	0
INF	0.416	9.914	0	0.673	6.053	0
SUM	0.287	6.832	0	0.533	4.794	0
SEQ	0.664	15.817	0	0.871	7.843	0
Average	0.366	-	-	0.666	-	-

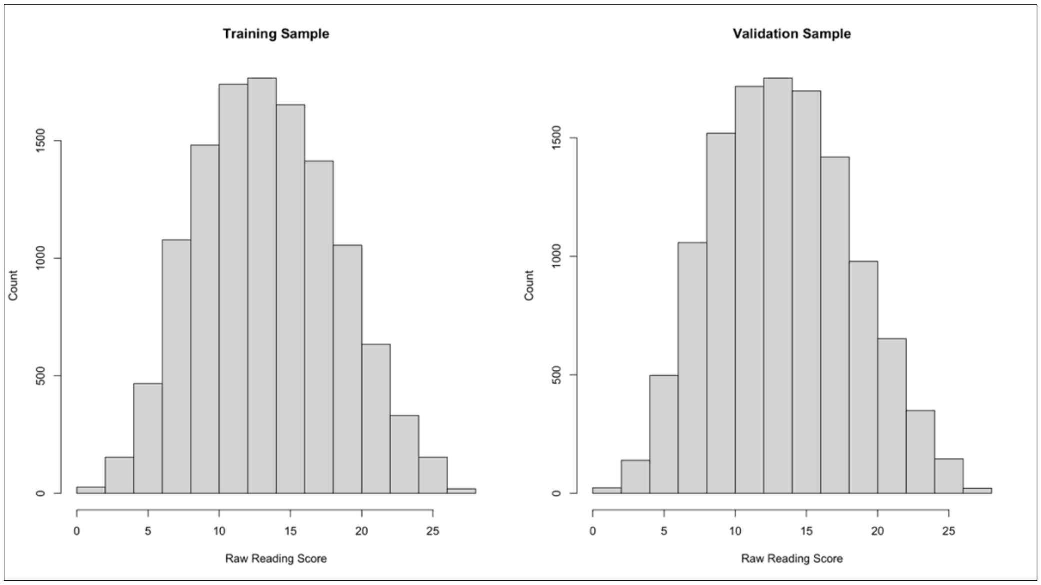
Note. <0.00 = Poor, < 0.20 = Slight, 0.21- 0.40 = Fair, 0.41 - 0.60 = Moderate, 0.61-0.80 = Substantial, 0.81 -.1.00 = Perfect (Landis & Koch, 1977, p. 165). VOC= Vocabulary, GRM= Grammar, EXP= Explicit Information and Details, INF= Inference, SUM= Summary, SEQ= Sequences and Process.

Empirical Validation and the Final Expert-defined Q-matrix

Upon creating the initial Expert-defined Q-matrix, it was validated with the GDI method. However, to ensure the modifications were not due to chance, the sample was divided into two equal samples as training and validation (N = 11, 971 for each). The GDI, as well as the model selection, was carried out with the training sample. The validated Q-matrix and the final model was then applied to the validation sample to gauge whether the changes and selected model would hold[5]. The training and validation samples were obtained by random sampling. Figure 3 shows they were similar to each other with respect to score distributions.

[5] Additionally, the replication approach was tested. When the GDI was applied to the validation sample, the changes recommended were the same as the training sample, further supporting the recommendations were not due to chance.

Figure 3. Raw Score Distribution for the Training and Validation Samples



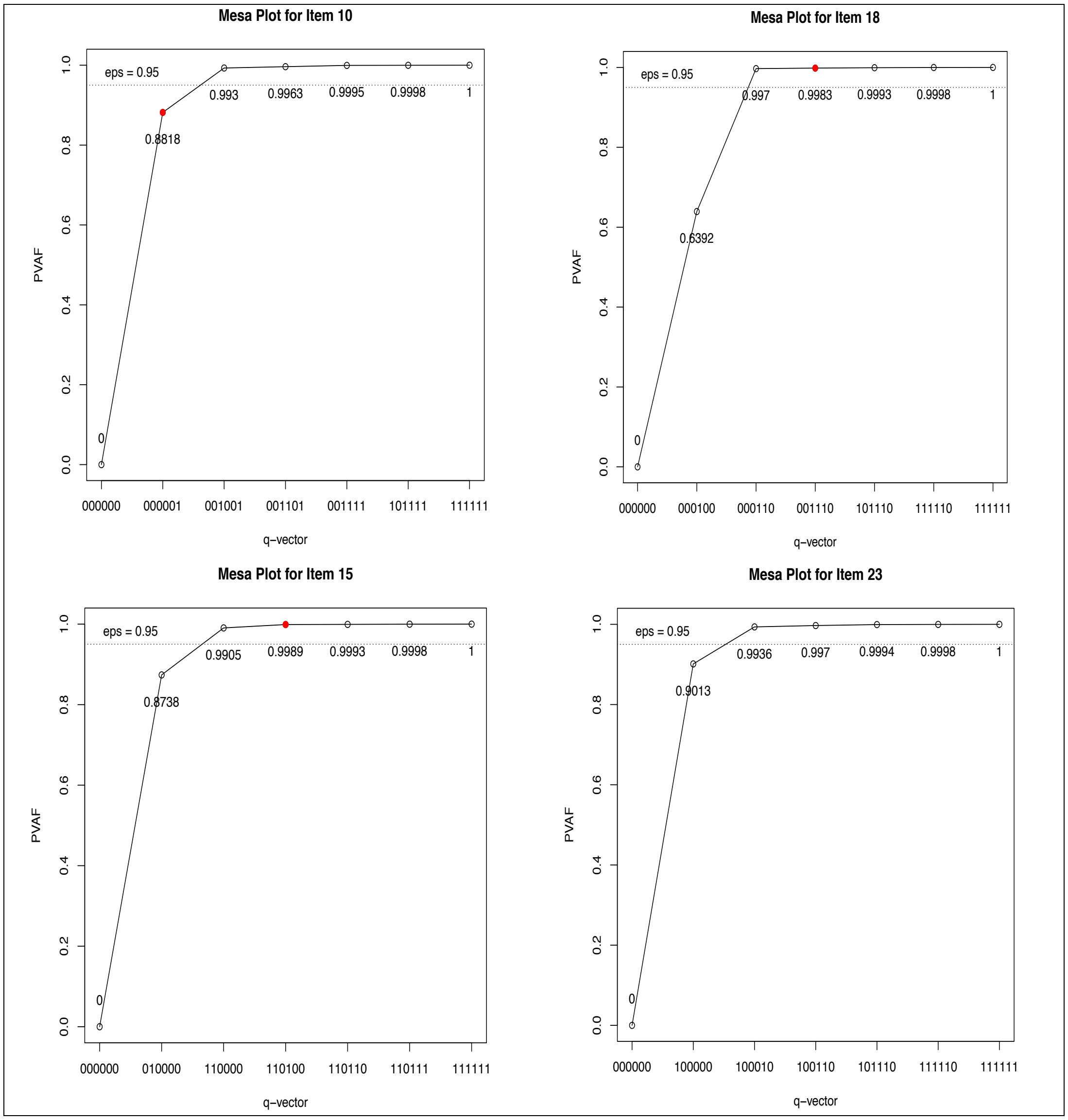
The GDI method helps identifying the simplest q-vector that yields the highest variance between the non-masters and masters of the attributes. de la Torre and Chen (2016) recommends the selected q-vector to account 95% of the variance between masters and non-masters ($\epsilon = 0.95$). In a simulation study, the authors report that it can correct 74% of the mis-specified q-vectors for the GDINA model. At the attribute level, the success of the correction was 80%, which is a fairly high proportion. However, Nájera et al. (2019) caution against the choice of ϵ, as it might itself lead to misspecifications during validation. Their simulation study shows the number of items and attributes, sample size, and item discrimination interact with the method. Hence, default ϵ might not be suitable for shorter tests, poor discriminating items, small samples, or large number of attributes or combination of these conditions. For conditions similar to those in this study (i.e., J = 30-60, N > 2000, low item discrimination) an $\epsilon = 0.85$ is recommended. In the light of their findings, both the default and recommended ϵ was applied in this study. When using $\epsilon = 0.85$ the changes were consistent with the default but there were additional changes for four items (Appendix F, Figure 3). Nevertheless, those additional changes were not theoretically applicable and not considered for the study. Below, results for the default ϵ are presented. Modifications were recommended for items 10, 15, 18, and 23. In one instance, the q-vector had been underspecified (Item 10), whereas the other cases were overspecification (i.e., more attributes than necessary were defined).

Mesa plots developed by de la Torre and Ma (as cited in Ma, 2019) to accompany the method are displayed in Figure 4. In these plots, PVAF (y axis) was plotted for different attribute combinations on the x-axis, which is ordered (i.e., PVAF increases as more attributes are added) (Ma, 2019; Nájera et al., 2019). The original q-vector specified by the experts was marked with a red dot on the plot.

For instance, for Item 15, 3 attributes were specified initially (i.e. Attribute 1, 2, 4). Among these Attribute 2 was relevant but not enough, as there was a noticeable leap when Attribute 1 was also specified. However, adding Attribute 4 did not add above and beyond, and the plateau effect was evident. Thus, specifying Attribute 4 might not be necessary and the GDI suggested omitting it. A similar pattern is apparent for Item 18 and 23. As depicted in the graphs for both items, specifying Attribute 3 did not contribute to PVAF much. For Item 18, Attributes 4 and 5, and for Item 23, Attributes 1 and 5 met PVAF > 0.95 and two attributes were sufficient. A different pattern emerged for Item 10.

Attribute 6 was defined for the item primarily. However, adding Attribute 3 would increase PVAF by 0.11 points and was tenable based on the GDI.

Figure 4. Mesa Plots of the Four Items Flagged by the GDI



Because GDI is a statistical method, it is possible to create a plausible Q-matrix based on statistical evidence, which might not be necessarily tenable, from the theoretical perspective. Therefore, the content of the items, their blueprint descriptions as well as experts' initial ratings and rationale, were reviewed again to determine the plausibility of the suggestions.

Of the recommended changes, only two were applicable. For Item 15, deleting Attribute 4 (Inference) was not reasonable because all SMEs matched it to the item. The item blueprint also evinced that the item was designed to assess inferencing skill. For Item 10, it was not plausible to add Attribute 3 (Explicit Information) as the attribute would be specious. 3 SMEs perceived it could be associated with the item. However, the item required understanding of the whole process rather than specific details of the process. The blueprint description also confirmed the decision (i.e., sequence *sentences*). On the other hand, omission of Explicit Information from Item 18 and 23 could be supported. For Item 18, SME 6 marked both Explicit Information and Summary and their explanation was inferring details by synthesizing information. This explanation shows they recognized the information is across parts of the text and not just related to one specific detail. SME 7, who also marked the item as Explicit Information, alluded to synthesis in the description. They hinted that ELs need to connect information in different parts of the chart. The item blueprint also implied Inference and Summary only (i.e. prediction and summary). Likewise, the same SMEs reasoned Item 23 requires synthesis as ELs need to locate multiple sentences and connect them to arrive at the correct solution for the item. This description voids their selection of Explicit Information. In the blueprint, Item 23 was described as being related to part of a large process, which conveys that ELs need to comprehend the whole process.

Table 16. The Final Expert-defined Q-matrix after Validation

Items	VOC	GRM	EXP	INF	SUM	SEQ
1	0	0	1	0	0	0
2	0	0	1	1	0	0
3	1	0	1	0	0	0
4	0	0	1	0	0	0
5	1	0	0	0	1	0
6	1	0	1	0	0	0
7	1	0	1	0	0	0
8	0	0	1	0	0	1
9	0	1	1	0	0	1
10	0	0	0	0	0	1
11	0	0	0	1	1	0
12	1	0	1	0	0	0
13	1	1	0	0	0	0
14	0	0	0	0	1	0
15	1	1	0	1	0	0
16	0	0	1	0	0	1
17	0	0	0	0	1	1
*18	0	0	0	1	1	0
19	0	0	1	0	0	0
20	0	0	1	0	0	1
21	0	0	0	1	1	0
22	0	0	1	0	0	1
*23	1	0	0	0	1	0
24	0	0	0	1	0	0
25	1	0	0	0	1	0
26	1	0	0	0	0	0
27	1	0	0	1	0	0
Total	*11*	*3*	*13*	*7*	*8*	*7*

Note. * denotes the refined items based on the GDI. VOC= Vocabulary, GRM= Grammar, EXP= Explicit Information and Details, INF= Inference, SUM= Summary, SEQ= Sequences and Process.

After applying the two changes (Items 18, 23) also supported by substantive evidence (i.e., expert rationale and blueprint descriptions), the GDI was applied again to the data. However, it did not suggest further recommendations beyond the modifications in the first step. Thus, the Q-matrix was finalized as presented in Table 16. The composition was very similar to the initial Q-matrix that was previously depicted. After the changes were made, Explicit Information was measured by 13 items instead of 15. Different from the initial Q-matrix, only 2 items instead of 4 items, were associated with 3 attributes.

The Model and Q-matrix Selection

Comparison of Models for the Expert-defined Q-matrix

After the Q-matrix was refined, based on the GDI with the training sample, the saturated model, the LCDM, and five specific models were fit to the training sample for model selection. The specific models were also fit within the LCDM framework (i.e., placing constraints on the general model). A caveat about the CDM package should be noted. The monotonicity constraint that ensures a higher probability for mastery of the additional attributes was specified for all models but the R-RUM. Setting the constraint, a model fits with a logit link. The CDM package fits the R-RUM through ACDM such that link function is log. The constraint for the R-RUM overwrites the link function and the model becomes the C-RUM (i.e., also fit through ACDM, link = logit). Therefore, the monotonic constraint was omitted for this model.

Relative fit statistics (Table 17) demonstrated that log likelihood (LL) and AIC were the lowest for the LCDM, which signifies a better fit. The C-RUM had the second lowest LL and AIC followed by the R-RUM. On the contrary, the C-RUM outperformed the LCDM and the R-RUM respectively, based on BIC, and CAIC. Namely, LL and AIC favored the LCDM and BIC and CAIC selected the C-RUM as the best fitting model. This shifting pattern was not unforeseen. Previous studies show that AIC picks up more complex models like the LCDM and GDINA, whereas BIC and CAIC select more constrained models as the true model (e.g., Chen et al, 2016; Henson et al., 2009; Kunina-Habenicht et al., 2012; Lei & Li, 2016). One reason for this was the larger penalty by BIC for more parameters (Kunina-Habenicht et al., 2012). Additionally, when a specific model fits as well as the saturated model, it might suggest that the specific model can be partly correct because it is the true model for some of the items (Chen et al., 2013). Relative fit indices also showed that the DINA, the DINO or the HO-DINA fit comparatively worse. Another sign for poor fit for the DINA, the DINO and the HO-DINA models were the number of items at the monotonicity boundary. 20 out of 27 items were at the boundary which implies that correct response probabilities were lower for mastering additional attributes for these items. Yet, there were no items at the boundary when the C-RUM and the R-RUM were fit. 5 items were also at the boundary in the LCDM. When these items were reviewed, they were all non-compensatory items, which absolved the concern.

Specific models were compared against the LCDM due to their nested nature (Table 18). Although the C-RUM was an acceptable model using BIC and CAIC, the LCDM fit significantly better than the C-RUM and all other constrained models.

Table 17. Relative Fit Indices for the Expert-defined Q-matrix

	Npar.	-2LL	AIC	BIC	CAIC
LCDM	124	**-195802.26**	**391852.511**	392768.901	392892.901
R-RUM	98	-195956.43	392108.859	392833.102	392931.102
C-RUM	98	-195910.85	392017.697	**392741.94**	**392839.94**
DINO	76	-196688.1	393528.199	394089.857	394165.857
DINA	76	-196702.84	393557.671	394119.33	394195.33
HO-DINA	66	-196894.37	393920.734	394408.49	394474.49

Note. Results are based on the training sample. Npar = number of parameters. Bold cells indicate the lower value which signifies the better model fit.

Table 18. The Likelihood Ratio Tests for the Expert-defined Q-matrix

Model 1	Model 2	χ^2	df	p
LCDM	R-RUM	154.174	26	<.00001
	C-RUM	108.593	26	<.00001
	DINO	885.844	48	<.00001
	DINA	900.580	48	<.00001

Note. Results are based on the training sample.

Table 19. Absolute Fit Indices for the Expert-defined Q-matrix

	$M\chi^2$	MADcor	SRMSR	MADres	MADQ3	RMSEA
LCDM	121.345	**0.0167**	**0.0231**	**0.3749**	0.0215	**0.021**
R-RUM	137.8477	0.0183	0.0248	0.41	0.02	0.02
C-RUM	131.182	0.0174	0.0238	0.39	0.0212	0.021
DINO	143.8516	0.022	0.0281	0.4943	0.0176	0.023
DINA	130.5527	0.0224	0.0287	0.5029	0.0178	0.022
HO-DINA	167.935	0.0243	0.0311	0.5524	0.018	0.02

Note. Note. Results are based on the training sample. Mχ^2 = Maximum χ^2 among item pairs. For all other indices <0.05 shows good fit.

Absolute fit indices did not provide any counter evidence. Differences across models were not dramatic and all models almost fit equally well. MADres values were higher, which might be attributed to the design of the test. Some items were based on the same text in the test or they were related to the same topic. Thus, there were some dependencies among the items that MADres was picking up on. In other words, residual dependency was high based on MADres. Similarly, despite that maximum χ^2 is expected to be non-significant (Rupp et al., 2010), it was significant across all models. It was speculated that there were misfitting item pairs. However, χ^2 tests are sensitive to sample size. A very large sample (N = 11,971) was used in this application. According to Rupp et al. (2010), extremely small p-values are anticipated because the statistic is susceptible to even small differences between observed and expected frequencies, which also explains the significance in this application. The authors also indicate χ^2 can be relevant to determine dependency rate. When the item pairs with large χ^2 were examined the majority of them were based on the same reading text or they were related to the same topic. Some item pairs with a large χ^2 also contained a problematic item (i.e. Item 9).

In brief, three of the bivariate absolute fit indices (except for MADres) and average RMSEA were below 0.05 and were acceptable. Because the LCDM showed a slightly better performance than others based on relative fit, it was specified as the final model for the Expert-defined Q-matrix.

The LCDM was then fit to the validation sample. The comparison of absolute fit indices obtained from the validation sample against the training sample proved that the model was still acceptable (Table 20). The patterns across the samples were similar, and overall, the fit with the validation sample was deemed to be adequate. The item parameter estimates across the two samples were also compared (Figure 5). On average, there were small deviations. Intercepts and main effect coefficients were more comparable than the two way and three-way interactions. Item mastery probabilities showed minimal deviations across the two samples. The average of mean absolute deviation of item mastery probabilities across all items was 0.023 (sd = 0.024) with a minimum of 0 and maximum of 0.161. In brief, the differences were negligible, and it was concluded that the changes in the Q-matrix and the final model would hold for the validation sample.

Figure 5. Comparison of Item Parameters between the Training and Validation Samples for the Expert-defined Q-matrix

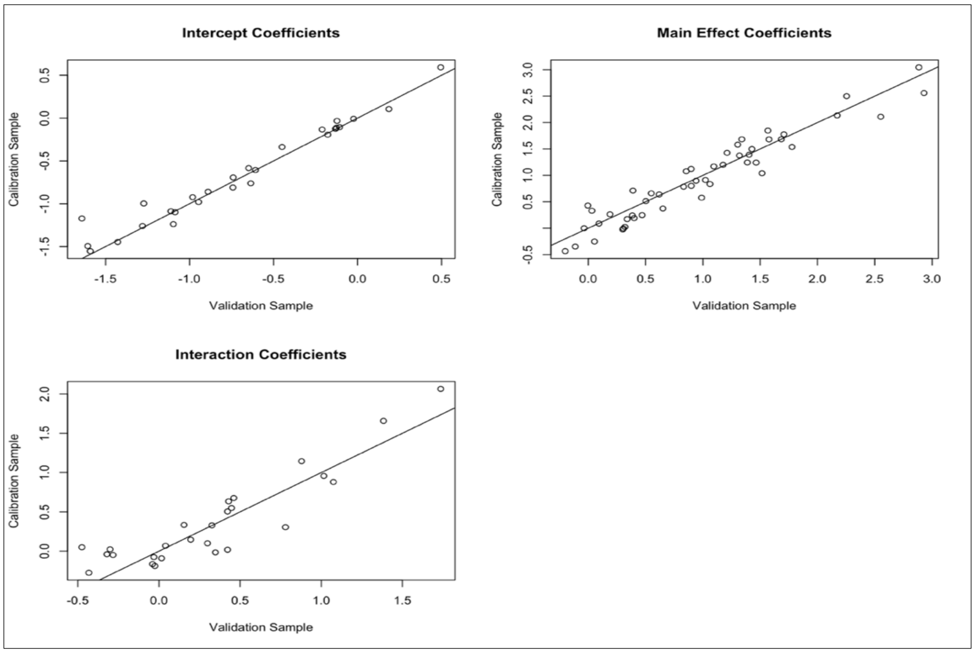


Table 20. Comparison of Absolute Fit Indices between the Training and Validation Samples for the Expert-defined Q-matrix

	LCDM Absolute Fit Indices					
	$M\chi^2$	MADcor	SRMSR	MADres	MADQ3	RMSEA
Training Sample	121.345	0.0167	0.0231	0.3749	0.0215	0.021
Validation Sample	129.163	0.0175	0.0237	0.3916	0.0234	0.023

Comparison of Models for the Standards-based Q-matrix

Because the Standards-based matrix was not statistically validated, the model fit analysis just based on the validation sample is presented here. It must be noted that model fit results on the training sample was comparable (Appendix F, Table 3) and that the final model fit both samples adequately.

Like the Expert-defined Q-matrix, the LCDM fit better than all models based on log likelihood and AIC. The C-RUM overperformed according to BIC and CAIC. The DINA and the HO-DINA were the worst fitting models among all. Likelihood ratio tests supported that the LCDM was significantly better than all constrained models (Table 22).

Table 21. Relative Fit Indices for the Standards-based Q-matrix

	Npars	-2LL	AIC	BIC	CAIC
LCDM	145	**-195658.4**	**391606.701**	392678.286	392823.286
R-RUM	118	-195748.1	391732.211	392604.259	392722.259
C-RUM	118	-195720.5	391677.093	**392549.141**	**392667.141**
DINO	91	-196885.4	393952.759	394625.271	394716.271
DINA	91	-196909.4	394000.693	394673.205	394764.205
HO-DINA	70	-197200.0	394539.954	395057.271	395127.271

Note. Results are based on the validation sample. Bold denotes lower values thus better fit.

Table 22. The Likelihood Ratio Tests for the Standards-based Q-matrix

Model 1	Model 2	χ^2	df	p
LCDM	R-RUM	89.7	27	<.00001
	C-RUM	62.1	27	.000139
	DINO	1227	54	<.00001
	DINA	1251	54	<.00001

Note. Results are based on the validation sample.

A similar pattern to the Expert-defined matrix was observed for the absolute fit results as well (Table 23). All models fit reasonably well based on MADcor, SRMSR, MADQ3, and average RMSEA. MADres was higher and maximum χ^2 was significant. Absolute fit indices did not present a clear division. Based on these results, the LCDM was specified as the final model also for the Standards-based matrix.

