Title

Reporting Fine-grained Feedback from a Summative Language Proficiency Assessment Using Diagnostic Classification Modeling (DCM): A Feasibility Study

Author

Setoguchi, Eric

Publication Date

2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Reporting Fine-grained Feedback from a Summative Language Proficiency Assessment Using Diagnostic Classification Modeling (DCM): A Feasibility Study

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Education

by

Eric Hiroyuki Setoguchi

2019

ABSTRACT OF THE DISSERATION

Reporting Fine-grained Feedback from a Summative Language Proficiency Assessment Using Diagnostic Classification Modeling (DCM): A Feasibility Study

by

Eric Hiroyuki Setoguchi

Doctor of Philosophy in Education

University of California, Los Angeles, 2019

Professor Noreen Webb, Chair

Assisting states to meet federal requirements related to EL students are summative English language proficiency (ELP) assessments, administered annually at the end of each school year. These assessments are increasingly expected to play a role in providing information about students' language proficiency to educators and other local stakeholders for informing instruction. However, this application raises several key questions. As large-scale instruments, do the summative assessments provide information that is both meaningful and useful to local educators and school administrators? Through what process are the summative assessments expected to "work together" with other assessments, including classroom assessments? What evidence exists to demonstrate how capable the current summative assessments are at fulfilling this role? To investigate these questions, this study sets out to explore a possible approach to improving the feedback provided by the reading subsection of an ELP summative assessment by retrofitting a diagnostic classification model (DCM),

a class of scoring model used to understand and report test data specifically in ways that can be interpreted and used by stakeholders to make decisions about student instructional needs, namely classification based judgments about what abilities and knowledge students do and do not have. The findings suggest that EL educators are struggling to see the current score reports provided by ELP summative assessments as interpretable or instructionally useful, and many are receptive to alternative approaches to getting feedback about their students, particularly, feedback in the form of fine-grained, skill-based information about what students strengths and weaknesses are. In addition, creatively applied DCMs could be used to generate finer-grain feedback about students' reading abilities than is currently being done, and the feedback provided is impressively reliable given that it was derived from a single test administration. However, the findings also make it clear that retrofitting of DCM models is not an uncomplicated procedure. There are a number of cautionary flags raised in this study in regard to this class of models and these particular assessments that requires further attention.

The dissertation of Eric Hiroyuki Setoguchi is approved.

Alison Bailey

Mark Hansen

Matthew Maddison

Noreen Webb, Committee Chair

University of California, Los Angeles

2019

Acknowledgements

It has been a remarkable journey to complete my graduate study at UCLA. At the top of a long list of people for me to thank is Dr. Noreen Webb, the best advisor and committee co-chair I could have ever hoped for. What an absolute privilege it has been to learn from Reenie's expertise, grow under her mentorship, and be inspired by her wisdom and patience. It is a special and rare gift for a student to get an advisor like Reenie, and to me she embodies everything that I got into this career to do and everything that I aspire to become. There is also no way this dissertation could happen if not for Dr. Matthew Maddison. I was fortunate to take his course on Diagnostic Classification Models, and it went on to become the inspiration for this dissertation. Matthew is one of the most tremendous and skilled teachers I've ever had, and his guidance and support have been priceless. Dr. Alison Bailey is one of the first people I contacted when applying to UCLA, and getting the opportunity to learn from her was one of the main reasons I joined the program. Alison has been an amazing resource for matters related to language learning and assessment, and I'm truly thankful for her insights and support throughout my time as a student. In my short time in the psychometrics field, I consider myself extraordinarily lucky to have had some truly fantastic mentors. Mentors like Dr. Mark Hansen. Mark welcomed me to work with him despite my being entirely unqualified to do so. The talent, passion, and ethics with which Mark approaches his work is remarkable and awe inspiring. I owe him a great debt for not only that opportunity, but the profound impact that he has had on my decision to make this work a part of my career. Others deserve my thanks as well: Dr. Felipe Martinez, Dr. Mike Seltzer, Dr. Mike Rose, my fellow students, and the amazing staff at UCLA. Of course, my degree would not have been possible without the support and understanding of my parents, family, and those close to me. To my wife Kana, the light of my life, I dedicate this dissertation to you and the months you worked supporting us without a single word of complaint. I think it's about time I finally stopped freeloading and got a job so you can take a bit of a break.

Table of Contents

List of Tables

List of Figures

Biographical Sketch

Eric Setoguchi obtained his Bachelor's degree in Biology from the University of California, San Diego in 2003. After a few years studying abroad in Japan, he obtained a Master of Arts degree in Second Language Studies from the University of Hawaii at Manoa in 2008, specializing in Language Assessment, Measurement, and Program Evaluation. Following a stint teaching English in Japan, he returned to California in 2014 to enter the doctorate program in Social Research Methodology in the Department of Education at the University of California, Los Angeles.

1. The Expanding Role of Summative English Language Proficiency Assessments

Diligent attention must be given to the instruction, assessment, and monitoring of the English proficiency of English learner (EL) students in the United States primary and secondary school system. Based on annual statistics collected by the Department of Education, EL students comprise approximately 10% of the K-12 population, equal to roughly 4.85 million students (Snyder, de Brey, & Dillow, 2019). Federal legislation requires that states address important educational needs related to the academic progress of these students in the current accountability-based education system, including the collection of evidence that demonstrates EL students' progress towards attaining English proficiency. This evidence in turn is used by federal, state, and school stakeholders in various ways to support EL students in their progress towards college and career readiness. Doing so, it is hoped, ensures that EL students do not fall behind their non-EL peers, and are receiving equal access to grade appropriate instruction in the academic content areas.

Assisting states to meet federal requirements related to EL students are summative English language proficiency (ELP) assessments, administered annually at the end of each school year. States have several options to choose from when selecting an ELP assessment to use, but the ways in which states use their chosen ELP assessment are largely consistent (Wolf et al., 2008). Firstly, the ELP assessment serves as a way for states to concretely operationalize a process for identifying and reclassifying EL students based on the definition provided in the Every Student Succeeds Act (ESSA) of 2015, the nation's education law designed to protect equal opportunity for K-12 students. According to ESSA:

> "[An EL student is] an individual whose difficulties in speaking, reading, writing, or understanding the English language may be sufficient to deny the individual – the ability to meet the challenging State academic standards; the ability to successfully achieve in

classrooms where the language of instruction is English; or the opportunity to participate fully in society." (ESSA Title VIII, 2015)

These classifications decisions are critical to get right. Scores on the ELP assessment are submitted by each state to the federal government under an accountability requirement in ESSA (ESSA Title I & Title III, 2015). In this way, ELP assessments make up not only a substantial proportion of the standardized testing responsibilities of each state, but importantly, a key component in decisions pertaining to the education of EL students as well. Subsequently much of the research on the development and improvement of the assessments have focused on these test uses in particular. However, summative ELP assessments are increasingly being expected to play a different and critical role in the education process, that of providing information about students' language proficiency to educators and for informing classroom instruction. The current homepages of each of the three most prominent ELP assessments, the English Language Proficiency Assessment for the 21st Century (ELPA21), ACCESS for ELLs, and the English Language Proficiency Assessments for California (ELPAC), refer to informing educators or classroom instruction as an intended use of the assessment (Table 1). Another place this trend can be observed is in documents and public presentations made by ELP assessment developers that have recently starting referring to the concept of a "balanced assessment system". The origin of this term appears to be taken from a brief mention in a subsection of the ESSA report related to the assessment activities to be carried out by the states:

"Developing or improving balanced assessment systems that include summative, interim, and formative assessments, including supporting local educational agencies in developing or improving such assessments." (ESSA Title I, 2015)

While further details regarding what these so called "balanced assessment systems" are meant to entail are not addressed in ESSA, others have since attempted to make a more detailed

description of the concept (The Los Angeles Unified School District, 2016; Educational Testing Service, 2018; The Center on Standards and Assessment Implementation, 2018).

Table 1.

References to Informing Educators or Classroom Instruction on ELP Assessment Homepages

ELPA21	"Our assessments are designed to give educators the information they need to help students unlock their potential with English language proficiency." Source: https://elpa21.org/assessment-system/assessments/
ACCESS for ELLs	"ACCESS for ELLs scores have many potential uses, from determining student placement to guiding the creation of new curricula. Test scores work best as a way to aid decision-making, in cases such as informing classroom instruction and assessment." Source: https://wida.wisc.edu/assess/access/scores-reports
ELPAC	"The [ELPAC Academy] focused on the ELPAC's implications for classroom instruction and student learning and how educators can use the ELPAC task types to improve learning." Source: https://www.cde.ca.gov/ta/tg/ep/elpacacademy1718.asp

There are a few notable similarities that appear standard across these descriptions. First, balanced assessment systems feature different categories of assessments, namely summative, interim, and formative assessments, that function in tandem to prepare K-12 students to be college ready. Second, the various assessments achieve this goal by providing "multi-tiered" information to a diverse population of stakeholders, including those at the, federal, state, and school level so that they can collaboratively improve teaching and learning practices and enhance educational collaboration system-wide.