Table 23. Absolute Fit Indices for the Standards-based Q-matrix

	$M\chi^2$	p	MADcor	SRMSR	MADres	MADQ3	Mean RMSEA
LCDM	75.290	0	0.0123	**0.0157**	0.2808	0.0239	0.016
R-RUM	41.124	0	**0.0119**	0.0159	0.2685	**0.0230**	**0.011**
C-RUM	57.054	0	0.0120	0.0159	0.2698	0.0245	0.014
DINO	81.366	0	0.021	0.0269	0.4706	0.0188	0.025
DINA	83.642	0	0.021	0.0274	0.4713	0.0183	0.025
HO-DINA	196.666	0	0.025	0.0318	0.5809	0.0176	0.037

Note. Results are based on the validation sample. Bold denotes lower values.

Comparison of the Expert-defined and Standards-based Q-matrices

All fit indices obtained from the LCDM estimation for the Standards-based and Expert-defined Q-matrices (i.e., with the validation sample) were evaluated to determine the most favorable Q-matrix (Table 24). The Standards-based Q-matrix fit slightly better than the Expert-defined Q-matrix. Relative fit indices were lower for the Standards-based matrix than the Expert-defined matrix. It must be highlighted that more attributes were associated with the Standards-based matrix and the model included more parameters. In addition, both matrices were acceptable based on absolute fit. However, upon examining the associations among the attributes of the Standards-based Q-matrix (Table 25), it was determined that the Expert-defined Q-matrix was more suitable.

Table 24. Comparison between the Expert-defined and Standards-based Q-matrices

	Standards-based Q-matrix	Expert-defined Q-matrix
Relative Fit		
-LL	**-195658.4**	-196322.85
AIC	**391606.701**	392893.709
BIC	**392678.286**	393810.099
Absolute Fit		
$M\chi^2$	75.290	129.163

MADcor	0.012	0.018
SRMSR	0.016	0.024
MADres	0.281	0.392
MADQ3	0.024	0.023
Mean RMESEA	0.016	0.023

Table 25. Tetrachoric Correlations among Attributes Obtained from the LCDM with the Standards-based Q-matrix

	LoSI	LoMA	LoLA	LoSC	LoSS	Recount	Explain
LoMA	0.751						
LoLA	0.867	0.795					
LoSC	0.798	0.877	0.738				
LoSS	0.768	0.895	**0.957**	0.846			
Recount	0.053	-0.066	0.234	0.135	0.168		
Explain	0.163	-0.088	0.336	0.052	0.222	**0.999**	
Argue	0.006	-0.033	0.179	0.082	0.129	**0.981**	**0.981**

Note. Correlations higher than 0.90 are bolded. The proportion of masters for Recount, Explain and Argue was 47%, 41%, 46%. LoSS and LoLA were mastered by 37% and 52% respectively. LoSI: Social instructional language, LoMA: Language related with math, LoLA: Language related with language arts, LoSC: Language related with science, LoSS: Language related with social studies

Correlations among the attributes related to the standards were high, with a range of 0.75-0.96 (Table 25). On the other hand, the key uses dimension was slightly correlated or not correlated with the standards dimension at all. The tetrachoric correlations ranged between -0.07 and 0.34. The highest correlation was between Explain, Recount and language related to language of arts, and Explain and language of social studies. However, even those were weak correlations. This was not surprising, as standards represent language related to the subject areas, while key uses signify language functions students need to respond correctly. To give an example, knowing math language does not suggest the student should master arguments as well. However, despite some degree of differentiation among the standards, key uses were perfectly correlated

(0.98-0.99). The proportion of masters for key uses were also similar. 47% of ELs mastered Recount, 41% mastered Explain, and 46% mastered Argue. The high correlations and similar proportions of masters suggest that attributes in the key use dimension are highly associated with each other. If an EL masters Explain, the function that is associated with understanding the relation between ideas, processes, and consequences, then they are also masters of Recount (understanding summary and details) and Argue (understanding arguments and judgements). This finding suggests that key uses or language functions cannot be separated from each other, or represented as separate attributes due to high associations. This pattern was also reflected in the class proportions. More than half of the students were classified in the profile where none of the attributes (20%), just the standards (13%), just the key uses (12%), or all attributes were mastered (11%). However, merging the key uses and representing them as a single attribute was not possible. It would reduce the Standards-based Q-matrix to a simple structure with the standards dimension only. Given the indivisibility of key uses in the Standards-based Q-matrix and relatively large number of profiles (i.e., 256 distinct classes) that is not practical for reporting, the Expert-defined Q-matrix was adapted as the final Q-matrix.

Evaluation of the LCDM for the Quality of Diagnostic Information

Findings Related to Items

As the LCDM showed reasonable fit with the Expert-defined Q-matrix, various model parameters and statistics were explored next. Table 26 presents the intercept, main effect, and interaction coefficients associated with each item. Table 27 gives essentially the same information by transforming the coefficients to the probabilities. In particular, it reveals the probability of correct response for non-masters and masters of each attribute and the interactions.

Low intercept but high main effect or interaction coefficients (i.e. low probabilities for non-masters and high probabilities for masters of the attributes) are desired to suggest masters of the attributes correctly respond to the items rather than non-masters. This also reflects the degree of association between the item and the specified attributes (Templin & Hoffman, 2013). The intercept coefficients ranged between -1.64 and 0.49 with a mean of -0.70. Most of the intercepts were low. The average probability of correct response for non-masters was 0.34. It was slightly higher than 0.25, which is the guessing probability for an item with 4 options.

There were some items with slightly higher coefficients that yielded a higher correct response probability for the non-masters. Specifically, for Item 10, despite not mastering Sequences the probability that an EL gets the item right was estimated to be 0.60. Similarly, non-masters of Vocabulary and Explicit information still had a 0.55 probability to correctly respond to Item 12. The review of the classical item statistics for these items showed they were the easiest items on the test (p-values = 0.70, 0.75 respectively) and that Item 10 was poorly discriminating between high and low performers. Furthermore, the content review of these items showed an EL with high science knowledge could answer the items right without even reading the text. These items were tapping the same standard, language of science and were accompanied by a diagram. Specifically, for Item 10, even a low performing EL could possibly eliminate two of the distractors by looking at the diagram, which partially explains the high intercept.

The probabilities for Items 1, 3, 4, 7, 11, 16, 19 varied between 0.45 - 0.49 for non-masters of the required attributes. These items were also comparatively easier items than other items on the test. The extent to which the probabilities drift from 0.25 (i.e., guessing) might imply that ELs were able to eliminate one to two distractors. The distractor analysis also supported that there was at least one distractor that was not as reasonable as the others, and attracted to a very small proportion of ELs. The characteristic of item 7, 11 and 19 were also similar to Item 10 and 12. Item 11 was related with items 10 and 12. The diagram might have cued the answer for some students. Similarly, ELs with high math knowledge would not need the reading stimulus to respond item 7 and 19 correctly, which was also noted by two of the SMEs. The correct response probability for the non-masters for the remaining items was within a range of 0.16 – 0.35.

Both the main effect and interaction coefficients were generally large. On average, the correct response probability for masters of the all required attributes across all items was 0.75. There were some easy and hard items. For example, the main effect for Item 4, which measures Explicit Information was estimated to be 2.88, yielding a correct response probability of 0.94 for an EL mastering the attribute. Similarly, by knowing Summary an EL has a correct response probability of 0.82 for Item 14 (λ_1= 2.254). Both of these items were simple items. Item 13 had a main effect of 2.55 for Vocabulary and 2.93 for Grammar. Knowing these two attributes, an EL has 0.71 and

0.78 probability of correct response. By knowing both attributes, an EL almost always gets the item right (0.96).

However, there were some exceptional items, like Items 2 and 9, for which the main effects or interactions were small ($\lambda_{j,1}$ or $\lambda_{j,2}$< 0.50). Item 9 was one of the most complex items on the test. An EL mastering one of the attributes has a 25 – 26 % probability of a correct answer (i.e., equal to guessing). Knowing all attributes increases the probability only to 0.48. Thus, even with knowing all three attributes, the correct response probability was still low, and the item was hard. This item also had a low p-value (0.34) and point biserial (0.11), and it was called out by several SMEs for being confusing. Although being less helpful to determine an EL's classification, there was still 26% difference between not knowing or knowing of the attributes.

Similarly, Items 17, 23, 26, and 27 yielded correct response probabilities lower than 0.60 for masters of all required attributes. They were difficult items, even for the masters. These items also had low p-values and low point biserial, similar to Item 9. Knowing the attributes contributed to the correct response probability but at a lesser degree. It is also recognized that Sequence for Item 8 and Inference for Items 15 and 27 were at the monotonicity boundary. The fact that knowing these attributes, in addition to others, increased the probability (i.e., non-compensatory relationship) and deemed this result admissible.

Table 26. Item Parameter Estimates Obtained from the LCDM with the Expert-defined Q-matrix

Item	Intercept ($\lambda_{j,0}$)	Main Effect ($\lambda_{j,1(k)}$)						Interaction ($\lambda_{j,2(k,k`)}, \ldots$)	
		VOC	GRM	EXP	INF	SUM	SEQ	2-way	3-way
1	-0.024			1.403					
2	-0.608			0.503	0.470			0.461	
3	-0.123	0.989		1.022				-0.475	
4	-0.131			2.884					
5	-0.449	1.516				1.684		-0.321	
6	-0.890	0.651		1.176				0.429	
7	-0.126	0.899		1.320				-0.025	
8	-1.111			0.619			-0.201	1.384	
9	-1.272		0.190	0.306			0.301		0.447
10	0.494						1.388		
11	-0.107				1.778	1.096		-0.283	
12	0.186	1.465		1.304				0.326	
13	-1.641	2.552	2.930					-0.432	
14	-0.742					2.254			
15	-1.097	0.031	0.390		-0.036				0.779
16	-0.211			1.707			1.340	0.423	
17	-1.590					0.550	0.385	1.016	
18	-1.606				0.055	0.400		1.736	
19	-0.178			2.171					
20	-0.740			0.832			0.855	-0.040	
21	-0.981				0.340	0.320		0.879	
22	-0.648			0.941			1.211	0.016	
23	-1.280	1.062				0.094		0.423	
24	-0.946				1.577				
25	-0.635	1.567				0.897		0.299	
26	-1.427	1.428							
27	-1.086	-0.003			-0.113			1.075	

Note. VOC= Vocabulary, GRM= Grammar, EXP= Explicit Information and Details, INF= Inference, SUM= Summary, SEQ= Sequences and Processes.

Table 27. Correct Response Probabilities for the Masters and Non-masters of the Attributes Obtained from the LCDM

Item	Non-masters	Masters							
		VOC	GRM	EXP	INF	SUM	SEQ	2-way	3-way
1	0.494			0.799					
2	0.353			0.474	0.466			0.695	
3	0.469	0.704		0.711				0.804	
4	0.467			0.94					
5	0.39	0.744				0.775		0.919	
6	0.291	0.441		0.571				0.797	
7	0.468	0.684		0.767				0.888	
8	0.248			0.379			0.212	0.666	
9	0.219		0.253	0.276			0.275		0.477
10	0.621						0.868		
11	0.473				0.842	0.729		0.923	
12	0.546	0.839		0.816				0.964	
13	0.162	0.713	0.784					0.968	
14	0.323					0.819			
15	0.25	0.256	0.33		0.244				0.632
16	0.447			0.817			0.756	0.963	
17	0.169					0.261	0.231	0.589	
18	0.167				0.175	0.231		0.642	
19	0.456			0.88					
20	0.323			0.523			0.529	0.712	
21	0.273				0.345	0.34		0.636	
22	0.343			0.573			0.637	0.821	
23	0.218	0.446				0.234		0.574	
24	0.28				0.653				
25	0.346	0.717				0.565		0.894	
26	0.194	0.5							
27	0.252	0.252			0.232			0.468	

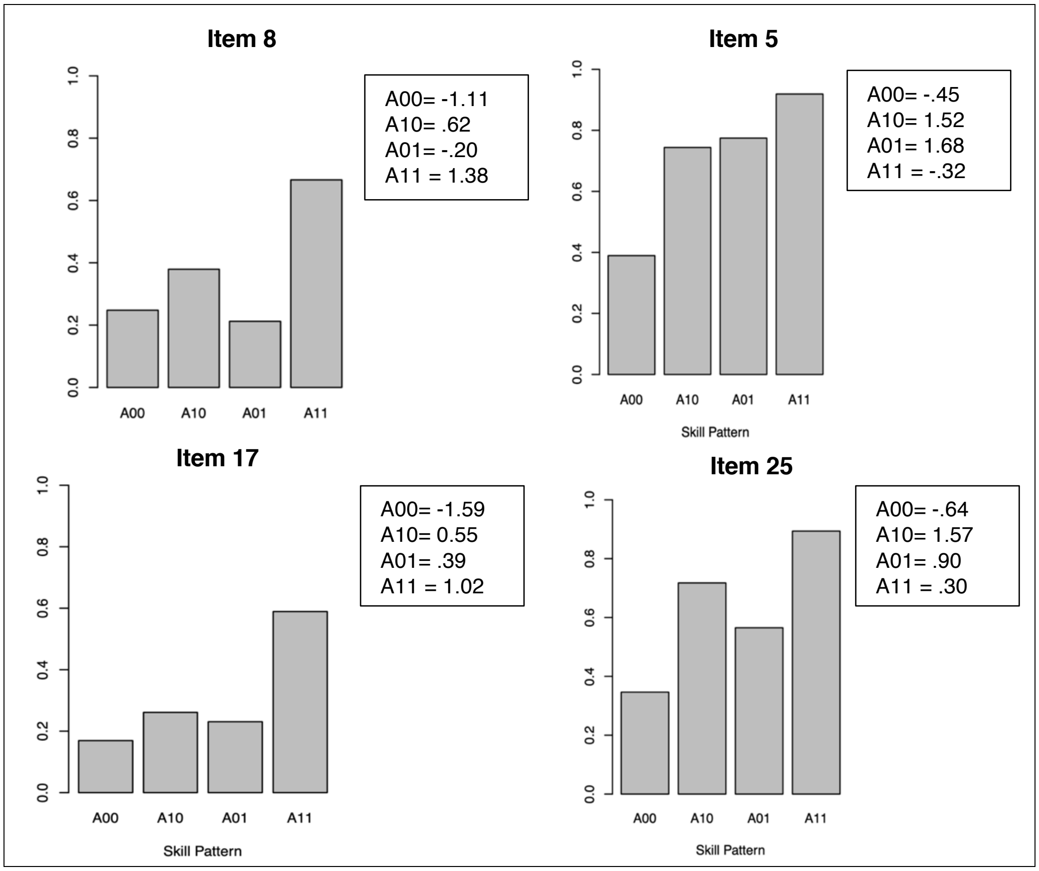
Note. VOC= Vocabulary, GRM= Grammar, EXP= Explicit Information and Details, INF= Inference, SUM= Summary, SEQ= Sequences and Process.

Item coefficients were also reviewed to understand the relationships among attributes, that is, how attributes interact with each other as they relate to the correct response probability. As seen in Table 26 (i.e., main effect and interaction parameters), compensatory and conjunctive relationships coexisted in the test, which aligns with fit results presented earlier (i.e., comparable performance of the C-RUM and R-RUM to the LCDM). Among the 20 items with complex structure, half were consistent with conjunctive structure (i.e., Items 2, 6, 8 ,9, 15, 17, 18, 21, 23, and 27), particularly the R-RUM (i.e., positive main effect and interactions), except for Items 15 and 27, which were more similar to the DINA (i.e., zero main effect and positive interactions). Two of these items (8 and 17) are presented in Figure 6. Apparently, there was a noticeable increase for the mastery of the both attributes. On the other hand, the remaining items were more consistent with a compensatory model like the C-RUM, with interaction terms close to 0, and positive main effects. So, there was only a slight increase for the mastery of both attributes (See Figure 6 for Items 5 and 25 as an example).

Wald's procedure (de la Torre & Lee, 2013) to compare whether the constraint model fit better than the saturated model at the item level was also applied. Results (Appendix F, Table 4) displayed that for twelve of the items the constrained model fit better at the item level, and eight of these items converged with the C-RUM, the remaining items converged with the DINA or R-RUM.[6]

[6] Item specific model could not be fit because the CDM package does not allow the R-RUM and the C-RUM to be fit at the same time along with monotonic constraints. The constraint is specified for the whole estimation and not item specific. The constraint is required for the identifiability of the LCDM (Henson et al., 2009). Wald's procedure was conducted using the GDINA package. CDM package only allows comparing LCDM, ACDM and DINA.

Figure 6. Item Response Functions for the Selected Items



Fit of the individual items and their discrimination capacity is shown in Table 28. The discrimination index (DI) for these items that is essentially the absolute difference of the probabilities of the non-masters and the masters of required attributes. The average DI was 0.42. Jang (2009b) and Kim (2015) considered 0.40 as large, while Nájera et al. (2019) proposed 0.60 and higher for discrimination. Given Jang and Kim's criteria, test items can be considered as having acceptable discrimination capacity. Item 13 produces the largest difference between masters and non-masters, with 0.81. The probability of masters was also 50-55% higher for Items 5, 6, 14, 16, and 25, and 40-48% higher for

Items 4, 7, 8, 11, 12, 17, 18, 19, 20, and 22. For the remaining items, masters differed with respect to their performance from non-masters at a lesser degree (< 0.40). The difference was especially less discernable for Items 9,10, and 27 (i.e. DI = 0.22-0.25) which also had low point biserial values. All items below 0.40 criterion were also either the easiest or the hardest items on the test.

When individual RMSEA values were reviewed all of them were smaller than 0.05. However, it must be noted that Item 9 had the largest RMSEA (0.05) and was on the borderline.

In summary, item estimates demonstrate that the specified attributes were associated with items and, overall, the majority of the items can be useful for diagnostic information, to some extent, as they were within the acceptable ranges. There were some items not performing as well as others and they might be more limited to separate masters and non-masters.

Figure 7. Correct Response Probabilities of the Masters and Non-masters

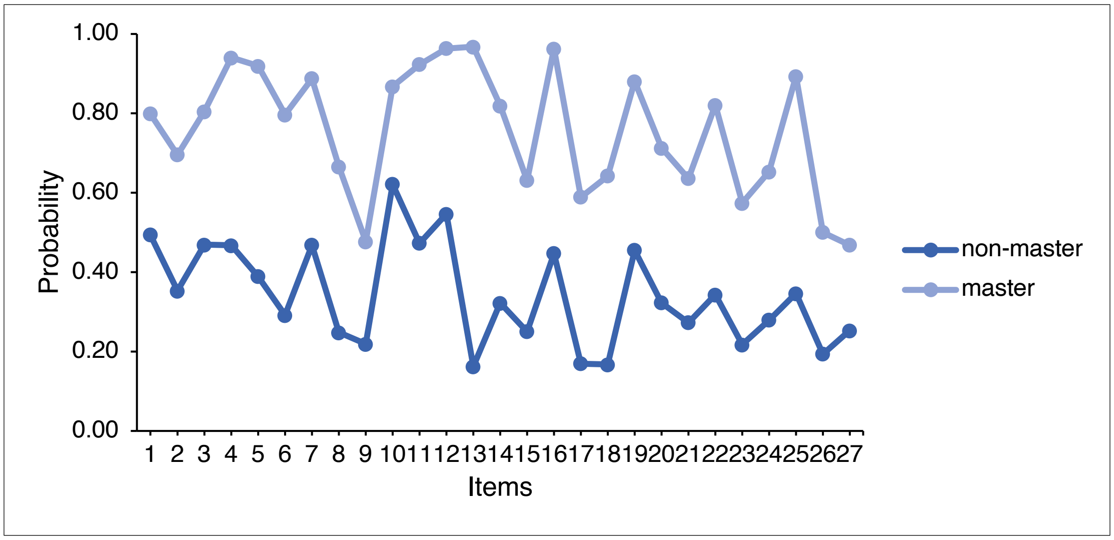


Table 28. Item Fit and Discrimination Statistics

Item	RMSEA	DI
1	0.036	0.305
2	0.013	0.343
3	0.018	0.335
4	0.036	0.473
5	0.032	0.530
6	0.017	0.506
7	0.015	0.419
8	0.012	0.418
9	0.049	0.258
10	0.042	0.247
11	0.016	0.450
12	0.016	0.417
13	0.008	0.806
14	0.042	0.497
15	0.028	0.381
16	0.019	0.516
17	0.011	0.420
18	0.007	0.475
19	0.046	0.424
20	0.011	0.389
21	0.012	0.363
22	0.008	0.477
23	0.016	0.357
24	0.028	0.373
25	0.024	0.547
26	0.031	0.307
27	0.029	0.216
Mean	*0.023*	*0.417*

Note. RMSEA is an item fit measure. DI= Discrimination Index (i.e. difference in the probability of masters and non-masters of all required attributes for the item).

Findings Related to Person Estimates and Latent Classes

Latent classes were scrutinized in an effort to understand the quality of classifications. There were 64 classes (i.e. $2^{6)}$ that ELs can be assigned to, and Table 29 shows 23 of the most frequent classes. Class probability denotes the probability that the profile can be observed, and the expected frequency is the proportion of the students in a class. For instance, 26% of the ELs were most likely to be classified in the class where none of the attributes are mastered. On the other hand, the profile that requires the mastery of all attributes had a probability of 0.16. That is 42% of the ELs were either the masters or non-masters of all attributes. It is not unexpected for these classes to emerge as the most frequent profiles, especially when DCM is retrofitted, which might signal unidimensionality. For a successful application, some variability is expected. Although these two classes were dense, and some classes were sparse or had no students, there some limited variability. It must also be noted that the number of profiles was still large, as there were many attributes. Specifically, about 9% of ELs were likely to be in the profile where only Grammar is mastered. Likewise, the profiles including Grammar-Explicit Information, and Vocabulary-Inference-Sequences were each likely to be mastered by 4% of ELs. ELs who mastered Grammar-Syntax-Summary also formed a large cluster with the probability of 0.18.

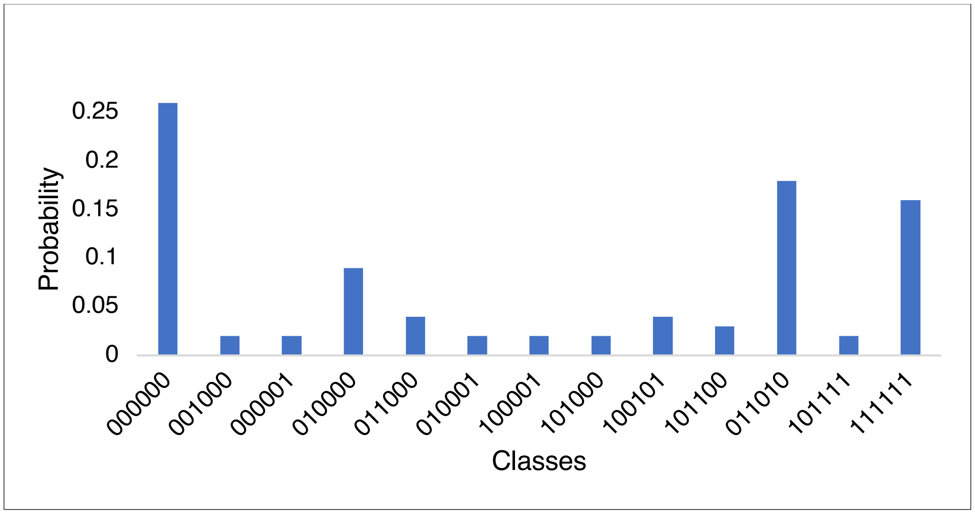
The probability of latent classes can also provide valuable information for attribute development. For example, there were almost no students mastering only Inference or Summary skills, hinting that these skills develop later than others or they co-occur with other skills. For instance, a large proportion of students were likely to acquire Summary after Grammar and Explicit Details were attained. This is reasonable, as Summary requires processing sentences rather than details located in a single sentence and understanding relations between the sentences. Likewise, Inference emerged with other skills such as Vocabulary-Sequences or Vocabulary-Explicit information.

Table 29. The Probabilities and Proportions of Latent Classes

Profile	Probability	Expected Frequency
000000	0.26	3076.78
010000	0.09	1103.07
000001	0.02	228.90
001000	0.02	183.85
100000	0.01	155.38
101000	0.02	174.16
011000	0.04	479.23
010001	0.02	230.44
100001	0.02	200.41
011010	0.18	2119.03
100101	0.04	452.97
101100	0.03	301.75
101010	0.01	111.76
011100	0.01	107.07
010101	0.01	67.16
101101	0.01	162.24
111010	0.01	141.54
011011	0.01	122.98
101111	0.02	215.68
111011	0.01	142.41
011111	0.01	85.41
111111	0.16	1855.48

Note. Less frequent classes (< 50 ELs) were omitted due to space limitations. Attribute order Vocabulary, Grammar, Explicit Information and Details, Inference, Summary, Sequences and Processes

Figure 8. The Proportion of the Most Frequent Classes



The accuracy and consistency of these classifications were also estimated to determine whether it would be appropriate to use the class (i.e., pattern) information as well as the individual attributes (Table 30). When prior information, how ELs performed on the test, was used both accuracy and consistency improved specifically for the pattern level. However, for a standardized test, the maximum likelihood (MML) method might be preferred over using prior information (i.e., MAP), as it represents actual behavior. With respect to the overall pattern, an EL's class was accurately estimated 26% of the time, whereas they were consistently classified into a class 11% of the time. These indices are related to each other, such that consistency (P_c) is equal to or smaller than accuracy (P_a) (Wyse & Hoa, 2012 in Wang et al., 2015). Despite being greater than random chance, the pattern level accuracy and consistency was not high, even when prior information was used. However, this was not surprising due to the large number of attributes, lower discrimination, short tests, or moderate to high associations that compromise both indices. Although 0.70-0.80 range is considered as acceptable, Cui et al. (2012) report P_a as 0.44 and P_c as 0.25 for 5 moderately correlated attributes that are associated with 20 low discriminating items. For the same conditions, Wang et al. (2015) estimate P_a as 0.20 and P_c as 0.09. Therefore, the results were consistent with earlier findings. In contrast, the probability of accurately and consistently determining students' mastery of the individual attributes was, overall, adequate. P_a for the individual attributes ranged between 0.72 and 0.86, which was adequate. The consistency of the Sequences, Inference and Grammar was slightly lower (i.e., 0.61, 0.63, 0.64), which was also the case for some attributes in Wang's (2015) study for the described conditions. MAP estimates for both classification and accuracy were higher and within the expected ranges for the individual attributes.

Table 30. Classification Accuracy and Consistency of the Overall Profile and Individual Attributes

	MAP		MLE	
	P_c	P_a	P_a	P_c
Pattern	0.580	0.535	0.256	0.110
VOC	0.854	0.810	0.804	0.709
GRM	0.793	0.735	0.739	0.640
EXP	0.898	0.833	0.861	0.767
INF	0.863	0.820	0.745	0.627
SUM	0.864	0.807	0.790	0.680
SEQ	0.833	0.818	0.719	0.606

Note. Pattern denotes the accuracy and consistency at the class level. MAP= Maximum a posterior, MLE = Maximum likelihood estimation. P_a= accuracy, P_c = consistency. VOC= Vocabulary, GRM= Grammar, EXP= Explicit Information and Details, INF= Inference, SUM= Summary, SEQ= Sequences and Process.

Findings Related to Attributes

Diagnostic feedback regarding individual attributes can be considered more useful than the profile information based on above results. The individual attributes were further examined to uncover their distribution, difficulty, and associations with other attributes. Table 31 demonstrates the attribute mastery status of ELs. They were based on posterior probabilities of individual attributes. As suggested by Jang (2009b), probabilities below 0.40 denote non-mastery, 0.40-0.60 range denotes uncertainty, and above 0.60 represents mastery. Ideally, the proportion of the students with uncertain mastery status should be low, and the probabilities should be closer to 0 or 1 (Lee & Sawaki, 2009) to make confident conclusions about the attribute status. The students with uncertain status are likely to benefit from additional support for the given attribute. However, when many students rest in this region, it would suggest that the precision of the information might be compromised for a large group of students.

In this case, a majority of ELs could successfully be identified as masters or non-masters (i.e., 88-95%). The largest undetermined category emerged for Grammar, which is also apparent from the distribution of the probabilities in Figure 9 (i.e., the second histogram in the first row). Although most students were still closer to 0-1, there were quite a few ELs with probabilities of 0.50 and 0.60., unlike the other attributes producing more U-shaped distributions.

Figure 9. The Distribution of Posterior Probabilities for Individual Attributes

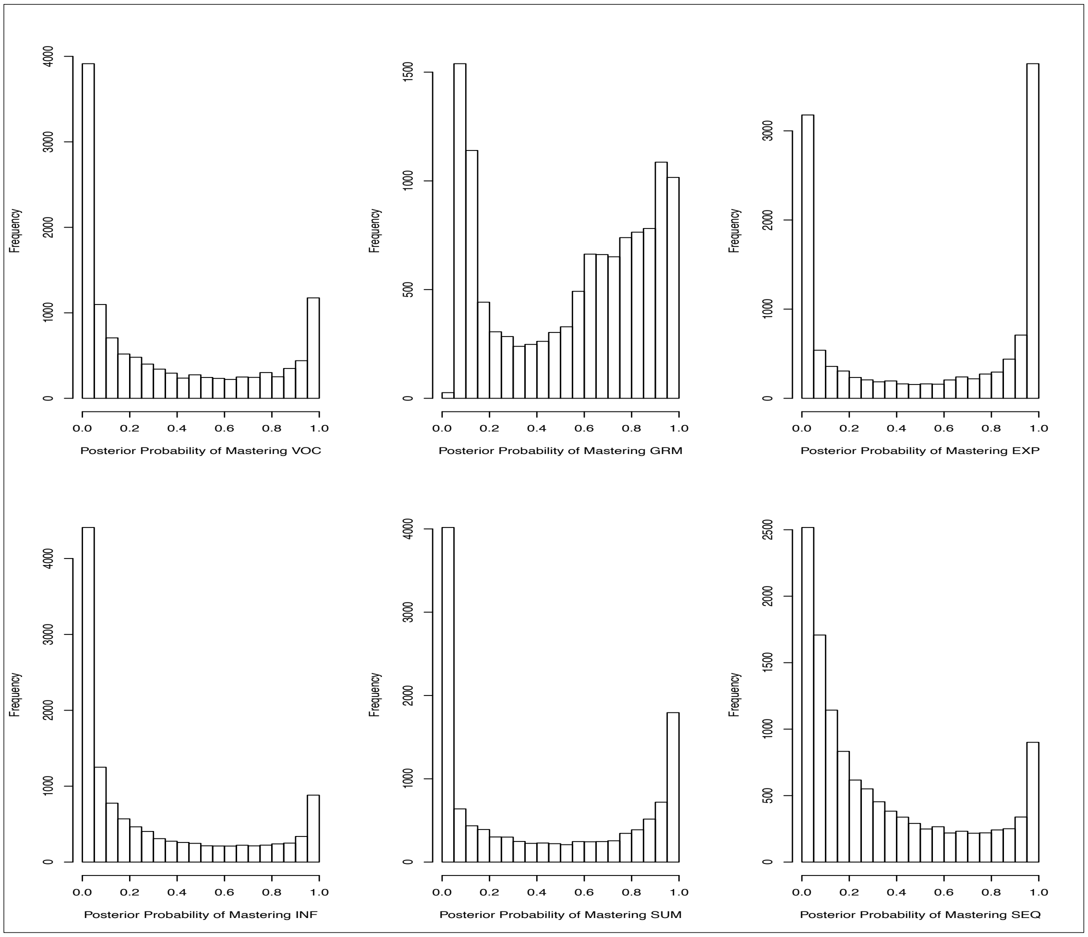


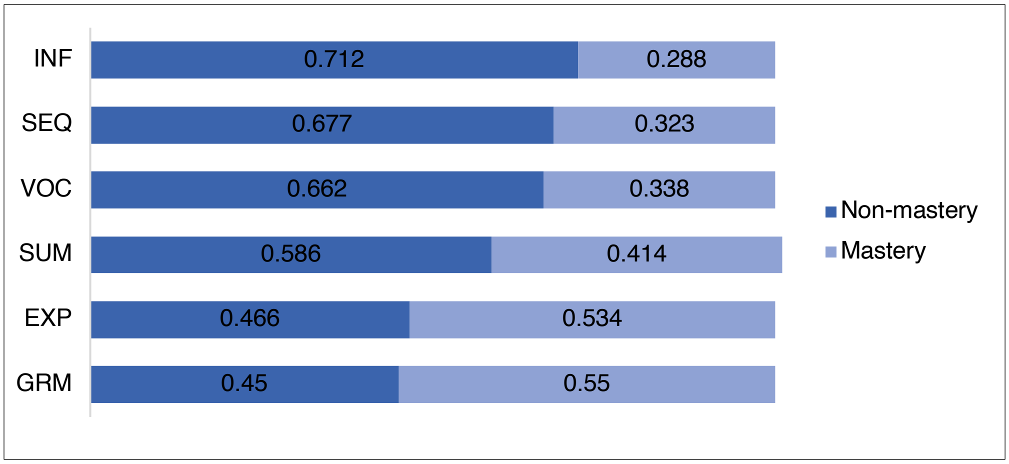
Table 31. Attribute Mastery Status

Attributes	Non-masters	Undetermined	Masters
VOC	7751 *(0.647)*	988 *(0.083)*	3232 *(0.270)*
GRM	4224 *(0.353)*	1386 *(0.116)*	6361 *(0.531)*
EXP	5202 *(0.435)*	637 *(0.053)*	6132 *(0.512)*
INF	8458 *(0.707)*	933 *(0.078)*	2580 *(0.216)*
SUM	6558 *(0.548)*	905 *(0.076)*	4508 *(0.377)*
SEQ	8206 *(0.685)*	1144 *(0.096)*	2621 *(0.219)*

Note. The numbers in the parenthesis indicate the proportion of the students. VOC= Vocabulary, GRM= Grammar, EXP= Explicit Information and Details, INF= Inference, SUM= Summary, SEQ= Sequences and Process.

The average of the posterior probabilities that conveys the difficulty of the skills also conformed with the expectations regarding the population. Figure 10 presents the probability of non-mastery or mastery of skills across all examinees. The attributes show varying difficulties, as anticipated. The level of difficulty associated with each attribute was also reasonable. Particularly, Grammar was the easiest attribute with a probability of 0.55, meaning most ELs can understand compound sentences and references. Explicit Details were likewise mastered by 53% of ELs and comparatively easier than other attributes. Note that this attribute entails understanding details that is transparent and (i.e. in some cases verbatim in the text and the answer choice). Summary was more difficult, with a probability of 0.41, because it involves integrating information across multiple sentences and parts of the text, and sometimes understanding the main idea of a whole paragraph. Vocabulary, with a probability of 0.34, was a difficult attribute as it was related to knowing more abstract, content specific words and synonyms, as well as Sequences, which has a probability of 0.32. Finally, Inference was the hardest attribute of the test, and only 29% of the ELs were likely to master it. Inference is a more abstract skill, as it is related to understating implicit information and making predictions justifying its difficulty.

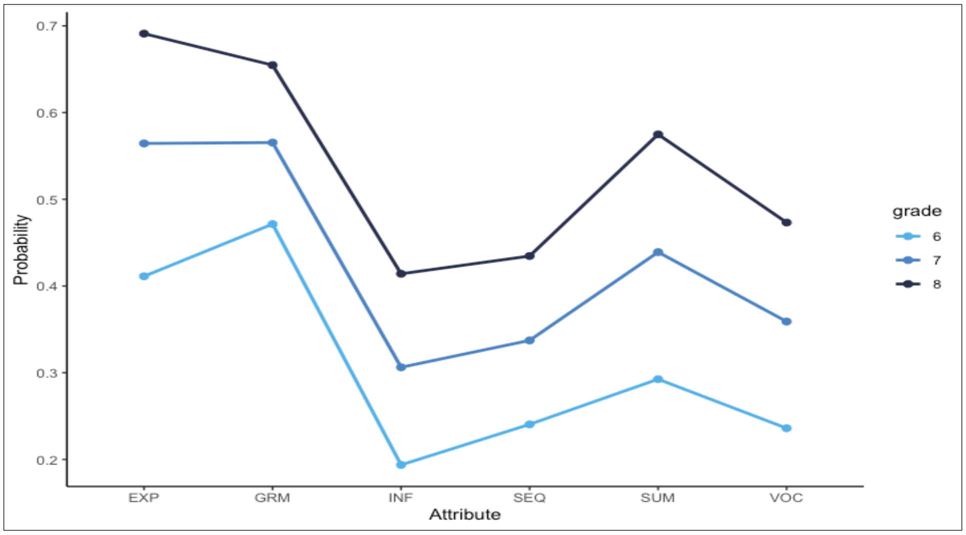
Figure 10. Average Posterior Probabilities of Individual Attributes



Note. VOC= Vocabulary, GRM= Grammar, EXP= Explicit Information and Details, INF= Inference, SUM= Summary, SEQ= Sequences and Process.

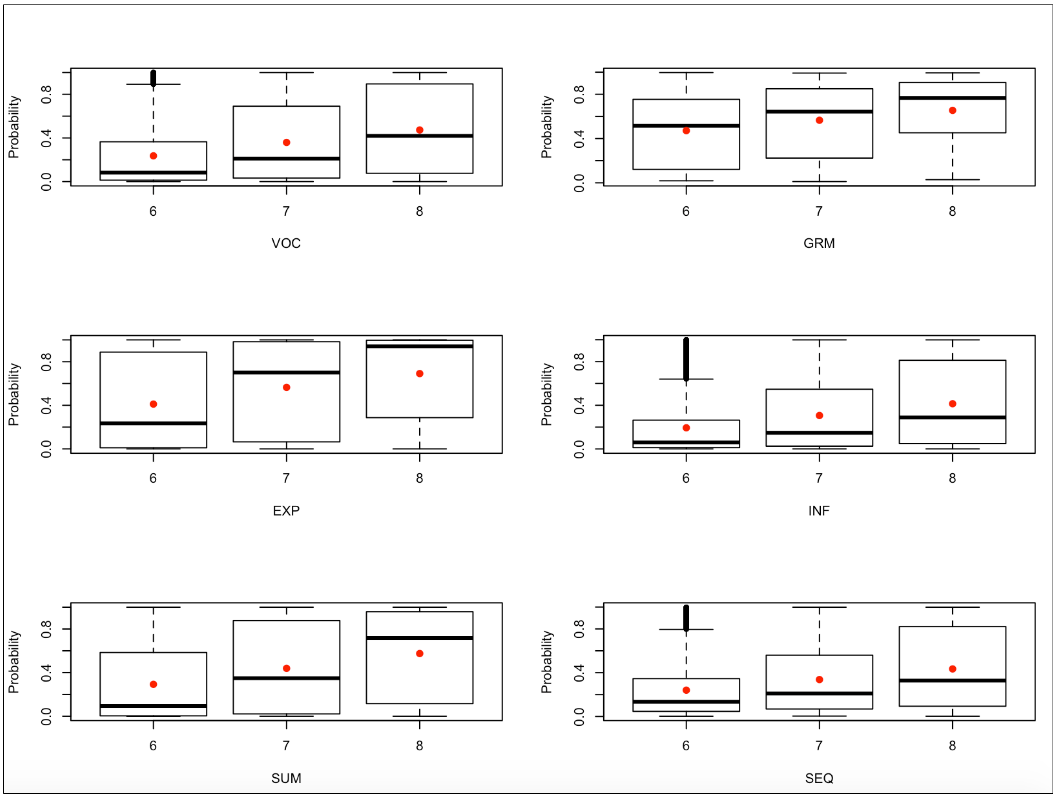
The skill probabilities were compared across the grades to uncover possible patterns. On average, 6th graders had the lowest probabilities, and 8th graders had the highest probabilities (Figure 11). Overall, this confirms that language ability improves over time. The patterns of the skills were consistent across grades, except for Grammar, which was easier than all the other attributes for only 6th Graders. Figure 12 shows that grade level differences were more distinct with respect to some attributes. For example, the variability (i.e., also the range) of mastering Inference was smaller among 6th graders. ELs' development of this abstract attribute might be more evident over time. It must be recalled that ELs enter the school system with varying proficiency. Thus, there were ELs in the 6th grade still having a high probability for the mastery of Inference and other attributes.

Figure 11. Posterior Probabilities of the Attributes across Grades



Note. Sample size for the grade 6, 7 and 8 are 5,088, 3,636 and 3,247 respectively. VOC= Vocabulary, GRM= Grammar, EXP= Explicit Information and Details, INF= Inference, SUM= Summary, SEQ= Sequences and Process.

Figure 12. The Distribution of the Posterior Probabilities of the Attributes within Grades



Associations among the attributes (Table 32) showed some variation (i.e., -0.01-0.99), yet the majority of the correlations provided some evidence for the viability of deconstructing the L2 reading construct. The different proportions of masters for the attributes also support the divisibility. The association between Vocabulary and Grammar was 0, meaning that understanding complex sentences does not necessarily mean an EL has knowledge of content-specific words or can recognize synonyms. Although Explicit Information is related to both of these attributes, it can be separated from them to some extent. It is also reasonable to assume that an EL should understand vocabulary and sentence structure to identify details in the text. Correlation between Inference and Vocabulary was 0.93 and it was high. Note that these attributes were among the most difficult. When an EL masters Inference, he/she presumably masters vocabulary. In addition, both skills relied less on direct information in the text. Inference was not highly correlated with other skills thus it might be possible to differentiate it. Summary was perfectly correlated with Explicit Information. Its association with Grammar was also high (0.87). As mentioned before, both Summary and Explicit Information rely on understanding the information in the text, yet Summary involves understanding groups of sentences. It is prudent to expect ELs who have mastered Summary to master Explicit Information. The skill necessitates recognizing sentence relations and references, which also explains its somewhat high association with Grammar. Summary was moderately correlated with the remaining skills, implying its distinctiveness. Sequences was highly related with Vocabulary (0.88) and Inference (0.90) but associated less with other skills. Some of these high correlations also explain the poor fit of the Higher order model.

Table 32. Tetrachoric Correlations among the Attributes in the Expert-defined Q-matrix

	VOC	GRM	EXP	INF	SUM
GRM	-0.013				
EXP	0.588	0.766			
INF	**0.932**	0.209	0.616		
SUM	0.498	0.865	**0.986**	0.494	
SEQ	0.876	0.239	0.338	0.900	0.492

Note. Correlations greater than 0.90 (bolded) are considered highly correlated (Sessoms & Henson, 2018). 0.80-0.90 range is considered typical (Madison & Bradshaw, 2015). VOC= Vocabulary, GRM= Grammar, EXP= Explicit Information and Details, INF= Inference, SUM= Summary, SEQ= Sequences and Process.

Relation to the Original Framework: IRT

Finally, posterior probabilities of the attribute mastery and final classifications were compared to the ability estimates and proficiency classification obtained from the original framework to further inspect DCM's viability. ELs' overall reading ability θ, increased as they mastered more attributes (Figure 13). However, for some students not mastering any attributes, θ can be as high as those mastering two attributes. This might be an effect of less diagnostic items which decreased the precision of attribute probabilities. A more detailed graph that shows θ against the specific classes (i.e., ordered according to number of attributes mastered) was included. Figure 14 substantiates that Grammar knowledge develops earlier, because for ELs just mastering this attribute, θ is relatively low. Whereas, Explicit Information might be acquired later as θ increases for masters of this attribute. Another example can be Vocabulary-Explicit Information-Inference that develops slightly later than Grammar-Explicit Information-Summary as the average ability is higher for the former. When such conclusions are made, one must consider the accuracy of profiles. It is not surprising to see some divergence for less accurate attributes.