Noteworthy is the enthusiasm with which summative ELP assessment developers have, via websites and promotional materials, committed their assessments to playing a role in these balanced assessment systems. However, such statements have been vague in regard to several key questions, particularly: (a) As large-scale instruments, do the summative assessments provide information that is both meaningful and useful to local educators and school administrators?, (b) Through what

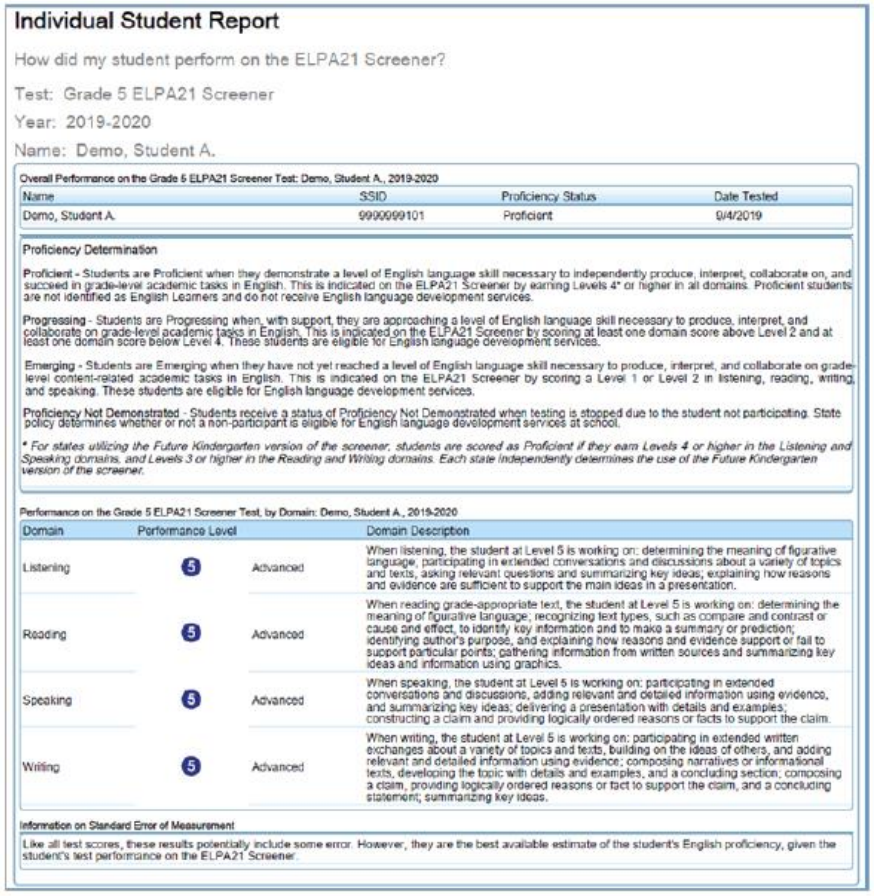
process are the summative assessments expected to "work together" with other assessments, including classroom assessments?, and (c) What evidence exists to demonstrate how capable the current summative assessments are at fulfilling this role? Indications certainly exist to suggest that better answers to these questions are sorely needed. For example, the ELPA21 currently provides information to educators in the form of an individual student report, or ISR (Figure 1). Included in the ISR is the student's proficiency classification, their ability level on each of the subdomain sections of the test (Listening, Reading, Speaking, and Writing), and short descriptions of each of the ability levels. While the ISR certainly provides educators with a variety different kinds of information, it is unclear whether the information would qualify as the type of information educators would consider interpretable or useful. For instance, teachers report wanting diagnostic information about their students (Huff & Goodman, 2007). Diagnostic information would include descriptions of the specific knowledge and abilities that students have demonstrated, and which areas they need improvement in, and it is likely that the large-grain, non-diagnostically derived scores and ability levels generated in reports like the ISR would not meet such a need.

To investigate the three questions raised in the preceding paragraph, the current study sets out to explore a possible approach to improving the score report shown in Figure 1 by using diagnostic classification models (DCMs; Rupp, Templin, Henson, 2010), a class of scoring models used to understand and report test data specifically in ways that can be interpreted and used by stakeholders to make decisions about students, namely classification based judgements about what abilities and knowledge students do and do not have. However, there is much that is unknown about the appropriateness of DCMs in the specific context of summative ELP assessments. The current study takes three pressing topics that are in need of further investigation as the basis of its research agenda:

- EL Educators' opinions about the interpretability and usefulness of the feedback that DCMs are designed to provide;
- The capacity for DCMs to provide feedback that is specific to skills and abilities relevant to EL curriculums;
- Whether DCMs fit to the summative ELP assessments can provide feedback about EL students that is reliable enough to be useful.

Figure 1.

An Example of a Summative ELP Assessment Score Report



Individual Student Report

How did my student perform on the ELPA21 Screener?

Test: Grade 5 ELPA21 Screener

Year: 2019-2020

Name: Demo, Student A.

Overall Performance on the Grade 5 ELPA21 Screener Test: Demo, Student A., 2019-2020

Name	SSID	Proficiency Status	Date Tested
Demo, Student A.	9999999101	Proficient	9/4/2019

Proficiency Determination

Proficient - Students are Proficient when they demonstrate a level of English language skill necessary to independently produce, interpret, collaborate on, and succeed in grade-level academic tasks in English. This is indicated on the ELPA21 Screener by earning Levels 4* or higher in all domains. Proficient students are not identified as English Learners and do not receive English language development services.

Progressing - Students are Progressing when, with support, they are approaching a level of English language skill necessary to produce, interpret, and collaborate on grade-level academic tasks in English. This is indicated on the ELPA21 Screener by scoring at least one domain score above Level 2 and at least one domain score below Level 4. These students are eligible for English language development services.

Emerging - Students are Emerging when they have not yet reached a level of English language skill necessary to produce, interpret, and collaborate on grade-level content-related academic tasks in English. This is indicated on the ELPA21 Screener by scoring a Level 1 or Level 2 in listening, reading, writing, and speaking. These students are eligible for English language development services.

Proficiency Not Demonstrated - Students receive a status of Proficiency Not Demonstrated when testing is stopped due to the student not participating. State policy determines whether or not a non-participant is eligible for English language development services at school.

** For states utilizing the Future Kindergarten version of the screener, students are scored as Proficient if they earn Levels 4 or higher in the Listening and Speaking domains, and Levels 3 or higher in the Reading and Writing domains. Each state independently determines the use of the Future Kindergarten version of the screener.*

Performance on the Grade 5 ELPA21 Screener Test, by Domain: Demo, Student A., 2019-2020

Domain	Performance Level		Domain Description
Listening	5	Advanced	When listening, the student at Level 5 is working on: determining the meaning of figurative language; participating in extended conversations and discussions about a variety of topics and texts, asking relevant questions and summarizing key ideas; explaining how reasons and evidence are sufficient to support the main ideas in a presentation.
Reading	5	Advanced	When reading grade-appropriate text, the student at Level 5 is working on: determining the meaning of figurative language; recognizing text types, such as compare and contrast or cause and effect, to identify key information and to make a summary or prediction; identifying author's purpose, and explaining how reasons and evidence support or fail to support particular points; gathering information from written sources and summarizing key ideas and information using graphics.
Speaking	5	Advanced	When speaking, the student at Level 5 is working on: participating in extended conversations and discussions, adding relevant and detailed information using evidence, and summarizing key ideas; delivering a presentation with details and examples; constructing a claim and providing logically ordered reasons or facts to support the claim.
Writing	5	Advanced	When writing, the student at Level 5 is working on: participating in extended written exchanges about a variety of topics and texts, building on the ideas of others, and adding relevant and detailed information using evidence; composing narratives or informational texts, developing the topic with details and examples, and a concluding section; composing a claim, providing logically ordered reasons or fact to support the claim, and a concluding statement; summarizing key ideas.

Information on Standard Error of Measurement

Like all test scores, these results potentially include some error. However, they are the best available estimate of the student's English proficiency, given the student's test performance on the ELPA21 Screener.

Before delving into the specifics of how DCMs could be applied to the context of ELP assessments to investigate these topics, it would be worthwhile to first step back and provide greater explanation of why the topics mentioned above are relevant and important, especially within the specific context of K-12 ELP assessment in the era of ESSA.

2. A Systematic Approach to Assessing English Learners in the K-12 Context

The push for balanced assessment systems in K-12 education brings an expectation that summative ELP assessments play a role in the classroom context. Taking a step back to clarify what is meant by the term "summative", traditional test development theory differentiates tests into separate archetypes depending on their intended users and uses. So called summative assessments, such as the large-scale standardized tests administered at the state level, are used to determine the level of success or competency attained by a student at the end of instruction. Summative assessments are seen as being situated in a separate category from the context of use of making instructional decisions in the classroom during the process of learning itself. The latter are more directly associated with so called formative assessments, such as locally-developed classroom tests of the sort traditionally made and used during daily instruction by educators and school staff. Importantly, these two terms, "summative" and "formative", are intended to describe the interpretations, decisions, and consequences that result from an assessment, rather than a quality of the assessment itself. For example, the summative ELP assessments are considered "summative" not so much because of the properties of the assessment itself, but because the primary intended purpose is for students' scores to be used for making end of instruction classifications about their proficiency. For this reason, some test scholars advocate for using the expanded terms "assessment for summative purposes" and "assessment for formative purposes" (Ussher & Earl, 2010). As the terms "summative" and "formative" are not static properties of assessments themselves, this opens

the door to the possibility that the summative ELP assessments, despite originally being intended for summative uses, conceivably could also be integrated with the context of what educators do in the classroom. Bailey & Carroll (2015) point to the many components currently involved in K-12 ELP assessment, and the pressing need to align and improve these parts into a coherent system. While integrating the summative tests with the classroom context could be one path towards achieving the kind of alignment and systematicity between different assessments that a balanced assessment system calls for, there is a need to clarify more about what this would entail.

To begin with, what is not being proposed in the current study is that summative ELP assessments should, or are capable of, being used formatively in the classroom in the same manner as local assessments designed explicitly for that purpose. For one, having a "one size fits all" mentality to any assessment instrument is unwise from a test design standpoint (Shepard, 1997), where test items are intentionally designed to meet specifications related to their intended use, and therefore cannot be assumed to be fully effective when applied to another context without strong evidence. Secondly, formative assessment practice involves doing more than simply taking information gleaned from an assessment and using it to inform instruction. Bailey and Heritage (2008) define seven essential elements, of which four would be impossible given the current form of the summative ELP assessment (spontaneity, student involvement, time interval, & locus of control.) Rather, information provided by a summative ELP assessment might very well have the potential to be useful indirectly or directly in support of an educators classroom instruction of EL students by means of various channels, but a clear distinction should be made between these sorts of activities and the "doing" of formative assessment.

Also important is that by "systematicity", an acknowledgement be made for the need to remedy, rather than worsen, the historical trend for large-scale assessments to hold greater influence and priority in the education system when compared to locally developed assessments, as has been the

case starting with NCLB, the policy forerunner of ESSA (Figlio & Loeb, 2011; Hamilton, Stecher & Yuan, 2012.) NCLB dictated that schools failing to meet annual yearly progress (AYP), an indicator of the increase in the proportion of students meeting proficiency on state administered tests, would be subject to "improvement, corrective action, and restructuring measures aimed at getting them back on course to meet State standards" (NCLB, 2002). Resnick (1980) described the feelings of suspicion around these new systematic changes:

> "The local school district is no longer the exclusive agent in the evaluation process…This trend toward a larger federal and state role in the funding and evaluation of public education is not likely to be reversed, and no effective resistance to the misuse of tests at the local level can ignore the state and federal mandate under which much of that testing proceeds." (Resnick, 1980, p.24).