Figure 13. The Distribution of θ for the Mastery of the Different Number of Attributes

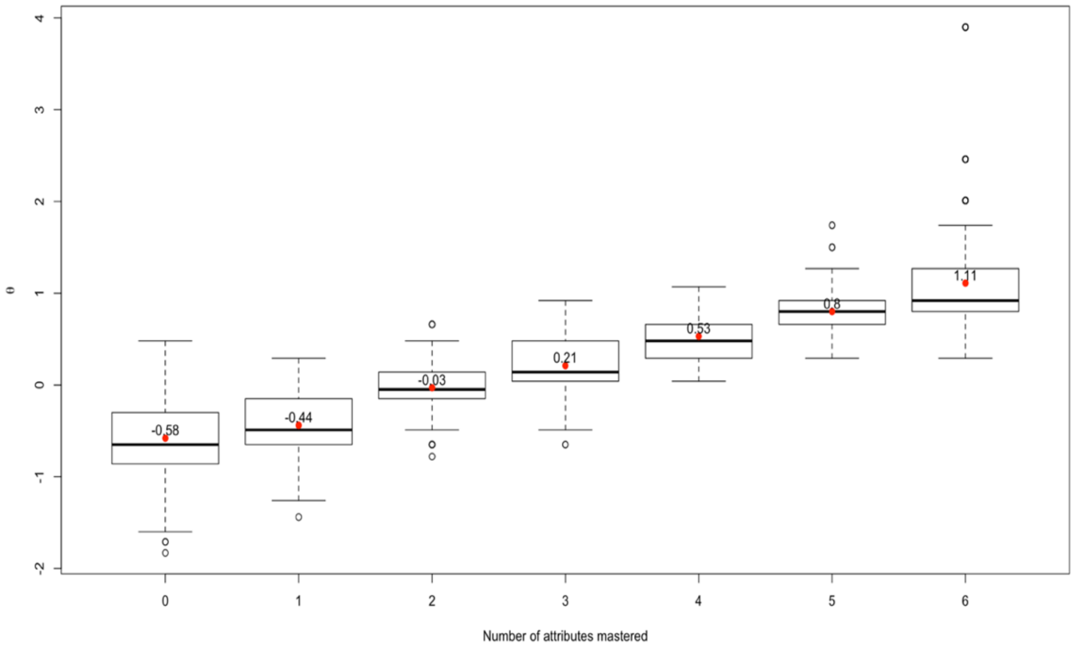


Figure 14. The Distribution of θ for the Most Frequent Attribute Profiles

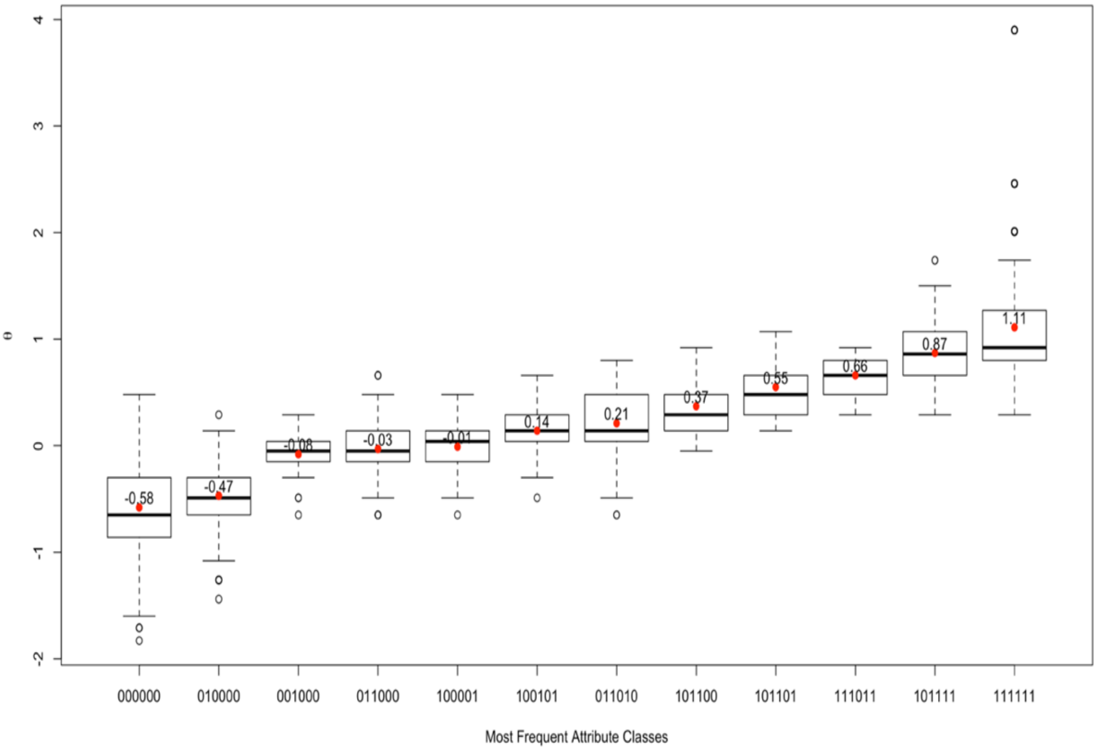
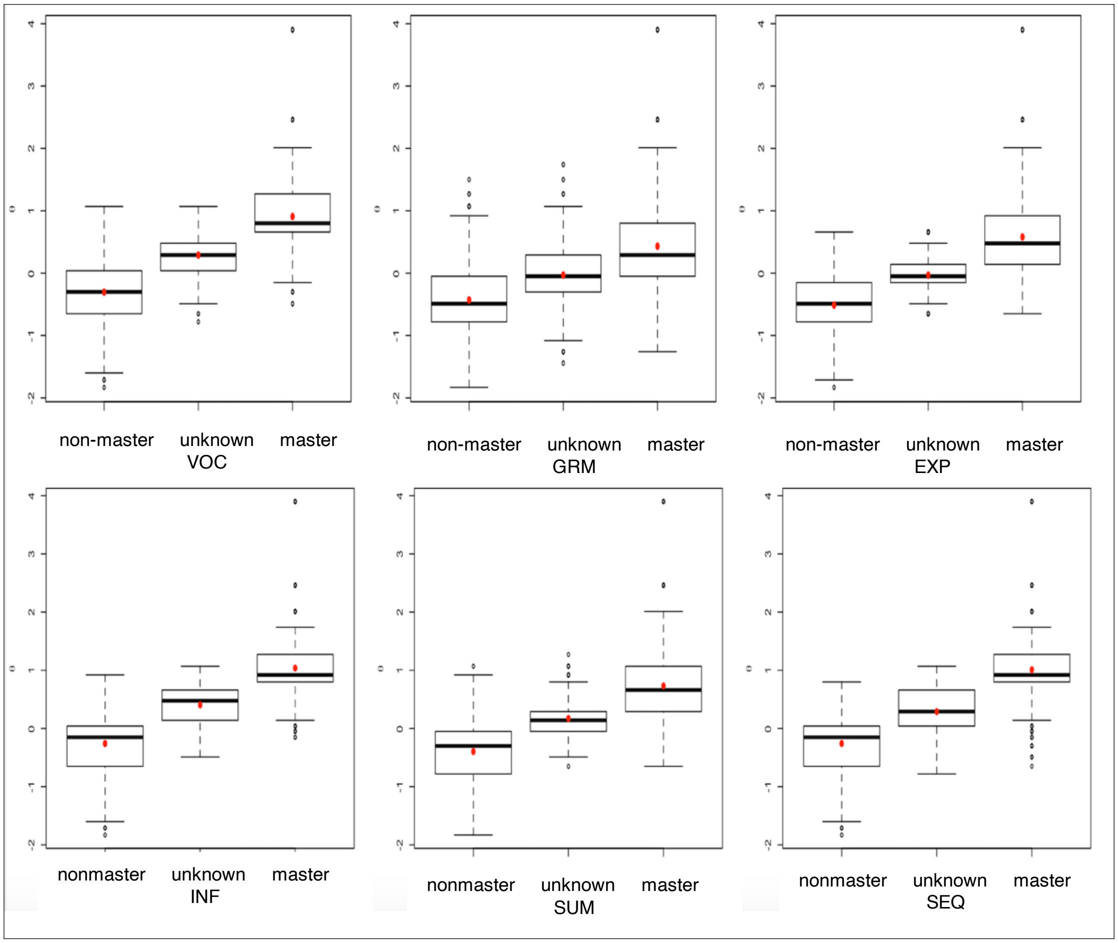


Figure 15 also shows θ distribution for the masters and non-masters of each attribute. Masters have a higher θ than non-masters and differences in the distribution of mastery probabilities between the groups is clear. A multiple regression analysis (i.e., posterior probabilities were treated as the predictors) also displayed that all attributes significantly predict θ and all together they explain 88% of the variability ($R^2 = 0.88$) (Appendix F, Table 5).

Figure 15. The Distribution of θ across the Different Mastery Status of the Individual Attributes



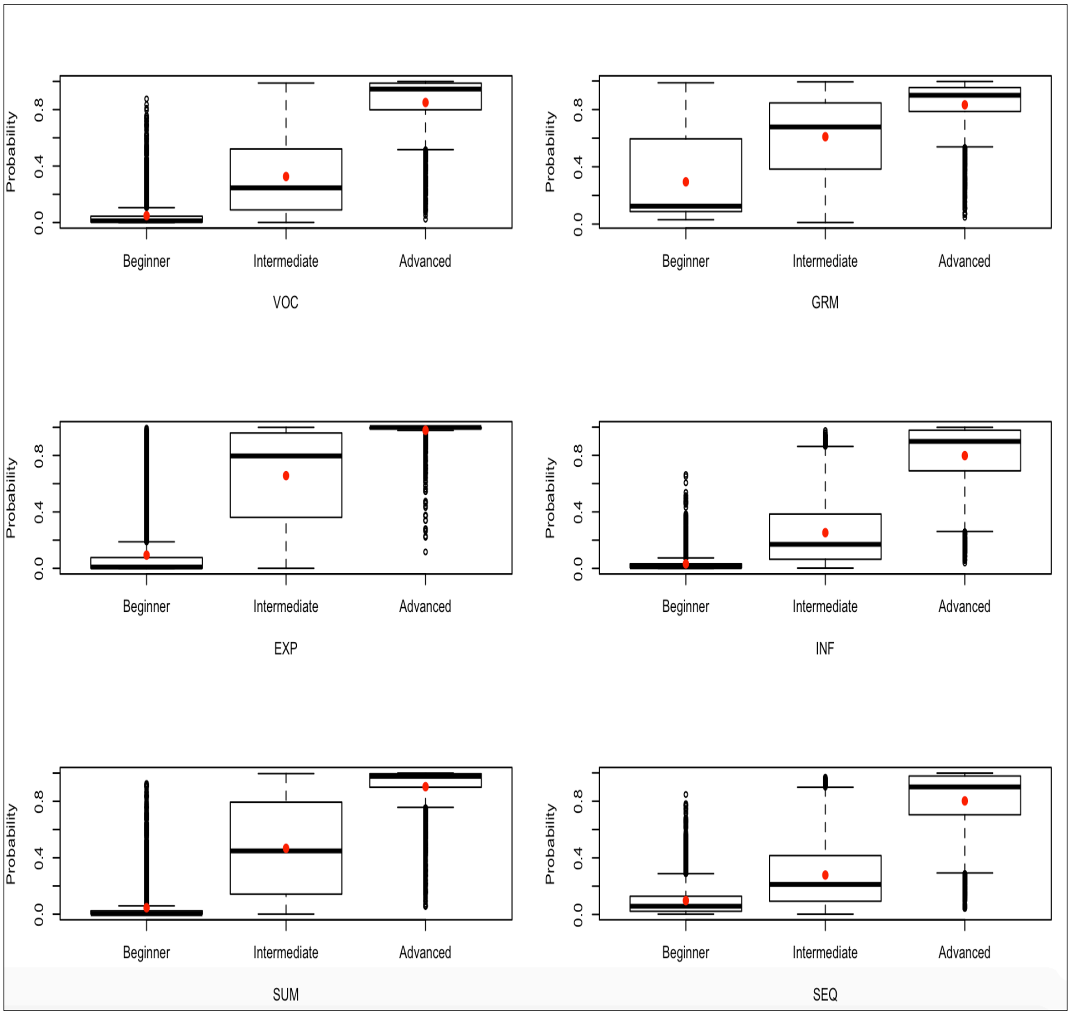
Note. Correlations between θ and the posterior probabilities of each attribute are VOC= 0.84, GRM= 0.66, EXP= 0.81, INF= 0.85, SUM= 0.83, SEQ= 0.82.

Students' DCM-based classifications were also compared to PL classification which was reported by the test developer. There were 6 PLs with more students achieving PL 2 or PL 3 (i.e., a total of 64%). PLs were grouped together, as shown in Table 33 (i.e., Beginner, Intermediate, Advanced), to simplify the comparisons. The posterior probabilities for each attribute across three groups were significantly different, and three groups varied with respect to mastery of the attributes (ANOVA results in Appendix F, Table 6). As seen in Figure 17, Advanced ELs' posterior probabilities were the highest, whereas as beginner ELs' probabilities were the lowest on average. Advanced ELs also had a higher probability to master all skills (i.e. average probability > 0.80). In other words, they are likely to master all skills. An opposite pattern emerged for beginners. These findings were aligned with the expectations. Figure 16 presents the distribution of probabilities for individual skills across different proficiency groups. Beginner and Advanced ELs' mastery probabilities for individual attributes were less spread compared to intermediate ELs who were the middle group. Distribution of individual mastery probabilities was clearly distinguishable, with majority of beginner ELs at the low end and advanced ELs at the high end. The proficiency of ELs in each profile was also examined (Table 34). Results mostly converged across two classifications. In line with the expectations, the majority of ELs in the profile where no attributes were mastered were the beginner ELs. There were no advanced ELs in this profile. Likewise, it was the beginners who mastered Grammar by itself. Intermediate ELs generally mastered 2-3 skills. The profile where all attributes were mastered consisted mostly of Advanced ELs, but surprisingly some Intermediate ELs, which might imply the importance of the accuracy of the overall profile. These results exhibit that two frameworks yielded comparable ability estimates, providing further evidence for the viability of DCM.

Table 33. The Number of ELs across Different PLs

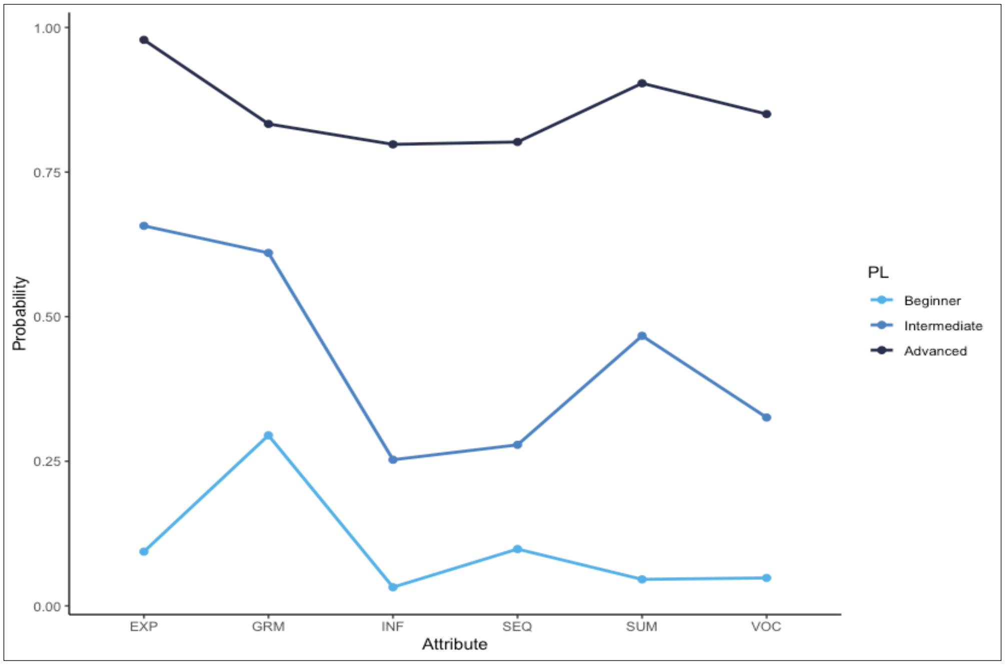
	Beginner (33%)		Intermediate (47%)		Advanced (20%)	
PL	1	2	3	4	5	6
N	293	3687	4013	1600	1565	813

Figure 16. The Distribution of the Posterior Probabilities of the Attributes within each PL



Note. VOC= Vocabulary, GRM= Grammar, EXP= Explicit Information and Details, INF= Inference, SUM= Summary, SEQ= Sequences and Process.

Figure 17. Posterior Probabilities of the Attributes across PLs



Note. VOC= Vocabulary, GRM= Grammar, EXP= Explicit Information and Details, INF= Inference, SUM= Summary, SEQ= Sequences and Process.

Table 34. The proportion of ELs with Different PLs across Different Attribute Profiles

Class	Beginner	Intermediate	Advanced
000000	2593	799	0
010000	1069	450	0
001000	13	68	0
100000	11	33	0
011000	52	195	4
010001	23	39	0
100001	12	41	0
101000	2	62	1
100101	23	447	28
011010	178	2728	264
101100	3	170	41
010101	0	10	0
011100	1	22	1
101010	0	4	1
101101	0	41	35
111010	0	15	4
101110	0	1	0

011011	0	11	3
101111	0	23	47
111011	0	13	13
011111	0	2	7
111110	0	1	0
111111	0	438	1929

Table 35. Posterior Probabilities and Characteristics of the Selected Students

ID	Background Characteristics				Posterior Probabilities					
	Grade	Time in ELP	PL	Raw Score	VOC	GRM	EXP	INF	SUM	SEQ
207846	6	6	2	8	0.023	0.677	0.143	0.011	0.013	0.261
216053	6	7	2	8	0.001	0.167	0.004	0.009	0.001	0.535
215100	6	7	3	12	0.080	0.721	0.095	0.335	0.030	0.666
75803	7	3	3	12	0.024	0.878	0.652	0.013	0.500	0.170
8414	7	2	4	17	0.200	0.929	0.981	0.103	0.955	0.112
8473	6	1	5	17	0.920	0.075	0.908	0.633	0.145	0.155
99	7	7	4	18	0.84	0.24	0.90	0.73	0.20	0.19
262	8	10	4	18	0.476	0.857	0.999	0.402	0.941	0.403
20239	6	2	6	22	0.997	0.451	0.995	0.976	0.556	0.867
19113	8	2	6	22	0.957	0.890	0.999	0.906	0.985	0.936

In summary, the application of the LCDM to ACCESS reading was useful from some aspects. The majority of the items could differentiate between masters and non-masters, to some extent. Performance on the items could be associated with the mastery of the attributes, which could be accurately and consistently estimated. There were some aspects that picked up on the unidimensional structure of the test, such as the variability of latent classes. Despite the limitations, DCM could still be reasonable to provide some low-stakes diagnostic information. Table 35 exemplifies how information obtained from DCM can be useful. The table presents probabilities of individual attribute for selected students. Despite obtaining the same score, individual students have different patterns for the mastery for the attributes. For example, a 6th grade student with ID 20239 who got the majority of the items right on the test, has a high probability of mastering four of the attributes and the status cannot be determined for Grammar and Summary attributes. On the contrary, an 8th grade student with ID 19113 has a high probability of mastering all attributes. They were both in the support services for two years. However, between the first two students with a score of 8, student 207846 might be a master of only Grammar, while student 216053 is probably master of none of the skills. They were both in support programs for 6-7 years. It must be noted that the response patterns of these students should be scrutinized because the diagnostic capacity of the items these students correctly responded is influential on mastery probabilities (Jang, 2009b). Had the items been more diagnostic, the results might have looked different. Nevertheless, this information might still be helpful to provide support for individual students

CHAPTER V

DISCUSSION

This study showcased an implementation of a complete DCM methodology for the reading domain of a large-scale K-12 ELP test, and explored the viability of the methodology. The study was intended to answer the research questions:

(1) What are key underlying attributes represented in the ACCESS reading domain in middle grades for more advanced ELs?

(2) What DCM fits the data better?

   a. Does a general or specific restricted model better represent all items in the test?

   b. Does a Standard-based or an Expert-defined Q-matrix show better fit?

(3) To what extent is it feasible to obtain diagnostic information using DCM?

   a. What is the diagnostic capacity of the test items?

   b. To what extent can students be appropriately classified using the model?

This chapter provides a summary and synthesis of the major findings. The chapter starts with the summary of the research. In addressing the research questions, findings related to the (1) attribute specification, (2) Q-matrix development and selection, (3) model choice, (4) final model item, person and attribute statistics are discussed, especially in relation to the previous studies. Methodological considerations, as well as theoretical and practical implications, are woven into the discussion. The chapter concludes by considering the limitations of the study and directions for future research.

Summary of the Study

Despite the pressing demands and needs for diagnostic information in education (e.g., ESSA, 2015; Huff & Goodman, 2007; Lopez et al., 2019), there is a lack of diagnostic assessments (de la Torre et al., 2010) or sufficient diagnosticity in current reporting practices (e.g., Wolf et al., 2018). This gap led researchers to employ DCMs as for non-diagnostic assessments, in the interest producing diagnostic feedback. In line with this trend, the current study undertook the DCM methodology for ACCESS, an academic ELP assessment at K-12 level. If diagnostic feedback about ELs' language development could be generated, it would be extremely valuable for ELs (Wolf et al., 2016), who historically underperformed as emphasized by various researchers (e.g., Deville & Chalhoub-Deville, 2011). The reading domain for the middle graders (i.e., 6-8) was selected for the implementation as it is considered a relatively more important academic skill (Grabe, 1991) and given more weight for calculating ELs score. A test form with 27 multiple choice items, which was administered to 23,942 ELs, was used in the study. The contribution of the study was two-fold as it revealed useful information about (1) the representation of the L2 reading construct and its attributes, and (2) the extent and quality of diagnostic information using these components. The study explored the dimensions of the reading relying on the test standards and through content analysis. It established relevant attributes, and examined the nature of the relationship among them through various models. With respect to the viability, item parameters and their diagnosticity were evaluated. ELs' knowledge of each attribute (i.e., probability of mastery), overall profile (aka., classification), and average attribute mastery (i.e., difficulty of attributes) were generated. The accuracy, consistency, and plausibility of attribute probabilities and overall profile were evaluated to determine the instructional use. Finally, ELs' competence level yielded by the DCM and the original framework, IRT, were compared to provide further evidence for the appropriateness of the methodology, which uncovered important developmental patterns for L2 reading. It must be highlighted that the study was methodologically robust as it considered alternative Q-matrices, a statistical Q-matrix validation method, and a more recent modelling framework. The study reveals similar findings with prior research and conceptualizations of L2 reading. Yet, some of the findings are different due to the population studied. All these aspects are discussed in depth in the following sections.

Summary of the Findings and Implications

Attributes Underlying ACCESS Reading Domain

This study incorporated two alternative Q-matrices. There were 6 attributes in the final Expert-defined Q-matrix, which were Vocabulary, Grammar, Explicit Information and Details, Inference, Summary, and Sequences. The sample size (N = 23,942) was large enough to recover 64 classes (i.e., 2^6). The grain size can also be regarded as being proper, when the number of attributes in similar studies (i.e., 3-10) is taken into consideration. The attributes identified for the test also aligned with the definition of the academic L2 reading construct. In addition to the knowledge of vocabulary and grammar that are deemed necessary for reading comprehension (e.g., Grabe, 1991; Koda, 2007), the attributes were related to language functions. Indeed, K-12 academic language is operationalized as language functions by some language researchers (Wolf & Faulkner-Bond, 2016) such as sequencing, summarizing, inferencing, synthesizing, retelling, describing (e.g., Sato, 2007 in Frantz et al., 2014, p. 442). The attributes specified were also connected with the process of reading. For example, Koda (2007) suggests that reading comprehension involves drawing information from the text, and processing it by integrating, synthesizing, and using prior knowledge. The attributes were also akin to the attributes in some earlier DCM studies concerned with the college level L2 reading construct (e.g., Li & Suen, 2014; Sawaki et al. 2009). However, despite sharing similar processes, the complexity of the attributes varies at different levels. For instance, while an inference was associated with low-level predictions and implicit meanings in this test, it required drawing from background knowledge to arrive at conclusions at a college level test, similar to MELAB (Li & Suen, 2013). The fact that 3 independent SMEs specified attributes that corresponded to the definition of the construct in the literature implies that reading construct was well represented in this form. Moreover, the attributes defined by the experts matched with the task specifications of the test developer, which provides further evidence to their vigor.

It is also worth pointing that because vocabulary and syntax are critical for comprehension (Grabe, 1991; Harding et al. 2015; Koda 2007), they might relate to all items on a reading test. This concern was also raised by two SMEs, who later acknowledged that vocabulary and grammar were more relevant to specific items on the form, or that some items required more than baseline grammar and vocabulary knowledge. The item blueprints confirmed that some items differed, with respect to the requirement of these dimensions (e.g., understanding of complex sentence structure, technical vocabulary etc.). It is also suggested that K-12 academic English is characterized by complex structures, embedded sentences, various phrases, conjunctions, etc. (Frantz et al., 2014). Fostering awareness of academic language is regarded as being effective. Gee (2008) contends that students not only benefit from being presented a wealth of activities and examples of academic language, but also from drawing their focus on these features in a comprehensible manner. Giving feedback on how ELs performed on these attributes could initiate such awareness. Thus, the two attributes were kept in the study. However, it might still be difficult to elicit these attributes when a test is not intended to test them (Sawaki et al., 2009). In short, the usefulness of the attributes was also considered when selecting them, and because they are consistent with the L2 reading process and academic language, the attributes specified in this study can be helpful for teachers.

The attributes in the Standards-based matrix were informed by the standards of the test developer, which describe language use to communicate in academic context and content areas, and key uses that describe language functions (i.e., recall, explain, argue). Despite having more dimensions, the attributes themselves were more broadly defined, an expected issue debated by some researchers (Li & Suen, 2013; Leighton & Gierl, 2007). The standards themselves fall short of laying specific aspects of language, and carry the risk that they might be simply interpreted as vocabulary requirement for the content area in diagnostic reporting. In fact, Frantz et al. (2014) point that educators associate academic language with subject related vocabulary, which is just one aspect of it. For this reason, key uses of language were incorporated to add a little more specificity. However, key uses also collapsed some functions together and were broader than expert-defined attributes. For example, Recall was associated with identifying details and summarizing. However, the attributes in the Standard-based matrix have the advantage of being familiar to the potential users of the diagnostic information.

Q-matrix Development and Choice: Standards-based vs. Expert-defined Q-matrix

SMEs were overall assured in their mapping of the attributes. There was a consensus for the majority of the items (i.e., agreement among 5+ raters). Their selection diverged somewhat, especially for a second attribute. The rater agreement results also showed a fair amount of agreement among SMEs that is also consistent with previous research (e.g., 0.31- 0.38 among 4-5 raters in Jang, 2009a; Li & Suen, 2013; Kim, 2015). It must be noted that the SME group was larger (7 SMEs) in this study. Jang (2009a) comments that agreement rate is conditional on the size of the panel, number of items, and attributes. I also observed that there was almost perfect agreement for the Standards-based matrix because 3 SMEs only mapped items to 3 attributes. The study also hints that the agreement rate can be increased with training and discussion. The 3 SMEs who specified attributes had small group meetings and there was substantial agreement among them. Weir et al. (1990) argue that for the sake of consistency of decisions about reading skills, experienced experts from testers and linguists should be selected. The SME panel in the study was unique in such that experts from the test developer were involved, and all SMEs had expertise in language testing as well as content (i.e., English as a second language). For this reason, they were provided with statistical information for the items and distractors. The group was tacitly relying on such information when selecting attributes. For example, SME 2 explained he decided to select vocabulary for an item because the most picked distractor showed technical vocabulary was creating confusion for ELs. Thus, statistical information might be helpful for correct specification of the Q-matrix in retrofitting studies.

When attributes are identified retrospectively, it is challenging to meet the ideal Q-matrix conditions, such as ensuring that they are separated and also combined (Madison & Bradshaw, 2015). The sparsity of the attributes is another issue, which undermines the accuracy of classifications (Deonovic et al., 2019; Jang, 2009a). For the Expert-defined Q-matrix, the least measured attribute was Grammar (3 items), and it was not separated from other attributes. Therefore, results regarding grammar should be treated cautiously. Liu et al. (2018) alternatively recommends maintaining such attributes for the completeness of the Q-matrix but disregard them in interpretation. The rest of the attributes were associated with 7-13 items. All attributes were measured together, except for Sequences with Vocabulary and Inference. Despite that some of the desired conditions were met for the Expert-defined Q-matrix, results can still be improved in the future if the proportion of attributes can be balanced. For instance, more items were associated with identifying details. It must be confirmed whether this was intentional or consistent with coverage of the skill.

On the other hand, the structure of the Standards-based matrix was less desirable because it compiled two simple structure dimensions together. However, despite this property, it showed a better fit than the Expert-defined Q-matrix, which was also contrary to the expectation. The Standards-based matrix included more attributes and model parameters, which might account for the better fit. Nevertheless, the Expert-defined Q-matrix was selected as the final matrix for the study, because the attributes related to the key use dimension in the Standards-based matrix were perfectly correlated and similar proportion of ELs achieved the attributes. In their comparison of Standards-based vs. Expert-defined matrix Reid et al. (2018) also found better performance for the latter. However, the study findings should not be interpreted as the attributes in the Standards-based matrix are less related to the items or are not germane to performance of the ELs. The Expert-defined Q-matrix can be more suitable for diagnostic information (i.e., better differentiation of the attributes) and feasible for reporting (i.e., 64 vs. 256 classes).

Another aspect that differentiated the study was the statistical Q-matrix validation. Diagnostic modelling progressed, with respect to the Q-matrix validation as several methods have emerged. The study also employed an elaborate design and cross validated modifications for integrity. The method proposed changes for 4 items, and 3 of them were overspecification issues. Previous studies also reported overspecification as an issue, albeit for more items (e.g., 8 items in Kim, 2015; 7 items in Ravand, 2016). Language experts might tend to add more skills than miss them. However, overspecification might be an easier issue to resolve (Jang, 2009a).

Fewer modifications do not speak to a perfectly specified Q-matrix. Some misspecification might still be present within, which the empirical approach did not catch. The true Q-matrix is not known since the attributes were retrofitted. The final Q-matrix was the most optimal to account for 95% of the variance between the masters and non-masters of the attributes. It must be acknowledged that when the cut-off point was changed, 7 modifications were proposed by the method. However, the additional suggestions were not applicable. The Q-matrix validation is not only an iterative process, but also entails a holistic approach, as seen in selection of the cut off point (i.e., knowing discrimination power of the items). Moreover, statistical validation does not assess the theoretical grounds and compels substantive verification to complement it (Jang2009a). In this study, the experts' rationale for the attribute mapping and item specifications, were incorporated.

In summary, as highlighted by Koda (2007, 2012) and Hudson (1996) L2 reading research and defining attributes is an arduous undertaking. According to Sawaki et al. (2009) L2 reading maintains its obscurity and not all the aspects are fully understood. For this reason, authors note that separate processes might result in different attributes, definitions, and Q-matrices. The process obliges blending information from different sources (Jang, 2009a) as well as a good composition of experts, familiarity with the test and rounds of discussion. Considering the specificity, the divisibility of the attributes and practicality of reporting the Expert-defined Q-matrix was more appropriate for this study.

Model Choice: A General vs. a Constraint Model

Due to the lack of a theory for attributes (Alderson, 2005; 2007) the study applied the LCDM framework and fit a general and several constrained models. The LCDM was determined to better represent the Standards-based and Expert-defined Q-matrices at the juncture of relative and absolute fit indices, and likelihood ratio tests. However, the C-RUM and R-RUM were comparable models; some of the indices even picked these constrained models. Some prior DCM applications (e.g., Lee & Sawaki, 2009; Ravand, 2016) also confirm similar performance of compensatory and conjunctive models for L2 reading. It was reasoned that these models might be applying to different items on the test (Jang, 2009a; Chen et al., 2013; Ravand, 2016), therefore, they are "partially correct" at the test level (Chen et al., 2013, p. 136). On the other hand, the DINA, as a conjunctive model, and the DINO, a disjunctive model, underperformed in the study, which also corresponds with previous studies (e.g., Li et al., 2016; Ravand & Robitzsch, 2018; Yi, 2017). These two models employ constraints across attributes (Rupp et al., 2010), meaning, the contribution is fixed. Therefore, as more restricted models (e.g., Henson et al., 2009), they are also less suitable for the L2 reading. Similarly, the HO-DINA was an inferior model, which might be attributed to the high associations among some attributes (e.g., key uses dimension in the Standards-based Q-matrix).

The LCDM's flexibility to permit different sub-models for items is also favorable to better understand the cognitive processes (Henson et al., 2009). Specific models not only ensure easier interpretation (Lee & Luna-Bazaldua, 2019), but also improved fit (Rupp et al. 2010). Potential item level models in this study were determined by examining item parameters and also empirically tested (de la Torre & Lee, 2013). In line with the fit results, the C-RUM and R-RUM each aligned with 10 items, and for 12 of these items, sub-models were statically better than the LCDM. However, item-level modelling could not be carried out due to the limitations of the software. The close examination of each item manifested that compensatory or conjunctive relations might be contingent on the type and complexity of the skills, as well as item/task characteristics, which was also stressed by Ravand (2016). For example, all Inference items picked a conjunctive model, meaning an EL will have a lower correct response probability if this more abstract attribute is not mastered. Jang (2009a) also observed students cannot make up this skill by relying on other comprehension skills in her cognitive surveys. One exception was Item 11. A diagram accompanied this item and presumably cued the correct answer for some students. This item also required lower level inference. Similarly, Summary and Sequences that are also more complex attributes had a conjunctive relationship. On the contrary, Explicit Information and Sequences, and the combination of Vocabulary with all other attributes (i.e., except Inference) showed a compensatory relationship. Namely, despite the non-mastery of the attribute, an EL could still have a high correct response probability. Two items departed from this pattern (Items 6 & 23) because they required knowledge of multiple words and synonyms. As a result, when the complexity of Vocabulary knowledge increased, the high probability was conditional on mastery of the attribute. There were sufficient complex items in the study (i.e., 20 items out of 27). However, there is a need to balance attribute combinations before drawing final conclusions about skills associations. Yet, item specific models are still useful and can be inspected for plausibility (i.e., intended or not). They can facilitate improvement for item design and writing. Dubious item level models can draw attention to the aspects of the cognitive processes that are not fully grasped (Henson et al., 2009).

The performance of relative and absolute fit indices used in conjunction for model selection in this study aroused several important implications. The study supports earlier findings that relative fit indices are prone to model complexity, and that AIC and BIC are not always consistent with each other (Henson et al., 2009; Li et al., 2016; Lei & Li, 2016; Kunina-Habenicht et al., 2012; Yi, 2017). When this happens, likelihood ratio tests can be employed (e.g., Liu et al., 2018). However, note that saturated models are highly parameterized. Item level models can also be examined to verify the final model selections. Several researchers also recommended absolute fit indices for complementary evidence for model selection (Lei & Li, 2016; Kunina-Habenicht et al., 2012). In this study, absolute fit indices converged with relative fit results and did not show counter evidence. However, they were less useful by themselves in line with earlier research (e.g., Kunina-Habenicht et al., 2012), because they did not show much variation across models. Among them MADcor, SRMSR, MADQ3, RMSEA values were below the cut offs. This shows that some absolute fit indices might also be less suitable depending on the data features. In the study, $M\chi^2$ was picking up even small deviations due to a large sample. Likewise, MADres was apparently large, similar to in some other studies, including testlet based items (Li et al., 2016; Liu et al., 2018; Ravand & Robitzsch, 2018). Therefore, the performance of this index should be investigated in other contexts where there may be less dependencies potentially due to a common stimulus. Also, as stated previously, when alternative Q-matrices are employed, relative fit indices might not be insufficient, and a more holistic model inspection is necessary.

In short, the LCDM fit the data reasonably and the study also substantiated that some model indices perform better than others.

Feasibility: Diagnostic Capacity of the Items

Inspection of the item parameters was considered equally important with model fit, as it indicates item performance for diagnostic purposes (DiBello et al., 2007). Consistently low parameters for a specific skill bring the existence of the attribute into question (Templin & Hoffman, 2013). In this application, the LCDM generally yielded large item parameters. There were 3 exceptional items (Items 2, 9, 15) with small main effect and interaction parameters (< 0.77). These items were associated with different skills, and their blueprint specifications also matched with the Q-matrix specifications (i.e., no potential misspecification). 2 items were the most complex items of the test, each having 3 attributes. Henson et al. (2009) recommend further exploration of the LCDM for complex Q-matrices and suggested that item parameter outcomes might vary for a large number of attributes. Item parameters were actually larger for simple items on the test.

Item response probabilities revealed the average correct responsibility was 0.34 for non-masters and 0.75 for masters. Getting items right, despite lacking the attribute, was slightly higher than guessing the items (i.e., 4 options). For about a third of the items, the correct response probability for non-masters was between 0.45 and 0.60 (i.e., except for two items where the probability was 0.45-0.50). Q-matrix misspecifications, item misfit, or item facility might foster larger probabilities for non-masters (Templin &

Hoffman, 2013). These items with large intercepts in the test shared some common characteristics, such as high p-values or implausible distractors. It is possible that ELs were able to rule out some distractors. Eliminating distractors was not considered as an attribute in the study, hypothesizing that it might apply to all items. Some SMEs also expressed the strategy might be more relevant to some items. It could prove helpful in the future to explore the attribute, in order to capture aspects of item performance that were missed. Previous research has found that language learners make strategic use of distractors (e.g., Li & Suen., 2013). However, generalizing test taking strategies might be more difficult and require additional information such as verbal protocols. It is also a method effect and highlights the importance of writing good distractors. The content review of these items exhibits that ELs' knowledge of content might have played a role in their performance as well. Some of the items might have been easy for ELs with low English but high content proficiency, which was also articulated by some SMEs. Topic knowledge is shown to influence test performance but is distinguished from language knowledge in eminent language assessment frameworks (e.g., Bachman & Palmer, 1996; 2010). However, the context of the present study is quite distinct, where language proficiency is assessed within content areas (Brynes, 2008). Therefore, distinguishing content and language knowledge in this context remains a challenge in test development (Römhield et al., 2011; Llosa, 2017). Although this boundary might be unnecessary in instruction, it should be established and maintained in the assessment context for clear constructs (Frantz et al., 2014). Based on content review, it is speculated that the distinction might not have been retained for some items and affected the probability of correct response for non-masters. The analysis of domain-specific (content related)) and domain-general (language related) factors for ACCESS also substantiates this prediction. Römhield et al. (2011) report that for high proficiency test forms of ACCESS, such as the one used in this study, domain-specific knowledge is more notable and thus affects the performance more than domain-general knowledge, especially in middle grades. If content knowledge is to affect performance in reading, the performance of ELs cannot be fully captured with a single standard Q-matrix. Alternative ways to factor in such information might be necessary. However, item probabilities for non-masters were still lower when compared to previous retrofitting research (e.g., 0.47 in Kim, 2015; above 0.40 for half of the items in Li & Suen, 2013).

There were also some hard items (< 0.57) even for the masters. These items were complex except for one (Items 9, 23, 26, and 27) and were also associated with more demanding skills such as Inference, Summary, and Sequences. The low probabilities even for the masters brings misspecification or underspecified attributes to mind. However, the specified attributes were consistent with item specifications. For the simple item, adding a skill did not substantially improve the probability. These items were also the most difficult, but the least discriminating based on classical analysis, a condition also observed by Jang (2009b).

Item discrimination capacity of the items can also show the viability the DCM (DiBello, 2007). The average discrimination was 0.42, and approximately two third of the items more clearly separated the two groups. Earlier retrofitting studies also reported a difference of 40-45% between the two groups (Jang, 2009b; Kim, 2015; Li & Suen, 2013). In Kim half of the items, and in Jang 32% of the items did not differentiate the master and non-master equally well. In Jang, those items had either very low or high item facility. Similarly, in this study, items with low diagnosticity were easier or more difficult than others, and some of them also had low point biserial estimates. As already alluded, those items not performing well also had poor classical item statistics. This implies that, in addition to task features, test characteristics seem to influence diagnostic capacity and performance of the DCM. As Jang (2009b) stressed, varying item facility that is essential to order students on a continuous scale in standardized tests might not be ideal to maintain the desired diagnosticity. She suggests, item difficulty in diagnostic framework should derive from cognitive complexity. Variation among the items tapping the same attributes is also undesired (Liu et al., 2018). Poor diagnostic items do not contribute much to diagnostic information, yet the majority of the items can be useful to determine attribute mastery to some extent. The DCM analysis might also be rendered useful for item and test development (Templin & Hoffman, 2013) in this context, even though it is not used for diagnostic reporting. Q-matrix specifications mostly matched the test developer's item specifications, and analyses can be used to figure to what degree the intended skills are assessed by the items.

Feasibility: Appropriateness of Student Classifications

Characteristics of the profiles yielded important findings about the strength of the DCM. Of 64 classes, a large number of ELs (42%) were likely to be in a profile where none of the skills or all of the skills are mastered, with the former almost doubling the latter. This means that there were fewer proficient ELs. This was consistent with the characteristics of the test takers, as 20% were high ability and 33% were low ability ELs. Those mastering all skills, or all skills but one, made up 20%. When students form compact clusters (aka, flat profiles), it is suspected that the attributes are highly correlated (i.e., unidimensional) (Lee & Sawaki, 2009; Rupp et al., 2010). Previous DCM implementations for the L2 reading construct also resulted in dense, flat classes (53-75 %) (e.g., Jang et al., 2013; Lee & Sawaki, 2009; Li et al., 2016). Although the original framework of the test pointed to some issues for classification, as with the items, some limited classes had fair proportion of ELs, especially for those where 1- 3 attributes were mastered (e.g., Grammar: 9%, Grammar-Explicit Information-Summary:18%). It must be noted that variability might be dependent on the number of attributes and classes.

An accurate and consistent estimation of the classification is also vital for diagnostic skill inference (DiBello et al., 2007), which is assessed via accuracy and consistency indices in the study (P_a and P_c). At the class level, accuracy and consistency were lower (0.25 and 0.11 based on MLE), and making use of the attribute pattern was less suitable. Attribute level accuracy and classification were deemed acceptable. The developers of the indices acknowledge that the impact of different conditions, such as the complexity of a Q-matrix, on the consistency and accuracy. The magnitude of the indices in this study was consistent with the developer's anticipation for similar conditions. Moreover, the findings reveal that as the number of items associated with an attribute increases, the consistency and accuracy improves. For example, there were 3 items related to Grammar which had slightly lower accuracy and consistency. Nevertheless, individual skill classifications were more trustworthy, and they might be more practical for teacher use.

There is also a need to evaluate the performance of these indices in real data applications for language constructs. The range of indices in Ravand and Robitzsch's (2018) study with 5 attributes were close to this study. However, in two other studies for L2 reading, the pattern of P_a and P_c were reversed, which hindered the interpretability and comparability of the results.

Feasibility: Properties of the Attributes

In addition to holding higher accuracy and consistency, the six attributes separated masters and non-masters mostly. On average, only about 8% of the ELs fell into the uncertainty category. Jang (2009b) also report 6-15% uncategorized students in her application. In other words, a decision about the attribute mastery could be reached for the majority of students. Grammar was less stable because, as mentioned, there were fewer items related to this attribute. Tatsuoka argues the stability of the probabilities increases as a function of the increasing number of items for a given attribute (in Svetina et al., 2011). In this regard, the disproportionate attributes reveal shortcomings from several aspects.

The proportion of masters and non-masters (i.e., attribute difficulty) should also conform to substantive expectations (DiBello, 2007), and deviations could imply problems with the Q-matrix (i.e., misspecification). The L2 reading construct consists of skills with different levels (Harding et al., 2015), hence, the difficulty of the attributes is expected to vary. According to Grabe (2009) and Grabe and Stoller (2002), lexical and knowledge, and semantic understanding of sentences are lower level skills, while understanding the gist/summary, interpreting information, inferencing, and using background knowledge are the higher order skills. The order of the skills (i.e., easy to difficult) in the study was Grammar, Explicit Information, Summary, Vocabulary, Sequences, and Inference. The difficulty of the skills aligned with previous reading research, as syntax (e.g., Jang et al., 2013; Ravand, 2016; Svetina et al., 2011) and finding information (e.g., Jang et al., 2013; Kim, 2015; Lumley, 1993; Svetina et al., 2011) were comparatively easier than other skills. For instance, in her RSM analysis, Svetina et al. reported moderate difficulty for grammar (0.40). Similarly, more than half of the ELs mastered the attribute in the study. Explicit Information pertained to sentence or local level understanding, thus was easier than summary, which required global understanding of the texts or paragraphs. Several studies (e.g., Kim, 2015; Jang et al., 2013) also reported summary as a more difficult skill. According to Pressley (2002) summary, or understanding main ideas entails vocabulary, syntactic, and discourse knowledge, as well as comprehension strategies (as cited in Grabe, 2009), thus it is more complex. The most difficult attribute was Inference, as in other studies (e.g., Baghaei & Ravand, 2015; Jang et al., 2013; Kim, 2015; Ravand, 2016). Hammadou (1991) and Long et al. (1996) also assert that low performers fail to infer or cannot infer to the same extent with high performers. According to Mecartty (1998) inferencing varies in degree as simple and complex inferences. The attribute was associated with low level inferences in this study. Yet, it still entailed drawing from less transparent information and applying information to new situations. Thus, Inference was more demanding. Unlike the previous studies (e.g., Jang, 2009a; Kim, 2015; Svetina et al., 2011), vocabulary was difficult to master. The EL population is diverse, with respect to age and background, and they vary in their L2 comprehension behaviors and strategies (Koda, 2007). The population in this study was quite different, as the participants were younger ELs at the K-12 level, rather than college-level ELs. Thus, content related technical vocabulary might have been more challenging and abstract for the younger population. Compared to college level students, K-12 ELs are at early stages of their vocabulary development. According to Koda, the L2 reading difficulties ELs experience at different stages, and their L2 development, might vary. Weir et al. (1990) also note that it might still be difficult to achieve lower order skills. However, skill development across grades was as expected and mastery probabilities increased with the grade level. 8th graders were more likely to master skills than 6th graders in general, which supported the developmental nature of language ability.

As DCMs are suitable for multidimensional constructs, attributes should also fulfill the separability from each other. Otherwise, a student's mastery of an attribute is dependent on other attributes (Lee & Sawaki, 2009). The correlation among attributes supports that it is possible to separate them to some degree. Of 15 correlations, 4 were weak (0.-0.40), 6 were moderate (0.50- 0.76), 5 were strong (>0.87), and of those only two were above 0.90. In earlier DCM research for the L2 reading construct, higher associations were a common problem (e.g., 0.70-0.90). Correlations higher than 0.90 can be concerning for divisibility (Sessoms & Henson, 2018). The highest correlations were between Vocabulary-Inferences, Vocabulary-Sequences, and Summary-Explicit Information. Overall, these attributes showed moderate correlations with other attributes, so they can be differentiated to some extent. The high association between Summary and Explicit Information was straightforward, as they relate to deriving transparent information from different levels of the text (local vs. global). It is reasonable to assume that those ELs' who are successful at understanding global meanings can also understand local information. Also, the slightly higher relation between Summary-Syntax were apparent in other studies (e.g., 0.74 in Svenita et al., 2011). As mentioned earlier, summary entails various skills and strategies, and grammar can be a means for global comprehension (Grabe, 2009; Pressley, 2002). Some of the other correlations, such as the weak correlation between Vocabulary-Grammar and moderate correlation between Explicit Details-Grammar, were consistent with other studies (e.g., Zhang, 2012). In short, the study found some statistical evidence that the L2 reading may be multi-componential.