Largely a result of a backlash against such policy efforts, school personnel and the public seem to be in agreement that students spend too much time taking tests (Kappa, 2016; Rentner et al., 2016). Then-president Obama commented frequently on a pressing need to address problems with over-testing in schools (Rothstein, 2009). Furthermore, a growing number of prominent educational scholars have dedicated all or part of their professional time to writing or speaking out on the negative impacts of accountability-based testing (Zhao, 2014; Robinson, 2010; Popham, 2003). Large-scale testing's "public image problem" hasn't been helped by a proliferation of research literature focusing on its shortcomings. Common targets for criticism include the narrowing of curriculums to those subjects appearing on the tests (Berliner, 2011), the adoption of a "factory model" approach to educating children (Au, 2011), and the questionable defunding of public education in favor of charter schools (Di Carlo, 2011). A number of high-profile incidents have also undermined the credibility of the tests in the eyes of the educational community and in the public, including widespread cheating scandals (Blinder, 2015), the failure of value added models to

effectively rate teachers (Newton et al., 2010), and the questionable ethicality of the public release of teacher effectiveness ratings (Felch, Song, & Smith, 2010). Especially concerning, more recently parents have also been resisting in the form of grassroots "opt-out" movements, a growing trend where parents simply refuse to allow their children to be tested (Layton, 2013). In this environment, it is critical that work to "systematize" ELP assessment be centrally concerned with elevating the voices and needs of educators and local stakeholders in the conversation about the assessment of EL students, and not about increasing the influence that summative ELP assessments have on their classroom instruction in a top-down manner. Indeed, the inclusion in ESSA of the term "balanced assessment system" has been cited as an explicit means to publicly "devolve power over education out of Washington and return it into the hands of states and local educators." (U.S Department of Education, 2018).

In spite of the concerns raised by the preceding paragraphs, there is a value to work that explores how to better integrate summative ELP assessments with the classroom context when done with the right perspective in mind. For instance, a practical argument based on the resources being spent on large-scale assessments could be made, given that the average amount of classroom time and money has been steadily increasing (Hart et al., 2015; Supovitz, 2009). In addition, the two largest state-level test developers alone, Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced Assessment Consortium (SBAC), received awards of $361 million to deliver their versions of a standardized assessment (Onosko, 2011). These figures not only represent a significant increase in time and expense spent on standardized tests over a relatively short timeframe, they also are higher when compared to similar trends seen in other nations. Given this investment, it stands to reason that getting more value out of these assessments beyond making a proficiency classification would help justify these costs.

An additional, and perhaps more impactful argument could be made in regard to a specific issue with the difficulty of interpreting summative assessment scores from the perspective of the language development of EL students. As suggested by Figure 1, the current aggregate scores which summative assessment report are unlikely to be seen as relevant and interpretable to educators and other local stakeholders, who are more concerned with finger-grain, diagnostic conceptualizations of language proficiency. As a consequence, these stakeholders are often given little choice but to interpret what they can from the aggregate scores into something meaningful about the specific things that their students can and cannot do. This equating of large- to small-grain information is problematic, given that the skills measured by the summative tests are very likely to be underrepresenting the actual collection of skills and processes involved with the use of English in the academic classroom (Bailey & Duran, 2019.) In other words, because the current assessment reports scores only in terms of a domain (ex. a "reading" score) or an overall proficiency classification, educators and other stakeholders have a reduced means to make use of this information to understand their students' needs or to participate in the conversation around the meaningfulness and accuracy of their students' performance on the assessment. In part to address this problem, assessment items were based on ELP and content area standards, which are designed to be closer in nature to the aspects of language that are relevant to educators. However, this alone is insufficient, as student scores are ultimately not reported with consideration for these standards (and furthermore the relationship between standards and specific assessment items not even provided.)

In light of the issues raised in this chapter, the current study takes a perspective where "systematizing" summative ELP assessments should involve an improvement to the way that scores are reported to include standards-based, or otherwise finer-grain information about students' language proficiency than currently is being achieved. The intent of doing so is to better integrate the summative assessment with other ELP assessments, particularly those based in the classroom,

through using a common framework for talking about student performance across various assessments. By having this framework come from the skills-based perspective of language learning in the classroom, the hope is that educators might be granted the capacity to collaborate in the summative assessment process, and the summative process in turn can more meaningfully support educators needs in the classroom. Summative assessments, and all assessments in the ELP system in fact, can contribute something more impactful than simply filling the role of a single test working in isolation. They can also act as sites of collaboration for diverse groups of users, bringing people together across educational disciplines and professional roles to solve system wide problems and bring about large-scale improvements. But to do so assessment feedback must take into consideration the perspectives of different user groups. This should be a key factor in the validity of such assessments, as Chapter 3 will discuss.

3. A Validity Argument for Integrating Summative ELP Assessments with the Classroom Context

The topic of validity holds a place of particular importance to those in the field of language assessment (Chapelle, 2012). *The Standards for Educational and Psychological Testing* defines validity as "the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (AERA, APA, NCME, 2014). Therefore, validity constitutes an appropriate platform on which to frame an investigation into the appropriateness of using summative ELP assessments for the purpose of supporting classroom instruction. Generally speaking, two sources in particular have come to serve as references of authority so far as educational test validity is concerned. The first is the aforementioned *Standards for Educational and Psychological Testing*, hereafter referred to as the *Standards*, a joint publication by the American Educational Research Association, American Psychological Association, and the National Council of Measurement in Education. The second is

the edited collection *Educational Measurement* (Yen, Fitzpatrick, & Brennan, 2006), which always includes a chapter devoted to the topic of validity. Given that the current study employs the validity argument approach discussed in the latter source, it will be worthwhile to spend time discussing it here.

3.1 Interpretive and Validity Arguments as a Framework for Validating Test Use

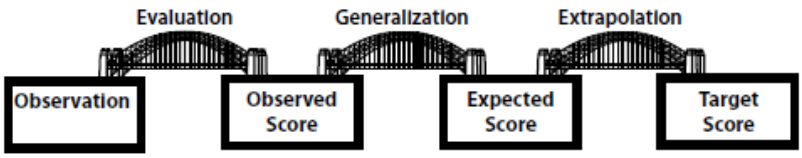
Kane's (2006) chapter on validation in the latest edition of *Educational Measurement*, perhaps comes the closest to describing, with a sufficient level of detail necessary for replication, what might be considered a consensus paradigm for approaching validity in any educational testing context. While dissent around the conceptualization of validity in the field remains a constant issue (Hammond & Moss, 2016), the chapter is popular (it has been cited 1678 times as of the writing of this report), and conforms to the description of validity currently laid out in the *Standards*. A thorough discussion of the work leading up to Kane's seminal chapter can be found in Chapelle, Enright, & Jamieson (2011). To summarize, building upon the ideas of some of the most prominent scholars working in assessment (Cronbach, 1988; Messick, 1989; Bachman & Palmer, 1996), Kane unified several different ideas concerning how validity should be evaluated into a single approach based on a need for carrying out of two complimentary processes: the construction of two arguments, an *interpretative argument* and a *validity argument*.

Interpretative arguments were developed to serve as conceptual tools that a test's stakeholders would use to construct and specify a logical argument for how the information generated by a test, observations of test performance, supports intended interpretations about test takers. This is done by visualizing such an argument as a step-by-step chain of reasoning that consists of two types of elements: constructs, such as test observations and scores, and inferences, processes that must take place in order to move from one construct to the next. When making an interpretative argument it is

common to start with the observations of test performance, to which constructs and inferences are sequentially added until the desired end point is reached, such as the interpretations that will be made about test takers. Kane's three-bridge argument (Figure 2), an early example of an interpretative argument, is relatively simple in that it consists of just four constructs and three inferences.

Figure 2.

Kane's Three-bridge Interpretative Argument (Kane et al., 1999)



Starting from the left side of Figure 1, after a test has been administered, the first inference to be made in Kane's three-bridge argument is that of an *evaluation*, the quantification of observations of performance into a useable form, in this case observed scores. Scoring rules or rubrics are developed by test stakeholders with the explicit intent to serve this purpose. This is followed by a second inference, *generalization*, use of the observed scores as estimates for what essentially is a universal missing data problem in testing, not knowing how test takers would have performed on parallel tests that were not administered. Parallel tests are tests that differ across some surface-level features, such as content, but measure the same domain of knowledge or skills. Generalizing observed scores to these missing, but still quite relevant expected scores, addresses the issue that limitations in time and testing resources means that typically very few test items are actually administered out of the universe of possible test items from the desired domain that is the target of the test. The final inference, *extrapolation*, involves the taking of expected scores and interpreting them into the target

setting, such as a measure for what test takers can and cannot do in a real-world setting. Consequently, extrapolation in this model must also serve as the justification of test use, as consequential decisions are likely to be made depending on interpretations of target scores and it is the final inference in the argument. This particular limitation of the three-bridge model will be revisited later in this chapter.

With an interpretive argument in hand, work on the second step in the process can begin, the validity argument. Constructing a validity argument involves establishing the plausibility of the individual inferences in the interpretive argument. This is accomplished by first identifying the assumptions that underlie the believability of each inference, then showing that sufficient evidence exists to support its credibility. To facilitate this process, inferences are often intentionally chosen that lend themselves readily to recognized sources of evidence that will be familiar to test professionals. Table 2, adapted from a larger table in Chapelle, Enright, & Jamieson (2011), shows how this alignment between inferences, assumptions, and supporting evidence could work in practice for the three inferences used in Kane's three-bridge framework.

Table 2.