Finally, ability estimation under IRT was congruent with DCM person estimates despite different assumption of the models. The agreement between the models provided further support for DCM's feasibility. Moreover, the comparison was beneficial to understand the development of reading attributes. ELs' ability tended to increase as they mastered more attributes. Thus, more proficient ELs mastered more skills. For individual attributes, a master's ability was clearly higher than a non-master's. The posterior probabilities had high correlation with ability, except for Grammar (i.e., moderate correlation). The association between θ and individual attributes was also higher than previous retrofitting studies for the L2 reading (moderate association in Lee & Sawaki, 2009; Svetina et al., 2011 and 0.45-0.95 in von Davier, 2008). The divergence of grammar from other attributes was striking; and it was also observed that some advanced and intermediate learners were missing this relatively easier skill. Brunfaut and McCray (2015) found that some higher proficiency L2 readers (e.g., B1 level) focus on higher order skills more than lower order skills. For example, although they have the capacity for syntactic parsing, they do not rely on this skill and do not use it, which results in their failure to comprehend the text and answer the items correctly. This might explain why more able students seem to be non-masters of grammar. Likewise, the PL examination also supported that the majority of the beginners consistently had lower attribute probabilities, whereas advanced ELs were likely to master all attributes. 65% of Beginner ELs were in a class where none of the attributes were mastered and 25% only mastered Grammar. Advanced ELs generally mastered 4-6 attributes (82%), and Intermediate ELs mastered 2-3 attributes. However, there were some irregular patterns. Some ELs' overall ability, who had mastered none of the skills, had a θ estimate, which is as high as those who mastered 2 skills. When compared to a PL analysis, they were suspected to be 15% of intermediate ELs who were classified as non-masters of any attributes. Similarly, 8% of them were in the class where all attributes were mastered. These irregular patterns might have sourced from the low diagnostic capacity of the items and thus their mastery probabilities were estimated less precisely (Jang, 2009b). As suggested by Deonovic et al. (2019), supplemental information is useful to evaluate accuracy of the findings.

With respect to the developmental patterns, the results imply that grammar is attained earlier. When ELs reach an intermediate level their probability to master extracting details and summarizing increases substantially. Simultaneously, the students still need to improve their vocabulary and inferencing skills, but those skills are mastered at later stages. However, it must be noted that some ELs have different patterns of skill probabilities Jang et al. (2013) showed that some ELs do not acquire the reading skills in order of hierarchical difficulty. For example, in this study, there were some beginner ELs who were better at summary. Thus, strong hierarchical assumptions might be difficult for language constructs. Also, lower order skills sometimes develop at the same with higher order skills (Harding et al., 2015). Yet, probing into these patterns might be beneficial to plan learning and teaching activities.

Limitations and Future Research Directions

The DCM implementation has proven to be useful to some extent for the reading domain of ACCESS. Nevertheless, the study is not without limitations. Future research can address some of these limitations or explore the aspects that the study was not concerned with for a better understanding of L2 reading and its development.

Although the instrument used in this study was informed by evidence-based assessment design, it was not developed to be a diagnostic assessment. Therefore, some aspects of the implementation which were covered in this chapter suffered from the limitations of retrofitting DCMs to a non-diagnostic assessment. However, the study did not intend to endorse DCM methodology as a substitute to IRT. The intent was to generate low-stakes feedback in the absence of a diagnostic assessment and to take advantage of the available resources to enhance supports for ELs' reading development that is critical for their academic learning (Koda, 2007).

The scope of the study was limited to the analysis of the reading domain for middle grades and only one form, which targeted more proficient ELs. The fact that the DCM was viable to some extent and yielded some useful information in this context does not suggest that it will function in the same way for others. Therefore, results do not necessarily generalize to other grades and reading forms of the test. It would be time and resource intensive to incorporate multiple forms and grades, especially from the point of Q-matrix development. Future LCDM implementations, especially for other grades, is encouraged for generalizations about the diagnostic quality of the ACCESS system and reading construct. As described previously, Römhild et al. (2011) found domain-specific knowledge to be a stronger factor for middle grade, high-proficiency ELs. It is hypothesized in this study that such factors reduce the diagnostic capacity of some items. Yet, for lower grades, domain specific knowledge was trivial (Römhild et al., 2011). Thus, it will be relevant to explore the LCDM and diagnostic capacity in different grades. Moreover, focusing on forms with varying proficiency levels for the same grade can be informative for construct representation. The DCM methodology can be used to understand whether different forms tap into similar processes and the extent to which they relate to the intended attributes.

Another limitation of the study was training and group discussion in the Q-matrix development process, which could potentially increase the consistency among the experts and help in refining the Q-matrix. However, the schedule and workload of SMEs did not allow large group meeting. In addition, there was no training session for test developer group for the Q-matrix coding. This limitation was intended to be mitigated by providing detailed attribute definitions, instructions, and examples. However, future research should establish group meetings whenever possible. Asking for a written rationale, at minimum, can be beneficial. In addition, despite not being asked to, some SMEs indicated their thoughts about the nature of interactions among attributes underlying an item, which converged with the LCDM estimation. Rather than just coding, other input, such as contribution of attributes, their interactions can be requested from all SMEs in the future studies and compared with statistical analysis.

Future implementation can benefit from verifying or refining the Q-matrix by other sources of information. Experts, as a more able group, might not capture all processes that students engage with (Li & Suen, 2013). Matthew (1990) also argues that skills and strategies employed by learners might vary, and they “interrelate differently” for learners (p. 515). Input from ELs, in the form of cognitive surveys, can help verify the attributes specified in the study. Such input can also reveal the existence of other strategies like distractor elimination brought up by some SMEs. Cognitive surveys can also uncover the influence of content knowledge, whether reading processes vary depending on the level of content knowledge and Q-matrices should be specified differently. Due to the absence of such information, the current study worked on the assumption that processes are the same across all ELs. Another group of stakeholders that can provide beneficial input for the Q-matrix in future studies are the teachers. They observe the reading processes of students every day and can provide valuable insights for the types of the skills they use. Involvement of teachers can also increase their understanding of the construct and improve their interpretations of the results (Wolf et al., 2016).

This study treated ELs as one homogenous group. However, ELs represent a heterogonous group of students, as they come from different cultural and ethnic backgrounds and speak various languages. Numerous factors play a role in development of the L2 reading (Koda, 2012). According to Chalhoub-Deville (2009) “Viewing ELL students as a homogenous group masks important differences with regard to ELLs’ language development…” (p. 287). Jang et al. (2013), using the Fusion model, uncover important developmental patterns of immigrant and non-immigrant students. Chalhoub-Deville also contends L1 and L2 literacy, length of residence in home and immigrated countries are critical variables to understand ELs and meet their needs. In the current study, only grade level and PLs were focused to provide very limited information about developmental patterns. Future research should evaluate the impact of other factors, such as time in language support programs, and first language on classification results, and attribute mastery. Such analysis can be helpful for learning trajectories and curriculum planning.

Feedback related with students learning and process in particular is considered critical (Harding et al., 2015). Yet the utility of the diagnostic information should not only be assumed, and it should be explored. Pedagogic action is separate from quality of the diagnostic information. Availability of feedback does not ensure its use. The users, such as teachers and students, can gauge the level of utility (Templin & Hoffman, 2013). This study was limited to evaluating the DCM methodology and did not delve into the use aspect. Thus, future research should consider generating report cards and collect input from teachers and students about different aspects of the information, such as skill specificity, appropriateness for teaching/learning purposes, and clarity. Previous research with college-level ELs and their teachers manifest that information can be confusing for some students or contradict teacher perceptions (Jang, 2009b). There is also need to evaluate K-12 teachers' and ELs' perceptions in this regard, specifically to understand whether teachers interpret the information accurately, or there is a need for professional development to avoid misinterpretations (e.g., mastery of attributes suggest competency in content area). It must be ensured that results are actionable but do not encourage unintended uses or lead to unintended consequences.

There are also methodological aspects that future studies can embed to improve the results. This study incorporated one statistical Q-matrix validation approach to ensure the quality of the Q-matrix. Although the method is claimed to identify most of the misspecification, there is not a way to assess the performance of this method, nor are there still misspecifications in the Q-matrix with a single approach. Other methods exist in the literature, and thus, future research can consider multiple Q-matrix validation approaches and compare them. Despite the fact that the data included some missing responses, no imputation strategy was undertaken in the study. Research suggests that missingness can negatively affect the Q-matrix recovery (Dai et al., 2018). Future research might consider data imputation techniques and investigate whether diagnosticity is improved.

Future research can also investigate the utility of other absolute fit indices, as two of the absolute fit indices in this study failed to be informative, given the data features. There are other bivariate fit indices proposed to work well (e.g., Fisher transformed of item pair correlations (r), and log-odds ratio (l) of item pairs, see Chen et al., 2013). Future research is recommended to compare their performance.

The current study also used a large sample (N=23,942) and employed cross validation, with training and validation samples which showed the model holds. Future research can test the stability of the models estimates with smaller samples, if DCMs are intended to be used by individual schools or districts.

A comparison of the DCM mastery probabilities with IRT ability estimates and PLs pointed to some divergence. It was speculated that divergence might have stemmed from low diagnostic items. However, some fit might have also occurred due to individuals. A final possible future research area can be incorporating person fit indices (e.g., Liu et al., 2009). The proportion of aberrant students can also hint to the viability of the methodology, as it indicates misclassification (Liu et al., 2018).

Conclusion

In this study, the suitability of ACCESS reading for diagnostic information was explored using a multidimensional methodology, the DCM. Such undertaking might be beneficial for low stakes purposes, but it was not without limitations. Specifically, the information yielded by the mastery probabilities were estimated with acceptable accuracy and were consistent with ELs' reported proficiency levels. Therefore, DCMs can be informative, to some extent, to uncover ELs' reading related problems and understand the development of reading skills for planning learning activities. By better proportioning the attributes, increasing the diagnostic capacity of the items, and especially by further reducing the impact of content specific knowledge, and creating distinguishing distractors, the quality of the information could be potentially increased. Aside from insights about ELs' reading performance and their instructional utility, this study provided critical information about the construct itself, and thus can be helpful for test design and development and construct representation (Liu et al., 2018; Rupp et al., 2010; Svetina et al., 2011). The study provided further evidence for the L2 reading representation, its divisibility, and the difficulty of the attributes. The attributes should be attended in test design at the very least not to underrepresent the construct (Alderson, 2000).

REFERENCES

Abedi, J. (2008) Measuring students' level of English proficiency: Educational significance and assessment requirements, *Educational Assessment*, *13*, 193-214.

Abedi, J. (2010). Research and recommendations for formative assessment with English language learners. In H. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp.181-197). Routledge.

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automated Control*, 19, 716-723.

Alderson, J. C. (1990a). Testing reading comprehension skills (part one). *Reading in a Foreign Language*, *6*(2), 425-438.

Alderson, J. C. (1990b). Testing reading comprehension skills (part two): Getting students to talk about taking a reading test (a pilot study). *Reading in a Foreign Language*, *7*(1), 465-503.

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment.* Continuum.

Alderson, J. C. (2007). The challenge of (diagnostic) testing: Do we know what we are measuring? In J. Fox, M. Wesche, B. Doreen & L. Cheng (Eds.), *Language testing reconsidered* (pp. 21-39). University of Ottawa Press.

Alderson, J. C. (2010). Cognitive diagnosis and Q-matrices in language assessment: A commentary. *Language Assessment Quarterly*, *7*(1), 96-103.

Alderson, J. C., Brunfaut, T., & Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics*, *36*(2), 236-260.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge University Press.

Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language*, *5*(2), 253-270.

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

Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing*, *8*(1), 41-66.

Antón, M. (2018). Dynamic diagnosis of second language abilities. In J. P. Lantolf, M. E. Poehner & M. Swain (Eds.) *The Routledge handbook of sociocultural theory and second language development* (pp. 310-323). Routledge.

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.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Baghaei, P., & Ravand, H. (2015). A cognitive processing model of reading comprehension in English as a foreign language using the linear logistic test model. *Learning and Individual Differences*, *43*, 100-105.

Bailey, A. L., & Carroll, P. E. (2015). Assessment of English language learners in the era of new academic content standards. *Review of Research in Education*, *39*(1), 253-294.

Bauman, J., Boals, T., Cranley, E., Gottlieb, M., & Kenyon, D. (2007). Assessing comprehension and communication in English state to state for English language learners (access for ells). In J. Abedi (Ed.) *English language proficiency assessment in the nation: Current status and future practice* (pp. 81-91). University of California, Davis.

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

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics*, *25*, 133-150.

Blais, J. G., & Laurier, M. D. (1995). The dimensionality of a placement test from several analytical perspectives. *Language Testing*, *12*(1), 72-98.

Bolt, D. (2019). Bifactor mirt as an appealing and related alternative to cdms in the presence of skill attribute continuity. In M. von Davier & Y.S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 395-420). Springer.

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

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice*, *33*(1), 2-14.

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes whilst completing reading tasks: A mixed-method eye-tracking and stimulated recall study* [ARAGs Research Report]. British Council. https://www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final_0.pdf

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning, 47*(3*),* 423-466*.*

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*(2), 119-157.

Buck, G., VanEssen, T., Tatsuoka, K., Kostin, I., Lutz, D., & Phelps, M. (1998). *Development, selection and validation of a set of cognitive and linguistic attributes for the SAT I verbal: Sentence completion section*. Educational Testing Service. https://www.ets.org/Media/Research/pdf/RR-98-23.pdf

Bunch, M. B. (2011). Testing English language learners under no child left behind. *Language Testing*, *28*(3), 323-341.

Byrnes, H. (2008). Assessing content and language. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (Volume 7) (pp. 37-52). Springer.

Camara, W. J., & Lane, S. (2006). A historical perspective and current views on the standards for educational and psychological testing. *Educational Measurement: Issues and Practice*, *25*(3), 35-41.

Carroll, P. E., & Bailey, A. L. (2016). Do decision rules matter? A descriptive study of English language proficiency assessment classifications for English-language learners and native English speakers in fifth grade. *Language Testing*, *33*(1), 23-52.

Center for Applied Linguistics [CAL] (2017). *Annual technical report for ACCESS for ELLs English language proficiency test, Series 400, 2015–2016 Administration* [Report No: 12B]. WIDA Consortium.

Chalhoub-Deville, M. (2009). Standards-based assessment in the U.S: social and educational impact. In L. Taylor & C. J. Weir (Eds.), *Language testing matters: Investigating the wider social and educational impact of assessment* (pp. 281-300). Cambridge University Press.

Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. *Language Testing*, *33*(4), 453-472.

Chalhoub-Deville, M., & Deville, C. (2008). Nationally mandated testing for accountability: English language learners in the US. In B. Spolsky, & F. M. Hult (Eds.), *The handbook of educational linguistics* (pp. 510-522). Blackwell Publishing.

Chen, H., & Chen, J. (2016). Retrofitting non-cognitive diagnostic reading assessment under the generalized dina model framework. *Language Assessment Quarterly*, *13*(3), 218-230.

Chen, J. (2017). A residual-based approach to validate q-matrix specifications. *Applied Psychological Measurement*, *41*(4), 277-293.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*(2), 123-140.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265-289.

Cohen, A. D., & Upton, T. A. (2006). *Strategies in responding to the new TOEFL reading tasks*. Educational Testing Service. https://www.ets.org/Media/Research/pdf/RR-06-06.pdf

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q 3: Identification of local dependence in the Rasch model using residual correlations. *Applied psychological measurement*, *41*(3), 178-194.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory Mason*. Cengage Learning.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302.

Cui, Y., Gierl, M. J., & Chang, H. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*(1), 19-38.

Dai, S., Svetina, D., & Chen, C. (2018). Investigation of missing responses in q-matrix validation. *Applied Psychological Measurement*, *42*(8), 660-676.

Davis, F. B. (1968). Research in comprehension in reading. *Reading Research Quarterly, 3*, 499-545.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Publications.

de la Torre, J. (2008). An empirically based method of q-matrix validation for the dina model: Development and applications. *Journal of Educational Measurement*, *45*(4), 343-362.

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*(3), 163-183.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179-199.

de la Torre, J., & Akbay, L. (2019). Implementation of cognitive diagnosis modeling using the gdina r package. *Eurasian Journal of Educational Research*, *80*, 171-192.

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical q-matrix validation. *Psychometrika*, *81*(2), 253-273.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333-353.

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.

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

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, *20*(2), 89-97.

de la Torre, J., & Michen, N. D. (2019). The g-dina model framework. In M. von Davier & Y.S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 155-170). Springer.

Deonovic, B., Chopade, P., Yudelson, M., de la Torre, J., & von Davier, A. (2019). Application of cognitive diagnostic models to learning and assessment systems. In M. von Davier & Y.S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 461-488). Springer.

Deville, C., & Chalhoub-Deville, M. (2011). Accountability assessment under no child left behind: Agenda, practice, and future. *Language Testing*, 28(3), 307-321.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26) (pp. 970-1030). Elsevier.

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

Edelenbos, P., & Kubanek-German, A. (2004). Teacher assessment: the concept of diagnostic competence. *Language Testing*, *21*(3), 259-283.

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

Faulkner-Bond, M. M. (2016). *Who is like whom? Reclassification and performance patterns for different groupings of English learners* [Unpublished Doctoral Dissertation]. The University of Massachusetts.

Faulkner-Bond, M., & Forte, E. (2016). English learners and accountability: the promise, pitfalls, and peculiarity of assessing language minorities via large-scale assessment. In C. S. Well & M. Faulkner-Bond (Eds.), *Educational measurement: From foundations to future* (pp. 395-415). The Guilford Press.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions*. John Wiley & Sons.

Fox, J., & Fairbairn, S. (2011). Test review: ACCESS for ELLs. *Language Testing*, *28*(3), 425-431.

Frantz, R. S., Bailey, A. L., Starr, L., & Perea, L. (2014). Measuring academic language proficiency in school-age English language proficiency assessments under new college and career readiness standards in the United States. *Language Assessment Quarterly*, *11*(4), 432-457.

Gee, J. P. (2008). What is academic language. In A.S. Rosebery (Ed.), *Teaching science to English language learners: Building on students' strengths*, (pp. 57-70). NSTA Press.

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.

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.

Gierl, M. J., Roberts, M., Alves, C., & Gotzmann, A. (2009, April 14-16). *Using judgments from content specialists to develop cognitive models for diagnostic assessments* [Paper Presentation]. Annual meeting of the National Council on Measurement in Education, San Diego, CA.

Glaser, R., & Nitko, A. J. (1970). *Measurement in learning and instruction*. University of Pittsburgh. https://files.eric.ed.gov/fulltext/ED038873.pdf

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly*, *25*(3), 375-406.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge University Press.

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading*. Routledge.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204-229.

Haberman, S. J., & von Davier, M. (2007). Some notes on models for cognitively based skills diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26) (pp. 1031-1038). Elsevier.

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

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, *18*(4), 5-9.

Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity arguments for accountability testing. *Yearbook of the National Society for the Study of Education*, *104*(2), 1-34

Hammadou, J. (1991). Interrelationships among prior knowledge, inference, and language proficiency in foreign language reading. *The Modern Language Journal*, *75*(1), 27-38.

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, *32*(3), 317-336.

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

Hedgcock, J. S., & Ferris, D. R. (2009). *Teaching readers of English: Students, texts, and contexts*. Routledge.

Henning, G. (1987). *A guide to language testing: development, evaluation, research*. Newberry House.

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, *9*(1), 1-11.

Henson, R. A. (2009). Diagnostic classification models: Thoughts and future directions. *Measurement: Interdisciplinary Research and Perspectives*, 7(1), 34-36.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*(4), 262-277.

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

Henson, R., & Templin, J. (2007, April 10-12). *Large-scale language assessment using cognitive diagnosis models* [Paper presentation] Annual meeting of National Council for Mesurement in Education, Chicago, IL.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191-210.

Hessamy, G., & Sadeghi, S. (2013). The relative difficulty and significance of reading skills. *International journal of English Language Education*, *1*(3), 208-222.

Hu, J., Miller, M. D., Huggins-Manley, A. C., & Chen, Y. H. (2016). Evaluation of model fit in cognitive diagnosis models. *International Journal of Testing*, *16*(2), 119-141.

Hudson, T. (1996). *Assessing second language academic reading from a communicative competence perspective: Relevance for TOEFL 2000*. Educational Testing Service. https://www.ets.org/Media/Research/pdf/RM-96-06.pdf

Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research, and Evaluation*, *15*, 1-7.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 19-60). Cambridge University Press.

Huhta, A. (2008). Diagnostic and formative assessment. In B. Spolsky & F. M. Hult (Eds.), *The handbook of educational linguistics* (pp. 469-482). Wiley Blackwell.

In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC test: A multiple-sample analysis. *Language Testing*, *29*(1), 131-152.

Jang, E. E. (2009a). Demystifying a q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly*, *6*(3), 210-238.

Jang, E. E. (2009b). Cognitive diagnostic assessment of L2 reading comprehension ability: validity arguments for fusion model application to LanguEdge assessment. *Language Testing*, *26*(1), 31-73.

Jang, E. E. (2017). Cognitive aspects of language assessment. In E. Shohamy, I. Or, & S. May (Eds.), *Language Testing and Assessment* (pp. 163-177). Springer.

Jang, E. E., Dunlop, M., Park, G., & van der Boom, E. H. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? *Language Testing*, *32*(3), 359-383.

Jang, E. E., Dunlop, M., Wagner, M., Kim, Y. H., & Gu, Z. (2013). Elementary school ELLs' reading skill profiles using cognitive diagnosis modeling: roles of length of residence and home language environment. *Language Learning*, *63*(3), 400-436.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272.

Kang, C., Yang, Y., & Zeng, P. (2019). Q-Matrix refinement based on item fit statistic rmsea. *Applied Psychological Measurement*, *43*(7), 527-542.

Kim, A. A., Kondo, A., Blair, A., Mancilla, L., Chapman, M., & Wilmes, C. (2016). *Interpretation and use of K-12 language proficiency assessment score reports: Perspectives of educators and parents*. University of Wisconsin-Madison. https://wcer.wisc.edu/docs/working-papers/Working_Paper_No_2016_8.pdf

Kim, A. Y. (2009). Investigating second language reading components: Reading for different types of meaning. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *9*(2), 1-28.

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

Kim, J. (2011). *Relationships among and between ELL status, demographic characteristics, enrollment history, and school persistence.* University of California Los Angeles. https://files.eric.ed.gov/fulltext/ED527529.pdf

Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. *Language learning, 57*(1), 1-44.

Koda, K. (2012). How to do research on second language reading. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 158-179). Blackwell Publishing.

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

Kunnan, A. J., & Jang, E. E. (2009). Diagnostic feedback in language assessment. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 610-627). Wiley Blackwell.

Kuriakose, A. (2011). *The factor structure of the English language development assessment: A confirmatory factor analysis* (Publication No: ED535975) [Doctoral dissertation, Arizona State University]. ProQuest Dissertations Publishing.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.

Lantolf, J., & Poehner, M. E. (2004). Dynamic assessment in the language classroom. The Pennsylvania State University.

Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing*, *11*(2), 144-177.

Lee, Y. W., & Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly*, *6*(3), 239-263.

Lee, Y., & Luna-Bazaldua, D. A. (2019). How to conduct a study with diagnostic models. In M. von Davier & Y.S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 525-548). Springer.

Lei, P. W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and q-matrices. *Applied Psychological Measurement*, *40*(6), 405-417.

Leighton, J. P. (2009). Mistaken impressions of large-scale cognitive diagnostic testing. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 219-246). American Psychological Association.

Leighton, J. P., & Gierl, M. J. (2007). Why cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: theory and applications* (pp. 9-18). Cambridge University Press.

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

Leung, C. (2007). Dynamic assessment: Assessment *for* and *as* teaching? *Language Assessment Quarterly*, *4*(3), 257–278.

Li, H. (2011). Evaluating language group differences in the subskills of reading using a cognitive diagnostic modeling and differential skill functioning approach [Unpublished doctoral dissertation]. Penn State University, State College, PA.

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.

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

Liu, H. H. T. (2014). The conceptualization and operationalization of diagnostic testing in second and foreign language assessment. *Working Papers in TESOL & Applied Linguistics*, *14*(1), 1-12.

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied psychological measurement*, *33*(8), 579-598.

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.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of q-matrix. *Applied Psychological Measurement*, *36*(7), 548-564.

Llosa, L. (2017). Assessing students' content knowledge and language proficiency. In E. Shohamy, I. G. Or & S. May (Eds.), *Encyclopedia of language and education* (pp. 3-14). Springer.

Long, D. L., Seely, M. R., Oppy, B. J., & Golding, J. M. (1996). The role of inferential processing in reading ability. In B. K. Britton & A. C. Graesser (Eds.), *Models of understanding text* (pp. 189-214). Psychology Press.

Lopez, A. (2019, March 4-7). *Empowering K-12 teachers to make better use of high-stakes summative elp assessments* [Paper presentation]. Annual meeting of Language Testing Research Colloquium, Atlanta, GA.

Lopez, A. A., Pooler, E., & Linquanti, R. (2016). *Key issues and opportunities in the initial identification and classification of English learners*. Educational Testing Service. https://files.eric.ed.gov/fulltext/EJ1124769.pdf

Luecht, R. M. (2003, April 22-24). *Applications of multidimensional diagnostic scoring for certification and licensure tests* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, Chicago, IL.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An eap example. *Language Testing*, *10*(3), 211-234.

Ma, W. (2019). Cognitive diagnosis modeling using the gdina r package. In M. von Davier & Y.S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 593-601). Springer.

Ma, W., & de la Torre, J. (2020). GDINA: *The generalized DINA model framework* (Version 2.8.0.). R package. https://CRAN.R-project.org/package=GDINA

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*(3), 200-217.

Macmillan, F. M. (2016). Assessing reading. In D. Tsagari & B. Jayanti (Eds.), *Handbook of second language assessment* (Volume 12) (pp. 113-130). De Gruyter Mouton.

Madison, M. J. (2019). Reliably assessing growth with longitudinal diagnostic classification models. *Educational Measurement: Issues and Practice*, *38*(2), 68-78.

Madison, M. J., & Bradshaw, L. P. (2015). The effects of q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement*, *75*(3), 491-511.

Matthews, M. (1990). Skill taxonomies and problems for the testing of reading. *Reading in a Foreign Language*, *7*(1), 511-517.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 71-101.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*(4), 305-328.

McDonald, R. P., & Mok, M. M.C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30*, 23-40.

Mecartty, F. H. (1998). The Effects of proficiency level and passage content on reading skills assessment 1. *Foreign Language Annals*, *31*(4), 517-534.

Menken, K. (2008). *English learners left behind: Standardized testing as language policy*. Multilingual Matters.

Menken, K. (2010). NCLB and English language learners: Challenges and consequences. *Theory into Practice*, *49*(2), 121-128.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (Vol. 1989) (pp. 13–104). American Council on Education.

Munby, J. (1978). *Communicative syllabus design*. Cambridge University Press.

Nájera, P., Sorrel, M. A., & Abad, F. J. (2019). Reconsidering cutoff points in the general method of empirical q-matrix validation. *Educational and Psychological Measurement*, *79*(4), 727-753.

National Center for Education Statistics [NCES]. (2019). *The condition of education: English language learners in public schools*. Institute of Educational Sciences. https://nces.ed.gov/programs/coe/indicator_cgf.asp

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

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Routledge.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing*, *20*(1), 26-56.

Poehner, M. E., & Infante, P. (2016). Dynamic assessment in the language classroom. In D. Tsagari & B. Jayanti (Eds.), *Handbook of second language assessment* (Volume 12) (pp. 275-290). De Gruyter Mouton.

Puhan, G., Sinharay, S., Haberman, S., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: An evaluation of methods using empirical data. *Applied Measurement in Education*, *23*(3), 266-285.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, *34*(8), 782–799.

Ravand, H., & Robitzsch, A. (2018). Cognitive diagnostic model of best choice: A study of reading comprehension. *Educational Psychology*, *38*(10), 1255–1277.

Reid, A. M., Hoeve, K., & Henson, R. A. (2018, April 12-16). *Fitting a diagnostic assessment to standards-defined skills versus expert-defined skills* [Paper presentation]. Annual meeting of National Council on Measurement in Education, New York, NY.

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

Römhild, A., Kenyon, D., & MacGregor, D. (2011). Exploring domain-general and domain-specific linguistic knowledge in the assessment of academic English language proficiency. *Language Assessment Quarterly*, *8*(3), 213-228.

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

Rost, D. H. (1993). Assessing different components of reading comprehension: fact or fiction? *Language Testing*, *10*(1), 79-92.

Rupp, A. A., & Templin, J. (2008). The effects of q-matrix misspecification on parameter estimates and classification accuracy in the dina model. *Educational and Psychological Measurement*, *68*(1), 78-96.

Rupp, A. A., & Templin, J. L. (2011). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*(4), 219-262.

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

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, *24*(3), 355-390.

Sawaki, Y., Kim, H. J., & Gentile, C. (2009). Q-matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly*, *6*(3), 190-209.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL internet- based test. *Language Testing*, *26*(1), 5-30.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological assessment*, *8*(4), 350.

Schwarz, G. (1976). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement*, *16*(1), 1-17.

Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *The Modern Language Journal*, *76*(4), 513-521.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107-120.

Sinharay, S. (2014). Analysis of added value of subscores with respect to classification. *Journal of Educational Measurement*, *51*(2), 212-222.

Sinharay, S., & Haberman, S. J. (2008). *Reporting subscores: A survey.* Educational Testing Service. https://www.ets.org/Media/Research/pdf/RM-08-18.pdf

Skaggs, G., Wilkins, J. L., & Hein, S. F. (2016). Grain size and parameter recovery with TIMSS and the general diagnostic model. *International Journal of Testing*, *16*(4), 310-330.

Slama, R. B. (2014). Investigating whether and when English learners are reclassified into mainstream classrooms in the United States: A discrete-time survival analysis. *American Educational Research Journal*, 51(2), 220-252.

Song, M. Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing*, *25*(4), 435-464.

Stiggins, R. J. (2001). The unfulfilled promise of classroom assessment. *Educational Measurement: Issues and Practice*, *20*(3), 5-15.

Stout, W., Henson, R. DiBello, L., & Shear B. (2019). The reparameterized unified model system: a diagnostic assessment modelling approach. In M. von Davier & Y.S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 47-80). Springer.

Svetina, D., Gorin, J. S., & Tatsuoka, K. K. (2011). Defining and comparing the reading comprehension construct: A cognitive-psychometric modeling approach. *International Journal of Testing*, *11*(1), 1-23.

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

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods*, *11*(3), 287-305.

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317-339.

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, *32*(2), 37-50.

Thissen, D. (2016). Bad questions: An essay involving item response theory. *Journal of Educational and Behavioral Statistics*, *41*(1), 81-8.

Turner, C. E., & Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & B. Jayanti (Eds.), *Handbook of second language assessment* (Volume 12) (pp. 255-274). De Gruyter Mouton.

U.S. Department of Education (n.d.). *Our nation's English learners*. https://www2.ed.gov/datastory/el-characteristics/index.html

U.S. Department of Education. (2004). Stronger accountability: *Testing for results: Helping families, schools, and communities under- stand and improve student achievement.* http://www.ed.gov/nclb/accountability/ayp/testingforresults.html

U.S. Congress. (2015). *Every student succeeds act*. Washington, DC: Author.

U.S. Department of Education, & U.S. Department of Justice. (2015). *Dear colleague letter: English learner students and limited English proficient parents*. http://www2.ed.gov/about/offices/list/ocr/letters/colleague-el-201501.pdf

U.S. Department of Education (2018). *A state's guide to the U.S. department of education's peer review process.* https://www2.ed.gov/admins/lead/account/saa/assessmentpeerreview.pdf

Urquhart, A. H., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice*. Routledge.

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

von Davier, M. (2014). The dina model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology*, *67*(1), 49-71.

von Davier, M. (2014b). *The log-linear cognitive daignostic models (lcdm) as a special case of the general diagnostic model (gdm)*. Educational Testing Service. https://onlinelibrary.wiley.com/doi/epdf/10.1002/ets2.12043

von Davier, M., & Lee, Y. S. (2019). *Handbook of diagnostic classification models*. Springer.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational measurement*, *24*(3), 185-201.

Walqui, A., Koelsch, N., & Hamburger, L, et al. (2010). *What are we doing to middle school* English *learners? Findings and recommendations for change from a study of California EL programs*. WestEd. https://www.wested.org/online_pubs/PD-10-02-full.pdf

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*, 165-187.

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, *52*(4), 457-476.

Weir, C. J. (2005). *Language testing and validation*. Palgrave McMillan.

Weir, C. J., Hughes, A., & Porter, D. (1990). Reading skills: hierarchies Implications relationships and identifiability. *Reading in a Foreign Language*, *7*(1), 505-510.

WIDA Consortium (n.d.). *ACCESS for ELLs Scores and Reports*. Board of Regents of the University of Wisconsin System. https://wida.wisc.edu/assess/access/scores-reports

WIDA Consortium (n.d.). *ACCESS tests*. https://wida.wisc.edu/assess/access/tests

WIDA Consortium. (2007). *English Language Proficiency Standards and Resource Guide, 2007 Edition, PreKindergarten through Grade 12*. Board of Regents of the University of Wisconsin System. https://wida.wisc.edu/sites/default/files/resource/2007-ELPS-Resource-Guide.pdf

WIDA Consortium. (2012). *2012 Amplification of the English Language Development Standards Kindergarten–Grade 12*. Board of Regents of the University of Wisconsin System. https://wida.wisc.edu/sites/default/files/resource/2012-ELD-Standards.pdf

WIDA Consortium. (2014). *The WIDA standards framework and its theoretical foundations*. Board of Regents of the University of Wisconsin System. https://wida.wisc.edu/sites/default/files/resource/WIDA-Standards-Framework-and-its-Theoretical-Foundations.pdf

WIDA Consortium. (2016). *Can do descriptors: Key uses edition*. Board of Regents of the University of Wisconsin System. Board of Regents of the University of Wisconsin System. https://wida.wisc.edu/sites/default/files/resource/CanDo-KeyUses-Gr-6-8.pdf

WIDA Consortium. (2019). *ACCESS for ELLs sample items user guide*. Board of Regents of the University of Wisconsin System. https://wida.wisc.edu/sites/default/files/resource/ACCESS-Paper-Sample-Items-User-Guide.pdf

WIDA Consortium. (2020). *ACCESS for ELLs interpretive guide for score report*. Board of Regents of the University of Wisconsin System. Retrieved from https://wida.wisc.edu/sites/default/files/resource/Interpretive-Guide.pdf

Willse, J. T. (2018). CTT: Classical Test Theory Functions (Version 2.3.3). R package. https://CRAN.R-project.org/package=CTT

Wolf, M. K., Farnsworth, T., & Herman, J. (2008a). Validity issues in assessing English language learners' language proficiency. *Educational Assessment*, *13*(2–3), 80-107.

Wolf, M. K., Kao, J., Herman, J., Bachman, L. F., Bailey, A., Bachman, P. L., Farnsworth, T., & Chang, S. M. (2008b). Issues in assessing English language learners: English language proficiency measures and accommodation uses. UCLA. https://files.eric.ed.gov/fulltext/ED502284.pdf

Wolf, M. K., & Faulkner-Bond, M. (2016). Validating English language proficiency assessment uses for English learners: Academic language proficiency and content assessment performance. *Educational Measurement: Issues and Practice*, *35*(2), 6-18.

Wolf, M. K., Guzman-Orth, D., & Hauck, M. C. (2016). *Next-generation summative English language proficiency assessments for English learners: Priorities for policy and research*. Educational Testing Service. https://files.eric.ed.gov/fulltext/EJ1124766.pdf

Wyse, A. E., & Hao, S. (2012). An evaluation of item response theory classification accuracy and consistency indices. *Applied Psychological Measurement*, *36*(7), 602-624.

Xie, Q. (2017). Diagnosing university students' academic writing in English: Is cognitive diagnostic modelling the way forward? *Educational Psychology*, *37*(1), 26–47.

Xu, X., & von Davier, M. (2008). *Fitting the structured general diagnostic model to NAEP data*. Educational Testing Service. https://files.eric.ed.gov/fulltext/EJ1111272.pdf

Yamaguchi, K., & Okada, K. (2018). Comparison among cognitive diagnostic models for the TIMSS 2007 fourth grade mathematics assessment. *PloS One*, *13*(2), 1-17.

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

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

Yi, Y. S. (2017). Probing the relative importance of different attributes in L2 reading and listening comprehension items: An application of cognitive diagnostic models. *Language Testing*, *34*(3), 337–355.

Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. *The Modern Language Journal*, *96*(4), 558-575.

APPENDIX A

DCM STUDIES USING LANGUAGE ASSESSMENTS

	Test	**Domain**	**Model**	***N***	**Q-matrix**	**Attributes**	**Model evaluation**
Buck et al. (1997)	TOEIC	Reading (J = 35)	Rule Space	5,000	• Literature Review • Teaching-testing Experience • Observation of Test-taking	15 primary attributes and 14 interactions	NA
Henson & Templin (2007)	ECPE	Grammar (J = 30)	NC-RUM DINA	2,922	• Exploratory Factor Analysis • Literature Review	(1) Morpho-syntactic Knowledge (2) Cohesive Knowledge (3) Lexical Knowledge	NA
von Davier (2008)	TOEFL iBT	Reading (J = 40) Listening (J = 34)	GDM GPCM	3,139	• SME Judgement	*Reading* (1) Word Meaning (2) Specific Information (3) Connecting Information (4) Synthesize and Organize *Listening* (1) General Information (2) Specific Information (3) Pragmatic & Text Structure (4) Inference & Connections	Test-rest reliability across forms, relative fit (-2LL), comparison of IRT and GDM estimates etc.
Jang (2009a)	LanguEdge	Reading (J = 37)	Fusion	NA	• Literature Review • Blueprint Analysis • Text Analysis • Statistical Item Analysis • Dimensionality Analysis • Think-aloud Protocols • SME Judgement	(1) Context-dependent Vocab. (2) Context-independent Vocab. (3) Syntactic, Semantic Links (4) Explicit Information (5) Implicit Information (6) Inferencing (7) Summarizing (8) Mapping Contrasting Ideas into Framework	Absolute fit such as comparison of observed and predict parameters, classification consistencies etc.

Jang (2009b)	LanguEdge	Reading (J = 39)	Fusion	2,703	• Same as Jang (2009a)	Same as Jang (2009a)	Convergence of MCMC, MADcor, classification consistency, distribution of skill profiles etc.
Sawaki et al. (2009)	TOEFL iBT	Listening (J = 34) Reading (J = 40)	Fusion	6,000	• SME Judgment • Literature Review • Test Framework Review	*Reading* (1) Word Meaning (2) Specific Information (3) Connecting Information (4) Synthesizing & Organizing *Listening* (1) Understanding General Information (2) Understanding Specific Information (3) Understanding Text Structure & Speaker Intention (4) Connecting Ideas	Comparison between observed and predicted parameters, classification consistencies etc.
Lee & Sawaki (2009)	TOEFL iBT	Listening (J = 34) Reading (J = 40)	GDM Fusion LCA	3,139	• Adapted from Sawaki et al. (2009)		Skill classification consistency, test-retest reliability across forms, distribution of skill mastery probabilities, comparison of examinees classified into profiles, RMSEA for item correlations etc.
Kim (2011)	TOEFL iBT	Writing (1 task)	R-RUM	480	• Verbal Protocols • SME Judgement	(1) Content Fulfilment (2) Organizational Effectiveness (3) Grammatical Knowledge (4) Vocabulary Use (5) Mechanics	Similar to Jang(2009b)

Jang et al. (2013)	K-12 Literacy Test	Reading Writing (J = 40)	R-RUM	18, 059	• Content/task Analysis with SMEs	(1) Textually-implicit Info (2) Textually-explicit Info (3) Inferencing (4) Grammar Knowledge (5) Summarizing Main Ideas (6) Vocabulary	Similar to Jang (2009b)
Li & Suen (2013)	MELAB	Reading (J = 20)	Fusion	2,019	• Literature Review • Think-aloud Protocols • SME Judgment	(1) Vocabulary (2) Syntax (3) Extracting Explicit Info (4) Understanding Implicit Info	Similar to Jang(2009b)
Jang et al. (2015)	K-12 Literacy test	Reading Writing (J = 40)	R-RUM	44	• Adapted from Jang et al. (2013)		NA
Kim (2015)	University developed ESL Placement Test	Reading (J = 30)	Fusion	1,982	• Literature Review • Construct Framework/ Model • SME Judgement	(1) Lexical Meaning (2) Cohesive Meaning (3) Sentence Meaning (4) Paragraph Meaning (5) Pragmatic Meaning (6) Identifying Word Meaning (7) Finding Information (8) Skimming (9) Summarizing (10) Inferencing	Similar to Jang (2009b)
Li et al. (2016)	MELAB	Reading (J = 20)	G-DINA DINO ACDM DINA R-RUM	2,019	• Adopted from Li & Suen (2013)		-2LL AIC BIC $M\chi^2$ MADcor MADres
Ravand (2016)	General English test of National University Exam	Reading (J = 20)	G-DINA	10,000	• SEM judgement • GDI	(1) Detail (2) Inference (3) Main Idea (4) Syntax (5) Vocabulary	$M\chi^2$ MADcor MADRES MADQ3 RMSEA

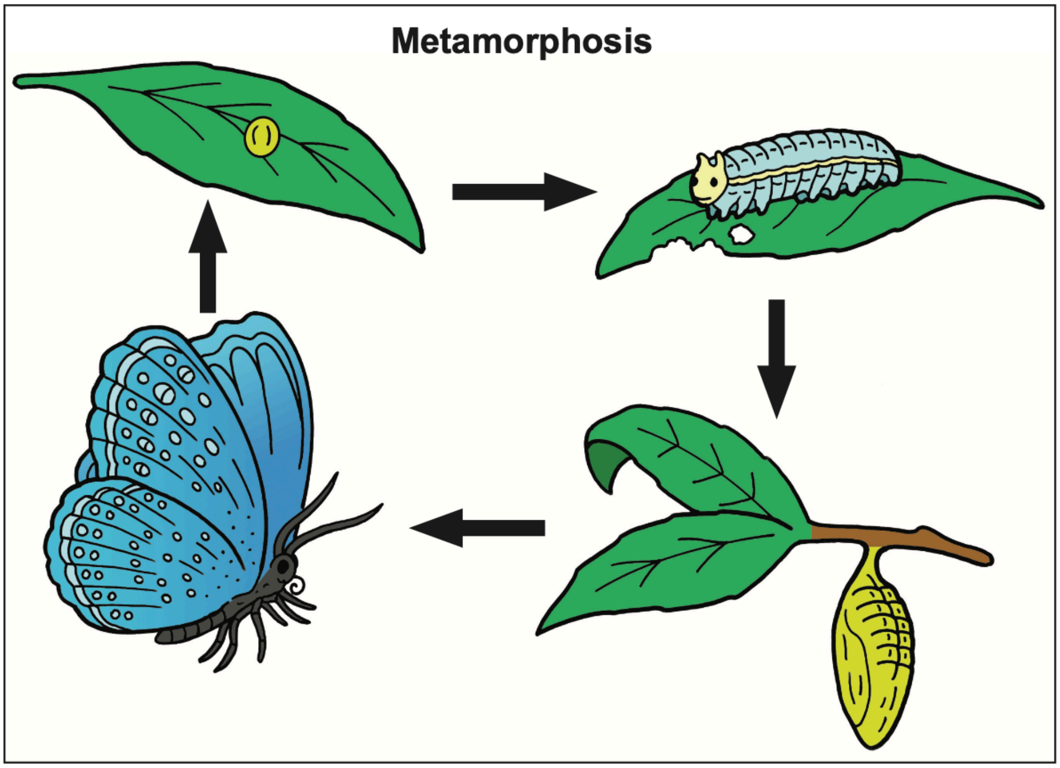
Yi (2017a)	ECPE	Grammar (J = 30)	LCDM DINA DINO NIDO C-RUM	2,922	• Adapted from Henson & Templin (2007)		AIC BIC Adjusted BIC Item-correlation RMSEs Distribution of attributes
Yi (2017b)	TOEFL	Listening (J = 34) Reading (J = 39)	DINA DINO NIDO C-RUM	3,139	• Adapted from Sawaki et al. (2009)		AIC BIC Item-correlation RMSEs
Xie (2017)	University developed placement test	Writing (1 task)	R-RUM	472	• Adapted from Kim (2011)		Similar Jang (200b)
Aryadoust (2018)	Singapore-Cambridge General Certificate Education	Listening (J = 32)	DINA G-DINA DINO HO-DINA R-RUM	205	• Literature review • Think-aloud Protocol • Eye-tracking	(1) Eliminating Inaccurate Information (2) Paraphrasing (3) Making Pragmatic Inferences (4) Using Word knowledge for Inferences (5) Making Inferences (6) Understanding Surface Information (7) Understanding contradiction (8) Making Anaphoric Moves (9) Catching Surface Details	AIC BIC CAIC $M\chi^2$ MADcor MADQ3 SRMSR
Ravand & Robitzsch (2018)	General English test of National University Exam	Reading (J = 20)	G-DINA DINA DINO ACDM R-RUM	21,642	• Adapted from Ravand (2016)		AIC BIC $M\chi^2$ MADcor MADRES SRMSR

APPENDIX B

SAMPLE ITEMS

Part D: The Life Cycle of the Butterfly

Butterflies are flying insects that undergo a complete physical change called a metamorphosis.



Metamorphosis

© 2020 Board of Regents of the University of Wisconsin System

10

A butterfly goes through many stages during its life. First, a larva comes out of an egg and starts to grow bigger. Next, during the pupa stage, the larva attaches itself to a twig and forms a hard outer shell. This shell is called a chrysalis. After many days, the butterfly emerges from its chrysalis as an adult.

According to the passage, which words show the correct order of a butterfly's metamorphosis?

○ Larva → Pupa → Egg → Adult

○ Egg → Larva → Pupa → Adult

○ Pupa → Larva → Egg → Adult

○ Egg → Pupa → Larva → Adult

11

Look at the diagram. A larva has simple eyes, small antennae, and many legs. The larva also has sharp teeth and powerful jaws for chewing food. After it forms a chrysalis, a total transformation occurs. This transformation results in a fully formed butterfly with compound eyes, long antennae, and only six legs. It also develops a proboscis, or flexible tube, for feeding.

What is one physical change that occurs during metamorphosis?

○ The number of eyes decreases.

○ The number of teeth increases.

○ The number of legs decreases.

○ The number of antennae increases.

12

When butterflies are young, their nutritional needs and eating habits are different from when they are adults. Larvae, or caterpillars, are ravenous eaters and quickly consume large quantities of food so they can grow rapidly and prepare for their metamorphosis. For example, as soon as a caterpillar is born, it eats its own eggshell and then begins to consume leaves on the plants around it. Caterpillars have specialized digestion systems that quickly process large amounts of food.

In adulthood, butterflies require less food than they did during the larva stage. Adult butterflies feed on liquids such as flower nectar to get the right nutrition so they can stay healthy.