Typical Supporting Evidence for Inferences in the Three-bridge Interpretive Argument

Inference	Assumption being made	Typical supporting evidence
Evaluation	Test administration conditions, the statistical characteristics of items, and scoring procedures including rubrics are appropriate for providing evidence for the target domain.	Prototyping studies, item and test analysis, rubric development
Generalization	The number of tasks, their configuration, and test form construction is appropriate for generalizing test takers' performances.	Generalizability and reliability studies
Extrapolation	Performance on the test is related to performance in the target context.	Criterion-related validity studies

In this way, the interpretive argument and the validity argument work as complimentary processes. The two processes serve a practical capacity in that they organize what can otherwise be a large and unwieldy collection of different kinds of evidence into one coherent narrative. The interpretive argument serves as an organizing framework by laying out what the structure of the overall argument for validity is, and making explicit the inferences that need to occur to accept a test in a given context of use as valid. The validity argument then attempts to establish a credible claim for validity by supporting each inference in the interpretive argument with relevant evidence.

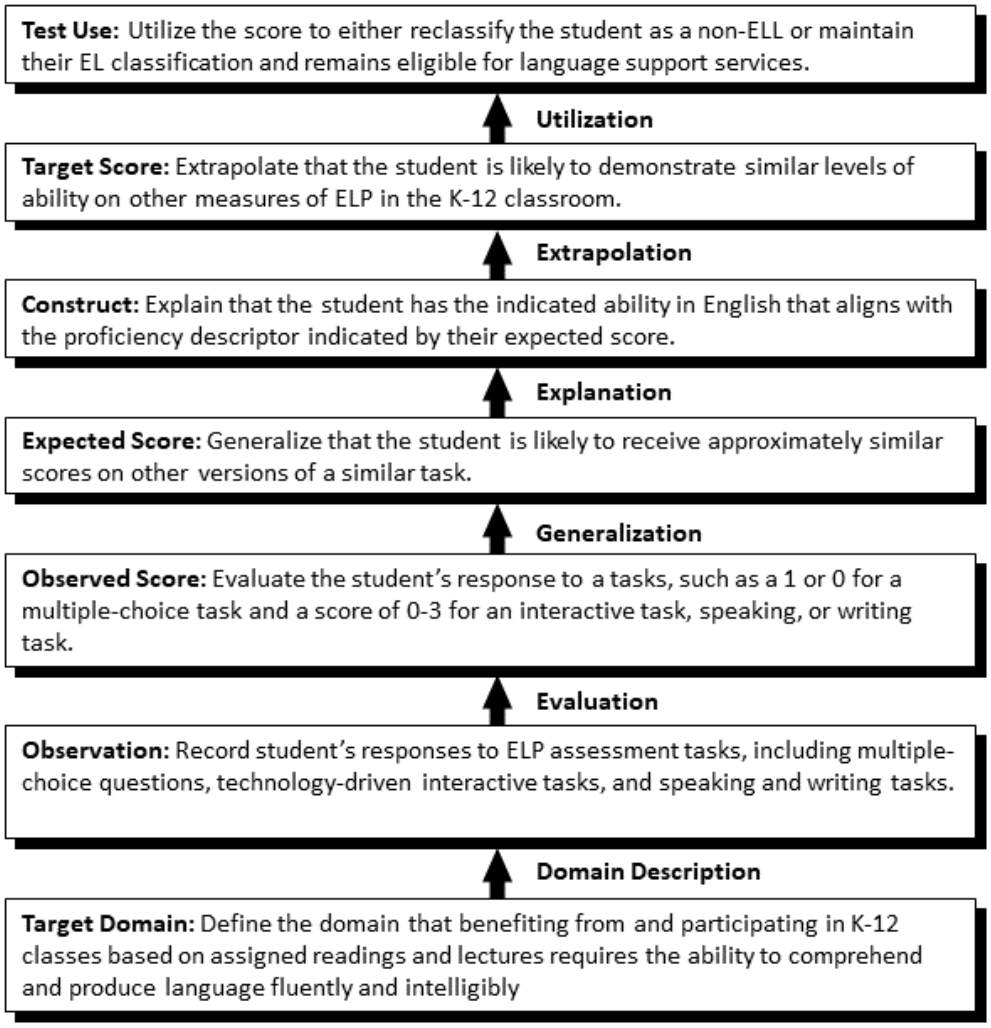
3.2 An Interpretive Argument for Summative ELP Assessments

Kane's approach to validity as described above was intended to serve as a starting point, a general reference to illustrate a system for constructing an argument through specifying a chain of inferences and supporting evidence. As test practitioners began adopting it to specific tests and testing contexts, many have chosen to incorporate a degree of customization and expansion to Kane's original arguments by modifying or introducing new inferences in response to identified needs. A well-known application of Kane's 2006 approach is the approach taken by Educational Testing Service with the new Test of English as a Foreign Language (TOEFL) (Chapelle, Enright, & Jamieson, 2011). Notable in the interpretive argument for the TOEFL is that while it has been modified with the addition of three inferences, *domain description*, *explanation*, and *utilization*, it largely retains the structure of Kane's original framework. As one of the highest profile large-scale English language tests, the approach taken by the TOEFL is likely to be familiar to and carry a weight of authority in the minds of many language testing professionals. For this reason, in this study the researcher selected it to be the basis for building a hypothetical concept for what an interpretive argument for the summative ELP assessments might look like (Figure 3). The argument shown in Figure 3 follows the same constructs and inferences of the TOEFL interpretive argument, adapting

each to the context of an ELP assessment. The argument lays out a logical process starting with the target domain, academic English language proficiency, and end with the tests current primary use, classification decisions for EL students. Note that this argument is only hypothetical, as while a great deal of evidence as to the validity of ELP assessments has been generated, this argument is only being used as an example and does not apply any ELP assessment specifically.

Figure 3.

Hypothetical Interpretive Argument for Summative ELP Assessments



Before this interpretive argument can be used to investigate the questions that are the focus of the current paper, namely whether it is valid to use summative ELP assessments to inform instruction, two potential shortcomings must be addressed. Firstly, while arguments like that in Figure 3 are well-suited for dealing with an isolated test with a single context of use, it is not clear in the framework how to address a situation where a test may be expected have more than one use. The argument as it is currently laid out has no clear place to put evidence supporting use of the test other than for classification decisions. As evidence that would support using the test to inform instruction may not be the same as for classification decisions, the argument needs a dedicated place for this evidence to go. Secondly, the kinds of evidence selected to support inferences in the current argument favor criteria used by a particular group of stakeholders, namely researchers and testing professionals. Indeed, it is often the case that these interpretive arguments favor the evaluation criteria of those in the professional testing community, often at the expense of criteria that might be more highly valued by different stakeholder groups. In order to provide information that will be directly usable by educators and school administrators in the classroom, the argument needs to somehow prioritize the perspectives of these stakeholders as well. It is clear that the interpretive argument proposed in Figure 3 will require adjustments to address both issues before it can be used to help investigate the question of whether the summative ELP assessments can help inform classroom instruction.

3.3 Making a Collaboration Argument for Summative ELP Assessments

In this section, in light of the unique challenge presented by making an argument for summative ELP assessments as part of a balanced assessment system, I propose the addition of a third argument to the interpretive/validity argument framework, a *collaborative argument*. The argument's purpose is to allow for the investigation of questions of whether summative ELP assessments can

also help inform educators in their classroom instruction. The name, collaborative argument, is explicit reference to the ultimate purpose in mind, the hope that tests like the ELP assessments be able to be used collaboratively by different groups of stakeholders and across different contexts of use. When considering and ultimately deciding on the form that a collaborative argument should take, two design features were deemed to be of high priority:

- The collaborative argument should be compatible with the interpretive/validity argument framework and share its key features: concepts, inferences, and evidence.
- A collaborative argument should help test developers incorporate and organize concepts, inferences, and evidence that are specific to uses and users of a test that fall outside the test's primary context of use covered in the interpretive argument.

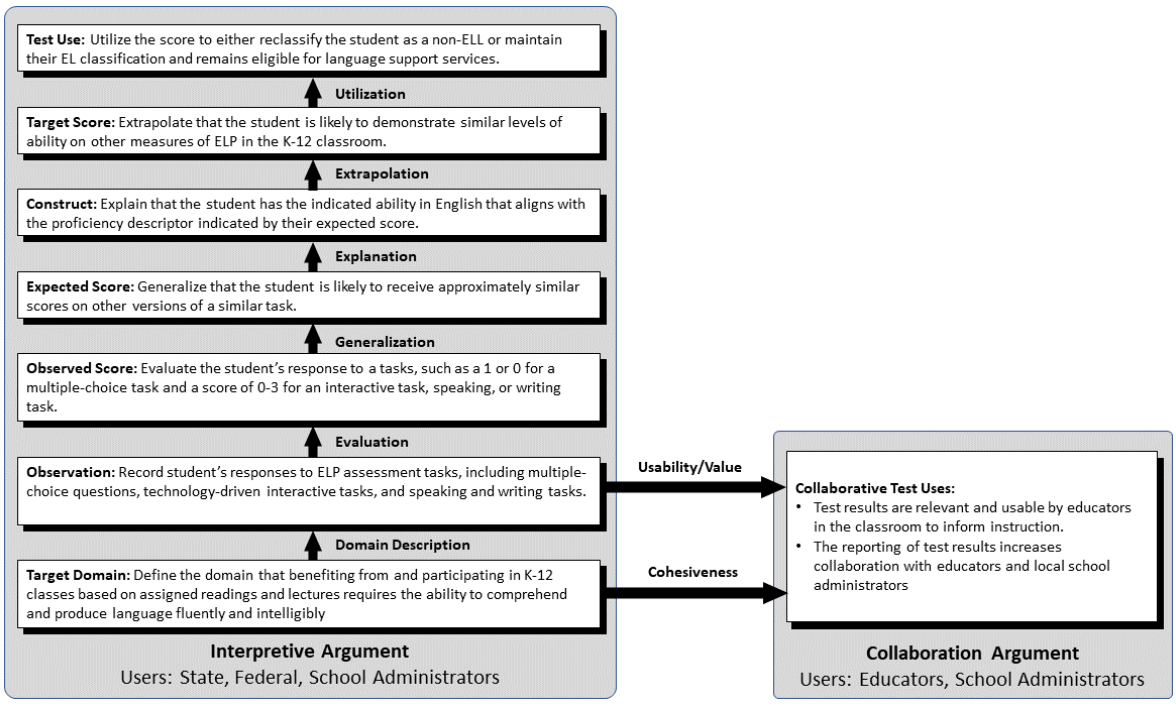
A proposed collaborative argument for ELP assessments is shown in Figure 4. Before going into a discussion of its individual components, some overall features of the collaborative argument should be highlighted. In this design, the collaborative argument is in parallel to but separate from the interpretive argument. The interpretive argument is unchanged to that shown in Figure 2 in its function as a chain of logical inferences underlying making classification decisions about EL students, the primary use of ELP assessments. In contrast, the collaborative argument specifically pertains to using the test to inform classroom instruction. In this way, each additional use for a test beyond its primary use can be accounted for with the addition of a corresponding collaborative argument. Theoretically, any number of collaborative arguments can be added to a test, although only a single collaborative argument will be considered here.

The conceptual and visual distinction made between the interpretive and collaborative arguments was deliberate. While there are connections between a collaborative argument and an interpretive argument, in recognizing that they pertain to a single test instrument and test administration, the two arguments are separated for several reasons. Firstly, this was done so that the task of making a

collaborative argument not be diminished in test developers' minds as simply an additional step in an interpretive argument. It represents a substantial undertaking to make a credible argument that using a test for a purpose unrelated to its primary intended use is valid, and it was felt appropriate that this idea be represented with a need to include a completely new argument. Second, the separation emphasizes that the credibility of a test's interpretive argument and the evidence supporting it is largely independent to that of its collaborative argument. In other words, while the interpretive argument for a test may be credible and composed of a great deal of evidence, this fact would not necessarily contribute anything to the credibility of the tests' collaborative argument. Similarly, the construction of the argument framework as in Figure 4 does not imply validity either. A credible validity argument is made only if evidence is found to sufficiently support the argument's assumptions.

Figure 4.

Proposed Interpretive/Collaborative Argument for Summative ELP Assessments



A final general feature of note concerning the collaborative argument is that collaboration requires consideration for the perspectives of different groups of stakeholders. Therefore, decisions related to concepts, inferences, and especially evidence that make up an argument framework requires incorporation of these stakeholder' opinions into the process. Attention to this fact is noted in the labeling of each argument as to the specific stakeholders they apply to in terms of the interpretive argument this would be: schools (to identify students who qualify for language services) and state and federal administrators when making accountability decisions and evaluating school ELP programs. For the collaborative argument this would be educators and school administrators (who are expected to use test results in their instruction).

Turning attention to the components that make up the collaborative argument in Figure 4, similar to how an interpretive argument is built, a collaborative argument is comprised of constructs and inferences, which together require the need for evidence to support them. Starting with constructs, these take the form of any outcomes that it is hoped will be achieved through a collaborative use of the test. In the case of ELP assessments, two have been specified here that call back to aims of the current study described in Chapter 2: (1) that test results are relevant to and useful by educators to inform classroom instruction, and (2) that this utility in the classroom have a positive impact on the collaboration between educators and local school officials with other stakeholders.

Inferences in the collaborative argument are similar to those in the interpretive argument in that they represent the processes that have to happen in order for the desired outcomes of the argument to occur. In other words, it is through investigation of and providing of evidence to support the inferences that an argument is made that a test is valid for producing the desired outcomes. These inferences are represented with horizonal arrows that link constructs in the interpretive argument to the desired outcomes of the collaborative argument. As the collaborative argument is dependent on the same test instrument as the interpretive argument, aside from its desired outcomes it does not

introduce new constructs and is instead reliant on using constructs from the interpretive argument to serve as the starting point of its inferences. The two specific inferences shown in Figure 3 where chosen as a result of a review of studies that looked at drawing links between large-scale assessments and educators activities in the classroom (Wilson & Draney, 2004; Black & Wiliam, 2004; LeMahieu & Reilly, 2004).

Based on this body of work, at least two processes have to occur for a large-scale summative assessment like ELP assessments to be used by educators in the classroom. First, the target domain of the test must overlap with the content of instruction, there must be cohesion between the content measured by the test and what educators are teaching. Second, students' performances on the test must be must be reported in ways that is both usable by educators and perceived by them to be valuable. Similar to inferences in an interpretive argument, these inferences require supporting evidence as a part of making a credible validity argument for ELP assessments. The critical importance of these inferences being a part of a collaborative argument is that it cannot be assumed that they are addressed as part of the interpretive argument. Thus, the collaborative argument outlines for test developers what assumptions need supporting evidence in addition to those in the interpretive argument to ensure that collaborative uses are credible. The next step is to define what that evidence should be. The following chapter of this report will briefly overview the general characteristics of DCMs before making an explicit connection between the kind of evidence a DCM approach can generate, the collaborative argument discussed above, and the specific research questions of the current study as they relate to the research agenda that was laid out in the introduction chapter.

4. Diagnostic Classification Models and Diagnostic Feedback

The ELP assessments are not the only context where the issues raised in the early chapters of this report are being faced. The educational measurement landscape as a whole is increasingly prioritizing large-scale assessments that maximize the instructional benefits of the information they provide (Liu, Huggins-Manley, & Bulut, 2017). It is not surprising, therefore, that the field of educational measurement has already been tackling the issue of exploring potential ways to get more detailed information about students from assessments that are summative, administered infrequently, or in many cases, both. A promising methodological development to come out of this work are diagnostic classification models (DCMs; Rupp, Templin, Henson, 2010).

As a type of diagnostic measurement model, DCMs are specifically designed to provide fine-grained feedback (i.e., multidimensional) and support criterion-referenced interpretations, precisely the sort of profile-based feedback being called for in the instructional context (Rupp, Templin, & Henson, 2010). Goodman and Huff (2006), for example, found that more than half of teachers surveyed do not believe that large-scale and commercial assessments provide detailed enough information about student's abilities. A high majority (85%) indicated that it is very important for feedback to provide detailed information about the strengths and weaknesses of students with respect to specific knowledge, skills, and abilities. Others have made a similar argument that profile-based, richer interpretations of student ability are more effective than other reporting methods when the purpose of a test is to directly support instruction and learning (Roussos et al., 2007). Also frequently cited is that feedback should be readily understandable and actionable, as in policy documents intended to advise states in the development of their ESSA assessment plans, including ELP assessment:

> A State might use State assessment grants to design easy-to-understand State and LEA report cards or improve the quality of individual student interpretive, descriptive, and

> diagnostic reports to help educators, parents, and families to understand and address the specific academic needs of students. (US Department of Education, 2016).

This has been interpreted to suggest that categorical-based scoring may also be an important characteristic for feedback to have, such as showing whether a student was a master or non-master at a skill rather than a scale score.

DCMs provide a number of statistical advantages when the intended use is to diagnose students' strengths and weaknesses at specific skills, namely those related to the dimensionality, reliability, and interpretability of the information provided. This is particularly true compared to models based on continuous scores, such as cut-score setting, which are prone to error (Templin, Cohen, & Henson, 2008). A number of studies have investigated the usefulness of DCMs to provide teachers with instructionally useful information, including an English language proficiency assessment (Jang, 2005), an assessment of basic geometry skills (Henson & Templin, 2008), and an international mathematics assessment (Tatsuoka, Corter, & Tatsuoka, 2004; Xin, Xu, & Tatsuoka, 2004). In addition, attempts have also been made to retrofit a DCM to an existing standardized test, as would be necessary in the case of the ELP assessments (Liu, Huggins-Manley, & Bulut, 2017; Lee & Sawaki, 2009). Before exploring further how this could be applied in relation to investigating the research questions proposed in the current study, it would be worthwhile to present in more detail the technical characteristics of DCMs.

4.1 Technical Characteristics of DCMs

DCMs are a type of latent class psychometric models that relate the directly observed responses of an individual, such as performance on a test item, to one or more unobservable latent characteristics, such as proficiency in an academic subject or skill. DCMs are similar to other psychometric frameworks in this regard, namely item response theory (IRT), factor analysis (FA),

and structural equation modeling (SEM). With a general audience in mind, only an overview discussion of the major technical characteristics of DCMs will be presented in this chapter, and for sake of simplicity will largely focus on aspects of the model that are relevant to the current study. There are however excellent resources available in the literature that provide more in-depth discussions of the DCM models and its major variants (Rupp, Templin, & Henson, 2010; Roussos, Templin, & Henson, 2007; Templin & Bradshaw, 2014).

Compared with DCMs, other models in the latent-variable model family share one key difference, the nature of the latent variable. A unique aspect of DCMs is that latent traits, referred to as attributes, are modelled to function as categorical variables with only a few distinct levels. In contrast, other models assume the latent traits to be continuous, normally distributed variables. This means that IRT for example, attempts to rank all students relative to each other based on their scale scores on a latent trait (θ_r), appropriate for making norm-referenced interpretations.

In contrast, though DCMs can and have been used for polytomous attributes, most standard applications have defined binary attributes having only two levels, a *mastery* level for individuals who are able to perform the skill, and a *non-mastery level* for those who are not. No ranking is modeled, and the mastery classification is appropriate for criterion-reference interpretations. The model then allows for taking a collection of observations for an individual and interpreting them as the result of a classification profile for that person that represents their estimated ability level on a pre-determined set of attributes. In determining these classification profiles, DCMs define the probability of observing an examinee's response pattern $P(X_r = x_r)$ as a function of two components, a structural component to explain the proportions of the latent attributes, and a measurement component to explain how responses on assessment items are related to the attributes (Equation 4.1).

$$P(X_r = x_r) = \sum_{c=1}^{C} v_c \prod_{i=1}^{I} P(X_{ir} = 1|\alpha_r)^{x_{ir}} (1 - P(X_{ir} = 1|\alpha_r))^{1-x_{ir}} \quad (4.1)$$

As the measurement component (Equation 4.2) is more intuitive to understand, and of greater relevance to the current study, its properties will be described first. The measurement is comprised of two probabilities. The first denotes the probability that examinee *r* responds to the item *i* correctly $P(X_{ir} = 1|\alpha_r)$ conditional on their attribute profile α_r. The attribute profile α_r can be thought of conceptually as analogous to the ability parameter θ_r used in IRT. The difference is that while θ_r is a single number that represents an examinee's relative location on a normally distributed latent trait, since the DCM is based on categorical classifications α_r takes the form of a vector of a length equal to the number of attributes represented on the test, with values of 1s or 0s indicating the attributes for which an examinee is classified as a master or non-master respectively. The DCM assumes that the probability of an examinees responding correctly to an item is dependent on the values in this vector, similar to how IRT assumes a response to be dependent on the value of the examinee's ability parameter.

$$\prod_{i=1}^{I} P(X_{ir} = 1|\alpha_r)^{x_{ir}} (1 - P(X_{ir} = 1|\alpha_r))^{1-x_{ir}} \quad (4.2)$$

The second probability in the equation, $(1 - P(X_{ir} = 1|\alpha_r))^{1-x_{ir}}$, represents that of an incorrect response, equivalent to 1 minus the probability of a correct response for dichotomously scored items. The exponents of the probabilities are necessary to dictate the rule for which probability is to be used for each item for each examinee, as determined from the examinees' score on that item, x_{ir}. As this term takes a value of 1 in the case of a correct response, the exponent of the probability for a correct response also takes a value of 1, meaning that it will be used in the

calculation, while the exponent for the probability of an incorrect response will take a value of 0, effectively cancelling it out. In the case of an incorrect response, the opposite will be true.