Why are the eating habits of caterpillars different from the eating habits of adult butterflies?

○ Because caterpillars need fewer nutrients

○ Because caterpillars are usually very healthy

○ Because caterpillars need a lot of food to help them grow

○ Because caterpillars digest nectar slowly to help them change physically

APPENDIX C

EXAMPLE L2 READING SKILLS, PROCESSES, AND STRATEGIES

Skill, Process, Strategy Labels and Examples	Sources
Word Meaning/Vocabulary Knowledge Examples: remembering words, understanding less frequent words, matching words and definitions, guessing/inferencing, deducing from context, identifying meaning of words, finding equivalent words, interpreting words	Alderson & Lukami (1989), Buck et al. (1998), Davis (1968), DELNA, DELTA, Grabe (1991), Kim (2015), Lumley (1993), Munby (1978), Phakiti, (2003), Rost (1993)
Grammar/Syntactical/Syntax/Structure Knowledge Examples: understanding complex structures, processing negation	Buck et al. (1998), Grabe (1991), Lumley (1993), Rost (1993), Phakiti, (2003)
Understanding Explicit Information, Identifying/Locating Information	Alderson & Lukami (1989), Davis (1968), Lumley (1993), Munby (1978), Phakiti, (2003)
Understanding Implicit Information	Munby (1978), Phakiti, (2003)
Drawing Inferences/Conclusions/Reasoning, Making Inferences, Inferencing, Understanding Inferred Meaning, Interpreting by Going Outside the Text, Using Background knowledge	Alderson & Lukami (1989), Davis (1968), DIALANG, DELNA, DELTA, Grabe (1991), Kim (2015), Lumley, 1993), Munby (1978), Phakiti, (2003), Rost (1993)
Identifying/ Understanding Main Ideas/Points, Summarizing Main Idea/Topic	DIALANG, Kim (2015), Munby (1978), Phakiti, (2003),
Distinguishing Main Points from Details	Davis (1968), DELNA, Lumley (1993), Munby (1978), Phakiti, (2003)
Identifying/Finding/Understanding Specific Information/Details, Understanding Supporting Details Example: extracting relevant information	DIALANG, DELNA, DELTA, Kim (2015), Munby (1978), Phakiti (2003), Urquhart & Weir (1998)
Connecting Ideas/Important Details/Sentences, Concepts, Ideas, Reading to Integrate	Buck et al. (1998), Davis (1968), Kim (2015)
Skimming	Kim (2015), Munby (1978)
Scanning for Specific Information	Kim (2015), Munby (1978), Urquhart & Weir (1998)
Distinguishing Facts from Opinion, Interpreting/Recognizing Attitude/Purpose	Davis (1968), DELNA, DELTA, Phakiti (2003)
Summarizing, Synthesizing/Evaluating, Surveying, Understanding Gist Example: extracting salient details	Alderson & Lukami (1989), Grabe (1991), Lumley (1993), Munby (1978), Phakiti (2003)
Identifying General Information	Phakiti (2003)
Paraphrasing	Kim (2015)
Pragmatic Meaning Example: contextual, sociolinguistic, sociocultural	Kim (2015)
Semantic Meaning Example: meaning of words, paragraphs, sentences	Kim (2015)
Locating Causes and Effects, Sequences, Contrasts/ Analyzing Elements in a Process	DELNA, Lumley (1993)

Organizing Information in Other Ways Example: map, diagram, chart	DELNA
Understanding Concepts, Discourse Structure/Markers/Cohesion Example: understanding cause, results, organization of the text	Grabe (1991), Munby (1978)
Interpretation Example: significance of information, relationships	Alderson & Lukami (1989)

Note: DIALANG, The Diagnostic English Language Needs Assessment (DELNA), The Diagnostic English Language Tracking Assessment (DELTA) are diagnostic language tests used at the higher education level. The skills for these tests were obtained from Harding et al. (2015). Munby's list is obtained from Alderson & Lukami (1989) and Alderson (2000). This list is not a comprehensive list. The intent was to familiarize experts with reading skills, labels and examples.

APPENDIX D

ATTRIBUTES AND CODING EXPLANATIONS

Attribute	Explanation	Additional Coding Considerations
Vocabulary (VOC)	Understanding the key words/phrases in the text dependent or independent of the context. The attribute also entails recognition and knowledge of synonyms, antonyms, and the association between similar words in the text and answer choices (i.e., paraphrase).	Note that vocabulary knowledge is necessary for correct response respond but consider students' grade level when selecting this attribute (e.g., if it is a very easy word, given students' grade level, you may not want to code the item for this attribute)
Cultural Conceptual References (CUL)	Understanding the idea of concept. The attribute is closely related with the vocabulary attribute, but it requires knowledge of the "extended meanings". Just knowing the meaning of the words might not be adequate and it might require understanding at a conceptual level. Some concepts might be rooted in the culture and may be unfamiliar to a student from a different culture (e.g., community service, leadership training). (Bachman, 1990, p.97)	Consider this attribute when the questions require understanding specific concepts some of which might be culture specific. Although students might know individual words/phrases, for a complete understanding, it is necessary to understand nuances.
Grammar (GRM)	Understanding and processing complex sentences (e.g., relative clauses), and compound clauses including numerous grammatical and cohesive devices such as conjunctions. The attribute involves recognizing pronoun references.	Consider this attribute when extracting meaning from structure is necessary to respond correctly. Some sentences may be compound or complex that students need to process, or require them to recognize pronoun references, conjunctions etc. to understand the meaning and correctly respond.
Explicit Information and Details (EXP)	Deriving and comprehending explicit important information and details from the text. The attribute involves scanning the text and finding the details, and/or matching (i.e., answer choice and sentence in the text).	Note that specific information or details necessary for correct response is transparent in the text, and sometimes verbatim.

Inference (INF)	Comprehending information by making inferences. The information is implicit or overtly stated in the text. For example, the attribute requires connecting information in the text with an example situation.	Note that details necessary for correct answer is less transparent but overtly stated in the text. Student is required to make some sort of inference but given the level of the students, it might be low level inference in some cases.
Summary and Synthesis (SUM)	Connecting and integrating information in adjacent sentences or parts of the text (e.g., paragraphs, charts). The attribute entails summarizing, understanding the gist of the paragraphs, or interpreting rhetorical relations (e.g., problem-solution).	Note that information that is necessary for correct answer is stated in different sentences, paragraphs or cells of a chart. It might also require summary/synthesis of the information.
Sequences and Processes (SEQ)	Understanding sequential language, steps or order in a process or cycle. Information presented includes description of a sequence/steps and/or sequential language (e.g., first, second, eventually) that needs to be processed for a correct response.	Information includes description of a sequence/steps and/or sequential language (e.g., first, second, eventually) that students need to process and understand for correct response.

APPENDIX E

AN EXAMPLE EXPERT-DEFINED Q-MATRIX

	VOC	GRM	EXP	INF	SUM	SEQ	CUL	Notes/Rationale	CONF
Q1	0	0	1	0	0	0	0	Definitely extracting info, but I was also torn about including vocab and even inferencing. I ultimately decided that the first paragraph did not have any challenging crucial vocab. I kind of think an inference is involved, but no one else thought so, so I'll acquiesce.	3
Q2	0	0	1	1	0	0	0	Student has to extract definition of X and see if each example matches the definition (inference)	4
Q3	1	0	1	0	0	0	0	Higher level vocab in the final paragraph, of which the key is essentially a paraphrase, so extracting info	5
Q4	0	1	1	0	0	0	0	The answer is in the text in the form "After A, B", so students must know that this grammar structure is implying a causal relationship between A and B	5
Q5	1	0	0	0	1	0	1	Must know or infer the meaning of X and the key is a summary of a longer sentence in the text.	4
Q6	1	0	1	0	0	0	1	Students might need to understand the meaning of X, which could be somewhat culturally-specific	3
Q7	0	0	1	0	0	0	0	Need to extract definition of X, might also need to synthesize info from the visual	3
Q8	0	0	1	1	0	1	0	Students have to extract a rule about a process and apply it to a new situation	3
Q9	0	1	1	0	0	1	0	Must extract what the string represents, which requires parsing some grammar and understanding the process of X	4
Q10	0	0	1	0	0	1	0	I'm very conflicted about this one. Seems like examinee might need to know/extract the science vocab, which is related to a sequence, and pick the answer that summarizes the sequence. But then the item facility is quite high, so it seems unlikely that 3~4 attributes are all necessary.	2
Q11	1	0	0	1	0	0	0	Although the more complex vocab does not seem strictly necessary to answer correctly, the science vocab could create confusion/distraction (most-picked distractor is about antennae). Also have to interpret meanings related to increase/decrease and infer based on descriptions of X	4

Q12	1	0	1	0	0	0	0	Some advanced vocab but key is fairly similar to answer in the text	5
Q13	1	1	1	0	0	0	0	Vocab and syntax (negation, demonstrative adj). I am not as sure about extracting info, but it seems like they might need to search the text for an indication of uniqueness	3
Q14	0	0	0	0	1	0	0	I think just synthesis/summarizing would work, because the entire stimulus is about X, so the task is not so much selecting information from the text as deciding which option is an accurate summary.	4
Q15	1	1	0	1	0	0	0	Conditional and relative clauses, and inference that if X happens. The low p-value seems consistent with multiple attributes	5
Q16	0	0	1	0	0	1	0	Have to extract information from a sequence	4
Q17	0	0	1	0	1	1	0	I’m not sure about extracting information. The question asks for summarization of a process, so the other two attributes seem more straightforward. Also, the p-value is extremely low, suggesting that multiple attributes might be needed	4
Q18	0	0	0	1	1	0	0	Have to synthesize information from two parts of the table in the text and infer the implications for an example scenario	5
Q19	0	0	1	0	0	0	0	The key is copied almost verbatim from the text	5
Q20	0	1	1	0	0	1	0	Need to understand the process of X and extract a specific step. Also have to interpret the meaning of X from syntax. The low p-value is consistent with multiple attributes required.	4

Note. Only 20 items are presented. Conf denotes confidence with the selected attributes.

APPENDIX F

ADDITIONAL TABLES AND FIGURES

Figure 1. Distribution of Raw Scores

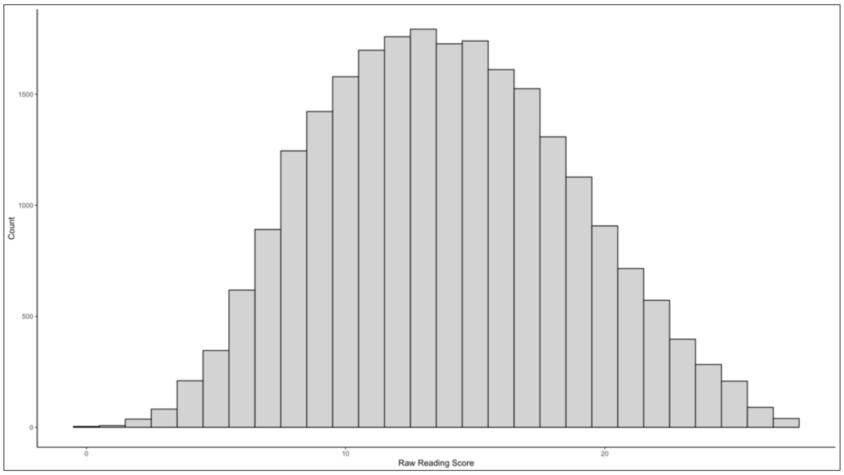


Figure 2. Distribution of Confidence Ratings across Content Areas

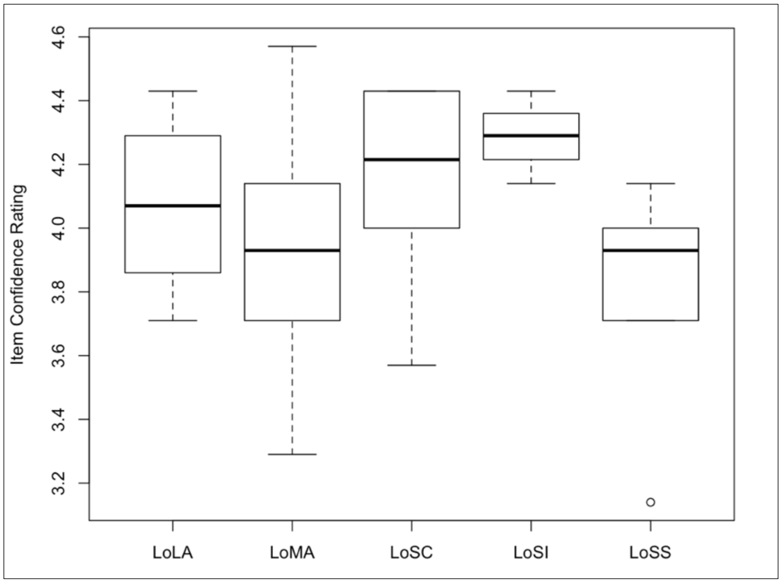


Figure 3. Mesa Plots when ∈= .85

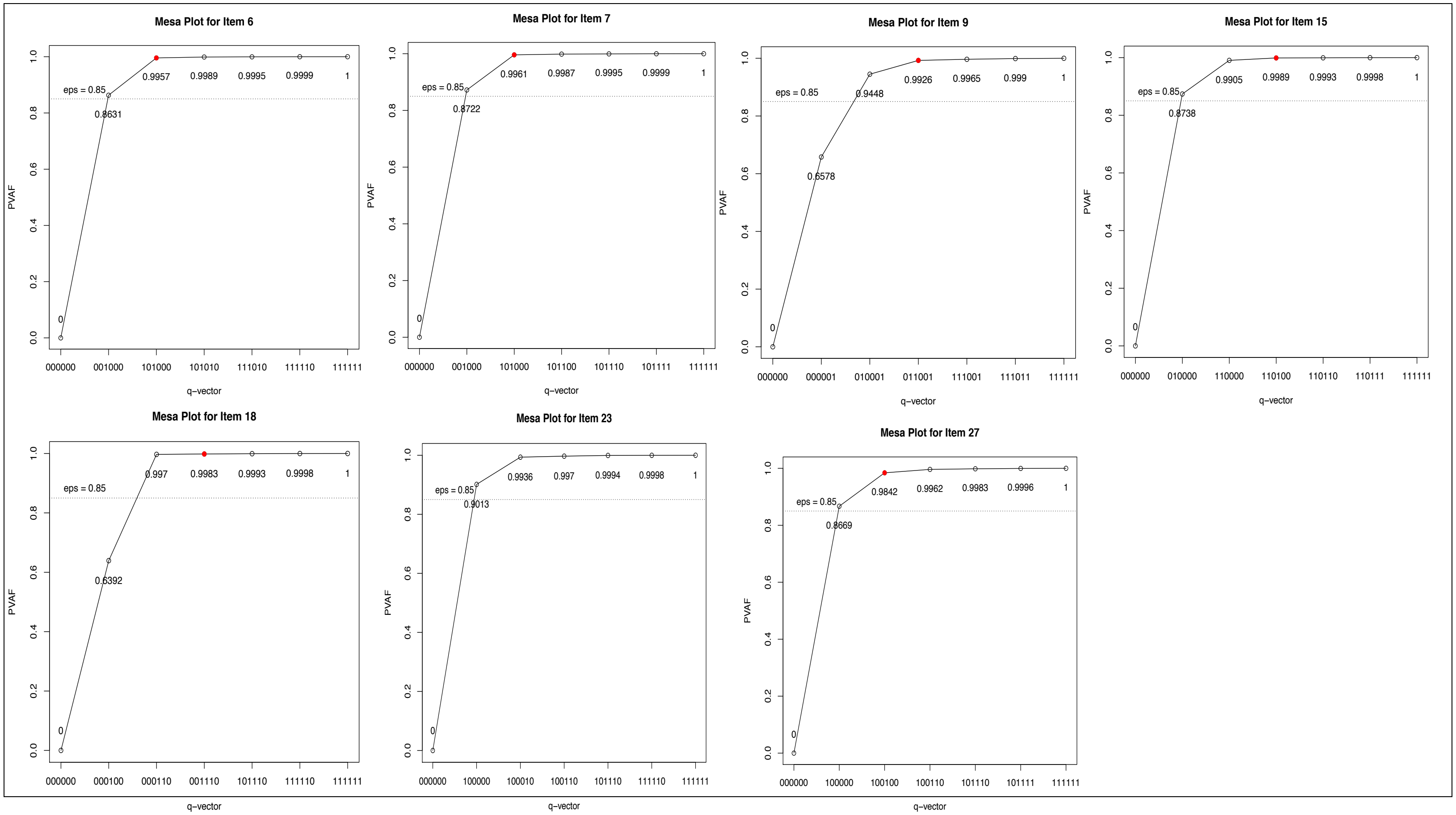


Table 1. Item and Distractor Statistics

Item	p	pbis	option		Item	p	pbis	option		Item	p	p.bis	option	
#				prop.	#				prop.	#				prop.
1	0.66	0.27	1	0.17	10	0.70	0.19	1	0.13	19	0.68	0.38	1	0.09
			2	0.66				**2**	0.70				2	0.14
			3	0.07				3	0.01				**3**	0.68
			4	0.09				4	0.15				4	0.07
2	0.48	0.24	1	0.30	11	0.65	0.33	1	0.06	20	0.49	0.26	1	0.14
			2	0.17				2	0.11				**2**	0.49
			3	0.48				**3**	0.65				3	0.24
			4	0.05				4	0.18				4	0.10
3	0.64	0.25	1	0.20	12	0.75	0.35	1	0.06	21	0.36	0.21	1	0.17
			2	0.06				2	0.08				2	0.28
			3	0.09				**3**	0.75				**3**	0.36
			4	0.64				4	0.10				4	0.16
4	0.72	0.41	1	0.11	13	0.63	0.43	1	0.14	22	0.55	0.30	**1**	0.55
			2	0.72				2	0.10				2	0.22
			3	0.10				3	0.12				3	0.13
			4	0.06				**4**	0.63				4	0.06
5	0.63	0.38	**1**	0.63	14	0.53	0.38	1	0.11	23	0.32	0.23	**1**	0.32
			2	0.15				2	0.17				2	0.28
			3	0.09				**3**	0.53				3	0.12
			4	0.14				4	0.17				4	0.24
6	0.51	0.35	1	0.13	15	0.35	0.20	1	0.12	24	0.39	0.26	1	0.13
			2	0.26				**2**	0.35				2	0.18
			3	0.51				3	0.37				**3**	0.39
			4	0.10				4	0.15				4	0.26
7	0.68	0.32	**1**	0.66	16	0.71	0.38	1	0.15	25	0.56	0.38	1	0.12
			2	0.20				**2**	0.71				**2**	0.56
			3	0.10				3	0.05				3	0.18
			4	0.01				4	0.07				4	0.10
8	0.38	0.28	1	0.23	17	0.28	0.27	1	0.24	26	0.30	0.23	**1**	0.30
			2	0.17				2	0.28				2	0.18
			3	0.38				3	0.18				3	0.33
			4	0.22				**4**	0.28				4	0.15
9	0.34	0.11	1	0.36	18	0.27	0.29	1	0.38	27	0.31	0.15	**1**	0.31
			2	0.34				2	0.16				2	0.26
			3	0.16				**3**	**0.27**				3	0.25
			4	0.14				4	0.16				4	0.24

Note. p = p-value, pbis = point biserial, option prop. = proportion of students choosing an option. Bold denotes the key.

Table 2. Agreement Rate among Test Developer SME Group

	Test Developer Group		
	Fleiss Kappa	z statistic	p-value
VOC	0.306	3.889	0.000
GRM	0.010	0.127	0.899
EXP	0.195	2.479	0.013
INF	0.286	3.637	0.000
SUM	0.247	3.141	0.02
SEQ	0.576	7.337	0.000

Table 3. Fit Indices for Standards-based Q-matrix (Calibration Sample)

	Npars	-2LL	AIC	BIC	CAIC
LCDM	145	**-195289.1**	390868.157	391939.742	392084.742
RRUM	118	-195308.7	390853.38	391725.429	391843.429
CRUM	118	- 195305.2	**390846.367**	**391718.416**	**391836.416**
DINO	91	-196501.6	393185.224	393857.736	393948.736
DINA	91	-196474.4	393130.767	393803.279	393894.279
HO-DINA	70	-196789.8	393719.519	394236.836	394306.836

Table 3 Cont. Fit Indices for Standards-based Q-matrix (Calibration Sample)

	Mχ^2	p	MADcor	SRMSR	MADres	MADQ3
LCDM	110.063	0	0.012	0.016	0.279	0.024
RRUM	56.3455	0	0.012	0.016	0.271	0.023
CRUM	54.0475	0	0.012	0.016	0.277	0.024
DINO	80.0818	0	0.021	0.027	0.476	0.019
DINA	73.0098	0	0.021	0.027	0.471	0.018
HO-DINA	211.5088	0	0.026	0.032	0.589	0.018

Table 4. Wald's Test for Item Level Model Selection

Item	Model	Wald Statistic	p-value
2	R-RUM	1.333	0.248
3*	C-RUM	6.200	0.013
5	C-RUM	2.346	0.126
6	R-RUM	1.107	0.293
7	C-RUM	0.048	0.827
8*	R-RUM	15.82	0.000
9	R-RUM	2.82	0.588
11*	C-RUM	3.884	0.049
12	C-RUM	1.592	0.207
13	C-RUM	0.894	0.344
15*	R-RUM	9.941	0.041
16	C-RUM	3.666	0.056
17*	R-RUM	10.419	0.001
18*	R-RUM	31.077	0.000
20	C-RUM	0.234	0.629
21*	R-RUM	13.213	0.000
22	C-RUM	0.078	0.78
23*	R-RUM	4.112	0.043
25	C-RUM	1.140	0.286
27	DINA	0.227	0.893

Note. Items 1, 4, 10, 14, 19, 24, 26 are simple items and excluded from the table. * denotes a non-significant result meaning the LCDM would be a statistically better choice.

Table 5. Multiple Regression Results

	Estimate	SE	t value	p-value
Intercept	-0.78815	0.00531	-148.442	2.00E-16
VOC	0.23036	0.01985	11.606	2.00E-16
GRM	0.17767	0.01332	13.343	2.00E-16
EXP	0.57303	0.01625	35.265	2.00E-16
INF	0.36523	0.02338	15.624	2.00E-16
SUM	0.1605	0.01741	9.218	2.00E-16
SEQ	0.64571	0.02252	28.671	2.00E-16

Note. Dependent variable= θ, R^2 = 0. 881, F-statistic = 1.47E+00

Table 6. One-way ANOVA Results

		df	Sum Sq	Mean Sq	F	p
VOC	PL	2	959.1	479.5	10342	< 0.001
	Residuals	11968	554.9	0.0		
GRM	PL	2	470.2	235.11	3640	< 0.001
	Residuals	11968	773.1	0.06		
EXP	PL	2	1326.6	663.3	9814	< 0.001
	Residuals	11968	808.9	0.1		
INF	PL	2	885.7	442.8	11886	< 0.001
	Residuals	11968	445.9	0.0		
SUM	PL	2	1125.1	562.5	9041	< 0.001
	Residuals	11968	744.7	0.1		
SEQ	PL	2	758.0	379	9721	
	Residuals	11968	466.6	0.0		< 0.001