To illustrate using an example, imagine a 3-item test for which an examinee answers the first two items correctly and the last one incorrectly. Their corresponding response pattern can be represented by the vector $x_r = [1,1,0]$. Referring back to Equation 5.2, the measurement component for this hypothetical examinee would estimate the probability of their response pattern as the product of the probabilities shown in Equation 4.3:

$$P(X_{1r} = 1|\alpha_r) \times P(X_{2r} = 1|\alpha_r) \times (1 - P(X_{3r} = 1|\alpha_r)) \quad (4.3)$$

We next turn to how the DCM expresses these probabilities. At this point there is a decision to be made, as different diagnostic models make different assumptions concerning these probabilities. For purposes of the current study, the focus will be on the use of one model in particular, the LCDM, or the Log-linear Cognitive Diagnosis Model (Henson, Templin, & Willse, 2009). The LCDM can be considered as an extension of an earlier model, the General Diagnostic Model (GDM), with two additional assumptions. Firstly, all responses are assumed to be binary, and secondly, conjunctive effects of an individual having two or more attributes are allowed (von Davier, 2014). In addition, the usefulness of the LCDM comes from its subsuming of other DCMs (Templin, 2016), meaning that as a general model, it dictates few restrictions about the nature of the attributes compared with other DCMs and can be expected to fit the data as well as them.

The LCDM does assume all responses to be binary, such as when a response to a test item is scored as either correct or incorrect. The LCDM uses a log-odds function, also known as a logit function, which is calculated as a function of the probability of either two possible responses (Equation 4.4).

$$\text{Logit}(X_{ir} = 1|\alpha_r) = \ln\left(\frac{P(X_{ir} = 1|\alpha_r)}{1 - P(X_{ir} = 1|\alpha_r)}\right) \quad (4.4)$$

Equation 4.4 can be read as the log-odds of a correct response being equal to the natural log of the probability of a correct response divided by the probability of an incorrect response. This equation is nearly identical to how probabilities are modeled in binary cases of item response theory, the sole difference being the use of an examinee's attribute profile α_r in place of their ability parameter θ_r. As one can expect, the rationale for using the logit function is shared by both models, namely it restricts the predicted probabilities to only take values that are realistically possible (i.e. values less than 1 and greater than 0).

The key function of the measurement component of a DCM is that it relates responses on items to the latent attributes of examinees. Responses on items are represented as the log-odds of a correct response as in Equation 4.4. In order to understand how the latent attributes are incorporated, it will be useful to consider an example from an actual testing context for which attributes are already known. For a test of arithmetic, an examinee's performance could be thought of as being dependent on their status on four attributes that underlie arithmetic proficiency: addition, subtraction, multiplication, and division. Performance on a specific item on the arithmetic test can be thought of as being dependent on the examinee's mastery status on each of the four attributes, as well as the specific combination of attributes related to the item, as all items must require at least one attribute to answer correctly, but not necessarily all of them simultaneously. For instance, an examinee's response to an item consisting of solving a problem with a subtraction and multiplication component would be dependent on the examinee being a master of both those attributes, but independent of their level on the addition and division attributes (Table 3).

Table 3.

Required Attributes for a Sample Arithmetic Item

	Required Attributes			
Item	**Addition**	**Subtraction**	**Multiplication**	**Division**
$3 \times 8 - 16 = ?$	0	1	1	0

The LCDM takes this information about which items measure which specific attributes and relates it to the log-odds of an examinee correctly responding to an item using a form of the general linear model (GLM). Equation 4.5 shows an example of the GLM for an item similar to the example, where two attributes are thought to underlie performance on the item.

$$\text{Logit}(X_{ir} = 1|\alpha_r) = \lambda_{i,0} + \lambda_{i,e(a1)}\alpha_{r1} + \lambda_{i,e(a2)}\alpha_{r2} + \lambda_{i,e,(a1,a2)}\alpha_{r1}\alpha_{r2} \quad (4.5)$$

Equation 4.5 can be read as the log-odds of a correct response being equal to the sum of four components: an intercept ($\lambda_{i,0}$), the base-level log-odds of an examinee answering the item correctly having mastered neither attribute, the main effect of attribute 1 ($\lambda_{i,e(a1)}$), the change in log-odds of answering the item correctly having mastered attribute 1, the main effect of attribute 2 ($\lambda_{i,e(a2)}$), the change in log-odds of answering the item correctly having mastered attribute 2, and a two-way interaction term ($\lambda_{i,e,(a1,a2)}$), the change in log-odds of answering the item correctly having mastered both attributes. As it is based on the GLM, the formula in Equation 4.5 is identical to a reference-coded ANOVA model with two categorical independent variables. The interpretation of the individual components, also referred to as item parameters in the LCDM, is similar as well. In this way, an examinee's probability of a correct response for any given item can be understood as a function of the item parameters as well as which attributes the examinee has mastered. Taking

Equations 4.4 and 4.5 and solving for $P(X_{ir} = 1|\alpha_r)$, the probability that examinee *r* responds to item *i* correctly, yields Equation 5.6 (in the case of an item measuring two attributes).

$$P(X_{ir} = 1|\alpha_r) = \frac{\exp(\lambda_{i,0}+\lambda_{i,e(a1)}\alpha_{r1}+\lambda_{i,e(a2)}\alpha_{r2}+\lambda_{i,e,(a1,a2)}\alpha_{r1}\alpha_{r2})}{1-\exp(\lambda_{i,0}+\lambda_{i,e(a1)}\alpha_{r1}+\lambda_{i,e(a2)}\alpha_{r2}+\lambda_{i,e,(a1,a2)}\alpha_{r1}\alpha_{r2})} \quad (4.6)$$

The probabilities that make up the measurement component of the DCM are calculated using Equation 4.6. Essentially what the measurement component of the model does is to take as inputs the observed examinee responses and attribute designations for each item (similar to Table 2), and iteratively estimate item parameter values and examinee attribute profiles in a way that maximizes the probability of the observed responses.

An important consideration related to the measurement component of DCMs is the question of how it is determined what attributes get assigned to specific test items. DCMs are confirmatory models in that they require as a prerequisite both specification of the attributes themselves and assignment of the attributes to items. In other words, the model itself is not intended to have the capacity to uncover from test response data the nature of attributes that may underlie test performance, nor the relationship between attributes and their association with individual items. Rather, these conditions should be pre-specified by the researcher prior to running the model. In DCMs, this specification takes the form of a *Q*-matrix, an *i* x *a* matrix where *i* is the number of items on the test and *a* is the number of attributes. The rows of the *Q*-matrix specify the item and attribute relationships, similar to Table 2 but with the number of rows corresponding to the total number of items on the test. An example *Q*-matrix for a 5-item test with 3 attributes is shown in Table 4.

As in Table 2, 1s and 0s indicate whether an attribute is measured by a test item or not. While it is possible for the numbers in a *Q*-matrix to take other values, such as when items may differ in the degree to which an attribute is measured, for purposes of the models used in this study only values of 1 or 0 are possible. While Q-matrices are intended to be expert defined based on a qualitative

review of items, a recent area of work is exploring the application of algorithms to identify potential Q-matrix structures from computer simulations (Chung, 2014.)

Table 4.

Sample *Q*-matrix for a 5-Item Test

	Attribute 1	**Attribute 2**	**Attribute 3**
Item 1	1	0	0
Item 2	0	1	0
Item 3	0	0	1
Item 4	1	1	0
Item 5	0	1	1

A final point to address before continuing onto the practical benefits of DCMs relates to the other component in the DCM, the structural component (Equation 5.7). As a full discussion of how this component functions is not necessary for the current study, it will only be briefly introduced here.

$$P(X_r = x_r) = \sum_{c=1}^{C} v_c \tag{5.7}$$

As was mentioned earlier in this chapter, the DCM works by estimating the probability of the observed response pattern made by examinees on a test. To do this the model uses not only the information contained in the measurement component, but also estimations of the overall probability in the population of being a member of each attribute profile. The structural component serves to add this information into Equation 5.1 with a summation function containing probability estimates for every possible attribute profile v_c, such that a total of c probabilities are estimated where $c = 2^A$ (with A being the total number of attributes). A test measuring three attributes, for

example, would have 2^3, or eight possible attribute profiles: 000, 100, 010, 001, 110, 101, 011, & 111, and eight probability estimates representing the overall probability in the examinee population of having each profile. Several valuable pieces of information about attributes and the relationships between them are recovered from these estimated overall membership probabilities, namely the probability of mastery of individual attributes and inter-attribute correlations.

Functionally, the structural component is useful when fitting DCMs to data as it can be manipulated in specific cases when researchers want to impose conditions on which parameters are estimated in the model, most often based on hypotheses they might have about the relationships between attributes. The technical specifics for how these manipulations are done is beyond the scope of the discussion in this paper (see Xu and von Davier (2008) for an example). For all models run in the current study no assumptions were made about the relationships between attributes. In other words, all structural component membership probabilities were estimated, an approach to structural model specification referred to as a *log-linear parameterization*.

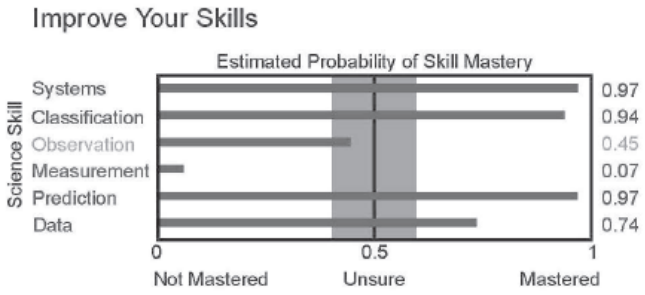
4.2 DCM Score Reports as Diagnostic Feedback

A distinguishing and advantageous feature of DCMs is that they provide the benefit of translating an examinee's score pattern into diagnostic feedback in the form of an attribute profile, or an estimate of whether the examinee is a master or non-master at specific skills that are measured by the test items. While the information provided by a DCM is therefore multidimensional, unlike a single numeric test score, its true value lies in the fact that it can report this information reliably. Other scoring models are capable of providing similar information in theory, but would require far too many items to be practical or feasible in the context of the ELP assessments. Furthermore, the interpretation of a student being a master or non-master at a skill is likely to be more grounded and relatable to the classroom context than a proficiency level or scale score, important when teachers

and other stakeholders have been calling for assessment systems that provide as much diagnostically relevant information as possible about their students (Leighton & Gierl, 2007). To visualize how this might look in practice, consider the example diagnostic score report showing a student's performance on a test that measures their ability in six science skills (Figure 5).

Figure 5.

An Example of a Diagnostic Score Report



The score report contains three key pieces of information, (a) the skills presumed to be measured by the test, (b) whether the student should be classified as a master or non-master for each attribute, and (c) the level of uncertainty associated with the classification. For example, it is readily clear that the report communicates information about the student's ability at the six skills listed on the left-hand side. This particular student has a high probability of being a master of four of those skills: systems, classification, prediction, and data, but the student also has a high probability of being a non-master of the measurement skill. Also evident in the report is that there is not enough information to make a determination whether the student is a master or non-master of one skill, the observation skill. Overall, the information in the report in Figure 5 takes up little space in efficiently communicating its' message in the kind of clear and simple language recommended for test score reporting in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014).

Compare the diagnostic score report in Figure 5 to one similar to those currently in use for reporting student performance on a major standardized ELP assessment, the ISR for the ELPA21 shown earlier in this report (Figure 1). The first thing that is apparent is the amount of information appearing on the report, in part due to the inclusion of lengthy descriptions of proficiency levels and domains. The layout of the report calls to mind the term "data dump", a word coined by developers of early test score reports to refer to the problematic habit of seeing a score report as an opportunity to cram as much information as possible onto a single page (Zenisky & Hambleton, 2012). An argument could be made that these kinds of explanations are not necessary in the diagnostic report, where skill mastery classifications may be more readily understandable by most stakeholders. In contrast, it is less clear for example what information is being communicated from the student having a proficiency status of "Progressing" without accompanying explanation. Similarly, in the domain level feedback provided in the bottom section of the report, it is not intuitively clear what is meant when a student possesses a "Level 2" reading level without further detail. The domain level feedback section of the report is further complicated by the inclusion of scale scores. While the scale scores contribute to the score report by adding a sense of the level of uncertainty by including a range of error around each score, these may be largely unnecessary for most stakeholders as it is not clear from the report how the scores translate into performance levels or how the uncertainty impacts this translation.

Ryan (2006) proposed a conceptualization of test score reporting as a form of communication, with a sender, message, medium, intent, and audience. In the case of ELP assessment score reporting, if the intent is to for ELP assessments to communicate usable information to an audience composed largely of educators, it is clear that much about the score report in Figure 1 is lacking. While no study has thus far investigated the instructional impact of diagnostic score reports compared to other report formats, there are several reasons to suspect that diagnostic reports would

represent at least some improvement. Firstly, as discussed above, attribute profiles present information about students in an efficient, clean format that minimizes the need for clutter in the form of explanatory text. Stakeholders of all types can more readily interpret what a report means when it says a student is a master or non-master of a particular skill more so than a scale score or proficiency level with little context for what it represents. Secondly, the concept of uncertainty can be more clearly communicated with less ambiguity in the case of a diagnostic score, where it simply means that it cannot be determined with confidence whether a student has mastered a skill. In the case of a scale score or proficiency level it is less clear how to interpret uncertainty or what its consequence should be. Finally, diagnostic reports address an ongoing issue related to the grain-size of feedback communicated in score reports. A requirement of NCLB dictates that individual student reports should allow teachers, principals, and parents to understand and address the specific needs of students (NCLB, 2002). With this in mind, it is unclear how effectively the domain-level feedback like that in the score report in Figure 2 would allow for an understanding of a student's language learning needs, as student needs are rarely if ever defined by the domains of language. In other words, it could be argued that students of all levels need reading, writing, listening, and speaking. Even if a student were deemed highly proficient in reading, it is unlikely that any knowledgably language teacher would deem that student as not needing reading instruction, or take action in the classroom to not teach reading to that student in favor of other language areas.

Put another way, the domain scores represent a trade-off of grain-size for precision that may not reflect the needs particular to educators. The domain scores likely represent the smallest grain information that can currently be provided because the report insists on providing all information in the form of continuous scale-scores. More tractable for instructional decision making may be the type of skill-based feedback provided by a diagnostic report, which would take the opposite approach and trade precision for grain-size. As DCM based score report would only be concerned

with reporting mastery or non-mastery, it would be able to report information such as specific skills within the domain of reading, something likely to be more directly relatable to can-do statements that connect to tasks that teachers are using in the classroom. The subsequent loss of the precise scale scores may be of little concern. Naturally, care would need to be taken to ensure that the particular skills chosen to appear on a score report are those that would be of value to educators in a particular school, but assuming this were the case, diagnostic score reports could serve a valuable role in linking standardized assessment performance to the classroom.

4.3 Concerns with Retrofitting a DCM to an Existing Assessment

As alluded to briefly at the start of this section, a retrofitting approach would be necessary in order to apply a DCM to a standardized ELP assessment and generate diagnostic feedback like that shown in the previous section. In a retrofit approach, existing test items are coded to attributes in a cognitive model in a *post hoc* fashion, and then analyzed for their capacity to provide meaningful diagnostic information (Jang, Dunlop, Wagner, et al., 2013; Li, Hunter, & Lei, 2015). This is due to the fact that unlike the case of a principled test design approach to creating a diagnostic assessment, where a cognitive model explaining the attributes measured by test items is first developed, followed by the writing of items based on that model, in the case of the ELP assessments the tests are already in use with development having happened for purposes other than those related to providing diagnostic feedback.

The topic of retrofitted DCM models is one with a relatively high degree of controversy in the literature. It is generally agreed upon that most applications of DCMs to date have by necessity been cases of retrofitting (Gierl & Leighton, 2007), in part due to the relatively recent development and limited cases of where DCMs have been applied in operational testing contexts. However, perspectives on the worthwhileness of retrofitting a DCM range from positive (Lee & Sawaki, 2009;

Davidson, 2010; Liu, Huggins-Manley, & Bulut, 2017), to neutral (Rupp, Templin, & Henson, 2010), to outright dismissive (Gierl & Cui, 2008; Alderson, 2010). The perspective taken in the current study as to the question of whether retrofitting a DCM can provide useful information is one of cautious optimism of the exploratory value that such an approach would contribute to an understanding of how to better use information from the summative ELP assessment. However, in order to diligently address the specific criticisms highlighted by the latter group of papers in particular, the currents study's response to a few of the most commonly cited challenges to retrofitting a DCM approach will be discussed to conclude this section of the report.

Retrofitted DCM approaches, especially those done on large-scale standardized tests like the ELP assessments, are often not seen as a viable way to fulfill educators' needs in the classroom for instructional feedback. In response, it is necessary here to revisit a point made earlier in this paper in regard to an important distinction between formative assessment, the primary means by which educators should elicit and act upon instructionally relevant information about their students, and the fine-grain feedback that could be possible on the summative ELP assessment using a DCM. Critics of the formative usefulness of this kind of feedback are correct in that the depth and quality of DCM-derived information provided by a standardized assessment cannot compare to what can be provided by other assessments in the educational system, namely interim assessments and teacher's ongoing classroom assessment practices. However, holding these assessments to the criteria by which formative assessments are defined paints a bleaker picture than what is likely the reality. The ways of reporting information explored in the current study may never fulfill everything that is needed out of a formative assessment, however, the aim of the current study is perhaps best expressed by Sawaki & Lee (2010) in that, "…virtually all types of language assessments can have some diagnostic value. Our task should be then to identify the factors that make a test more or less diagnostic…and how those factors interact among themselves." (p.109) Currently, summative ELP

assessments are likely providing little or no useful information to educators and other stakeholders that is applicable to the classroom context, but it is worth exploring whether they have the potential to do better.

Another criticism of the retrofit approach relates to the defensibility of researchers' *post hoc* selection of attributes and assignment of them to test items, given that the items were not developed under a cognitive model to begin with. In this sense there is truth in saying that even with rigorous protocols in place, the step of defining a Q-matrix in retrofitting is an inherently subjective process, and insufficient for the diagnostic feedback generated to serve as comprehensive evidence for understanding the underlying processes by which examinees' performing the tasks appearing on the test, in other words, a precise theory for which attributes students are using and to what degree. In response, the current study would emphasize that retrofitting a DCM to a standardized test is fully intended to be a fundamentally different undertaking than the principled construction of a fully diagnostic assessment, and therefore in agreement that only the latter test is capable of providing evidence about examinees' processing skills. The goal with a retrofit approach is less about uncovering new theory about the processes that examinees deploy when completing test tasks, and more about applying what theories already exist in service of generating the most useful (i.e. accurate) diagnostic information. There is little value in scrutinizing Q-matrices in retrofit approaches beyond the limited role they are intended to play, namely serving as the best possible item-level blueprint connecting the content of test tasks to the skills a test presumes to measure.

This focus on content can contribute further benefits to retrofit DCM approaches of standardized assessments beyond generating the best possible diagnostic information. As Davidson (2010) articulately stated:

> CDA [another name for DCM] is a procedure, and it does one very important thing that normative item analysis has long overlooked: It values the content of a test task. Even more,

> it asks that test developers portray that content in a well-reasoned and conscientious manner…CDA gives us a way to articulate rich discussions about test content in a procedural manner. (p. 106)

What Davison is referring to here is not meant to suggest that standardized tests like the ELP assessments are entirely unconcerned with content. Many tests in fact take steps to faithfully address content in their item design process, typically by following a protocol based on some kind of evidence-centered design. In the case of the ELP assessments, the ELP standards and development paths serve as a common underlying framework intended to connect the test to classroom content. But an issue with the standards is that while they were developed by panels of content experts, and test items written with coverage of them in mind, beyond that there is little way of empirically answering important questions about how well the standards are represented in the ELP assessments or how well the development paths actually capture how the typical EL student progresses through the standards. It is also a valid criticism that often times in the case of large-scale and standardized assessments that normative item analysis, particularly in cases where unidimensional scoring models are used in the evaluation of test items and the designing of test forms, acts to prioritize the statistical performance of items to distinguish examinees from one other over an item's content, potentially putting statistical calculations at odds with the judgements of content experts. A DCM approach to ELP assessments could be a place to start looking into content related questions like these, serving as a way to deepen conversations amongst test stakeholders about the appropriateness of test content, and supplement the evidence-centered design process with content-related empirical data. A second benefit to this approach would be to allow various stakeholders to better visualize what skills the summative ELP assessments sufficiently measures, and those that it does not. As alluded to earlier, it is likely that the assessment items are working to measuring some, but not necessarily all the skills that will be relevant for EL students'

learning. Aggregate scores like those appearing on the current ISR make it impossible for educators to interpret which skills may have been adequately measured by the summative assessment, giving them little context when designing and using their own formative assessments. However, reporting summative assessment scores as finer-grain feedback would give educators a more useful profile of their EL student's proficiency, including where additional assessment is needed, allowing them to better target their own classroom assessments around those areas.

One area of concern where both supporters and detractors of retrofitted DCM approaches are in agreement is a need to properly investigate whether a test provides sufficient coverage of attributes to generate reliable, consistent diagnostic feedback. DCMs are statistical estimation models, and as with any such models, there are limitations dictating the conditions under which they can provide accurate and reliable results. With retrofit DCMs, *Q*-matrix design is of particular concern in this regard. Ongoing research has uncovered a number of *Q*-matrix conditions that are believed to impact how effectively the model is able to reliably estimate diagnostic classifications, a selection of which are shown in Table 5.

Table 5.

Properties of *Q*-matrices that Impact DCM Estimation

Condition	**Source(s)**
Fewer than 10 total attributes measured	DiBello, Roussos, & Stout (2007)
At least 1 unique item per attribute	Madison & Bradshaw (2015)
(3) or more items per attribute	Madison & Bradshaw (2015)
Maximum of 2 attributes measured by a single item	Liu, Huggins-Manley, & Bradshaw (2017)

Meeting the conditions listed in Table 5 however is not a straightforward task, particularly in retrofit cases where test items and forms may already be largely decided. For one, a high degree of overlap across attributes is a commonly encountered problem in educational tests, where for example, a pair

of attributes may be highly related to one another and therefore appear frequently together in the same item. Where there is not good distribution of attributes across items, the DCM is likely to encounter difficulties when estimating item parameters, and the accuracy and reliability of examinees' classification profiles may suffer (Madison & Bradshaw, 2015). Another common issue is when an attribute may be associated with such a basic or foundational skill that all items on a test potentially measure it. On a test of reading ability for example, having an attribute defined as "can understand the meaning of short, simple words", would make it very difficult to imagine a type of item that would not measure such an attribute. A test where an attribute is measured by all items would complicate a *Q*-matrices' ability to meet one of the conditions, the need for at least 1 unique item per attribute.

It should be noted that the conditions in Table 5 are meant only as rough guidelines, and the actual numbers they dictate and their relative importance to having an impact on reliability noticeably differs from study to study. Not only do the conditions interact with one another, but they are highly dependent on contextual factors such as the type of test being administered and the examinee population. However, as unreliable classifications would represent a significant a threat to the validity of using diagnostic feedback in any context, any retrofit DCM approach should take these conditions into consideration.

4.4 Using a Retrofitted DCM to Make a Collaborative Argument for the ELP Assessments

Bringing together the major concepts discussed in the report thus far, Table 6 outlines a conceptual framework connecting retrofitted DCM approaches, the collaborative argument discussed in Chapter 3, and the questions raised in Chapter 1. The framework forms the basis for the approach to arguing the validity for using summative ELP assessments to inform classroom

instruction that is the focus of the current study. Next, the precise methods that were used to collect the supporting evidence shown on the right-most column in Table 6 will be discussed.

Table 6.

Evidence to Support Inferences in a Collaborative Argument for ELP Assessments

Inference	Assumption being made	Supporting evidence
Cohesiveness	The target domain measured by ELP assessments overlaps with the content being taught in the classroom.	• A DCM scoring approach to ELP assessments shows that a *Q*-matrix can map test content to skills that align to ELP learning standards • DCM scoring suggests that performance on the ELP assessments can be represented as a student's proficiency at language skills that align to ELP learning standards
Usability/Value	Student performance on ELP assessments can be reported in ways that are understandable to educators, reliable, and relatable to the classroom context.	• DCM scores can be used to create diagnostic score reports that educators find more interpretable than overall score reports • DCM scores based on performance on the ELP assessments are sufficiently reliable to serve as diagnostic indicators of students' abilities at specific language skills

5. Research Questions

The summative ELP assessments were not intentionally designed to provide fine-grain or skill based feedback. However, it may be possible to retrofit a DCM to an assessment that would allow it to provide this kind of information to educators about which language skills students have or have not mastered. The current study attempts to apply a DCM for this purpose to one summative ELP assessment, the English Language Proficiency Assessment for the 21st Century (ELPA21). The primary motivation for the study being to investigate a potential way that the ELPA21 could better support educators in their classroom instructional practices.

Based on the principles behind Kane's (2006) validity argument framework, Chapter 3 proposed the making of a collaborative argument to support the use of a DCM for allowing an assessment to provide diagnostic information. The argument consists of two inferences: (1) *Usability/Value*, and (2) *Cohesiveness*. In a validity argument approach, if a strong case is to be made for applying a DCM to the ELPA21 for this intended purpose, there is a need for research that would identify and provide credible evidence to support both these inferences. The diagram displayed in Figure 6 outlines the approach taken in the current study to support both inferences, focusing on three research questions.

Research Question 1

Do educators think ELPA21 diagnostic score reports would be usable and valuable for informing their classroom instruction? How does this compare to how educators feel about the current ELPA21 score report?

Research Question 2

How capable are ELPA21 diagnostic score reports at providing feedback that aligns to state ELP standards?

Research Question 3

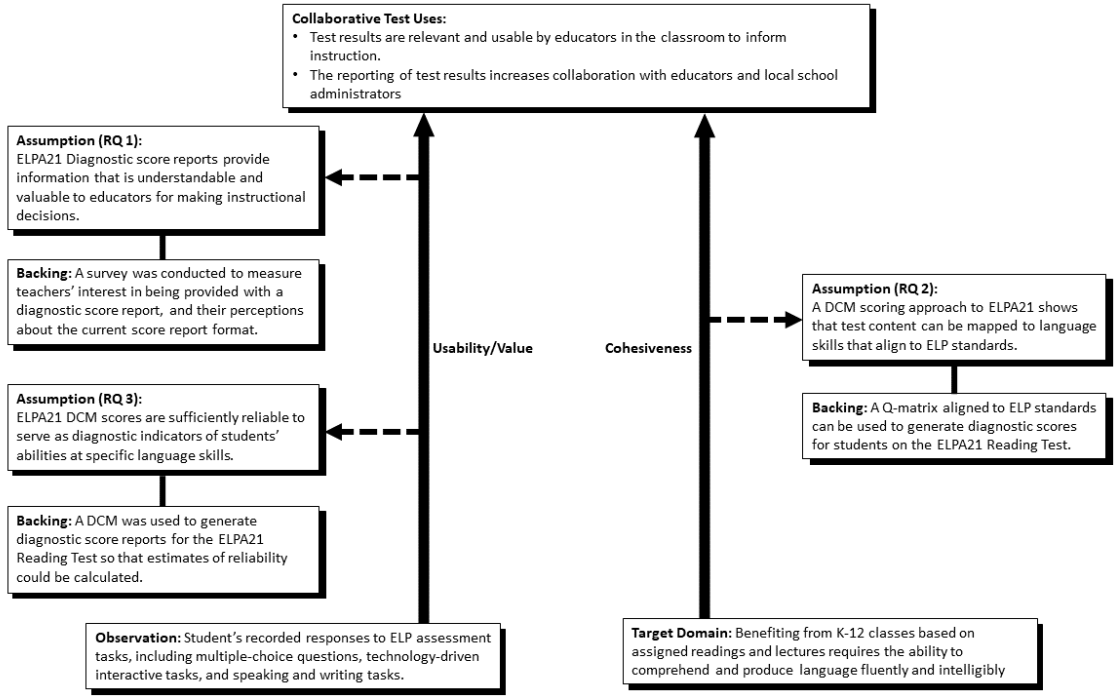
How capable is a DCM at capturing diagnostic information the ELPA21? Does it generate diagnostic scores that are sufficiently reliable for use as diagnostic information?

Each research question corresponds to a piece of evidence to be collected and analyzed as backing for the inferences. The first piece of evidence, corresponding to Research Question 1, was a survey to measure teachers' perspectives about the current ELPA21 score report and the potential interest in diagnostic score reports. The second, corresponding to Research Question 2, was the designing and analysis of a *Q*-matrix based on state ELP learning standards. Finally, an investigation of DCM model fit and score reliability was conducted on data from a recent administration of the ELPA21,

corresponding to Research Question 3. The following section of this report provides a brief overview of the ELPA21 and how it was used as the subject of investigation in this study.

Figure 6.

Evidence to Support a Validity Argument for Applying a DCM to the ELPA21



6. The ELPA21

The summative ELPA21 is an annual end-of-year assessment primarily used to determine the EL classification status of students and for reporting school, district, and state progress towards accountability targets. The test is based on state standards that align with the language skills that students would need to access content in the K-12 classroom related to college and career readiness standards in English language arts, mathematics, and science. A consortium of eight states collaborate together to share responsibility for the administration, maintenance, and ongoing

evaluation of the assessment. To ensure grade appropriateness of test content that must span the entire K-12 curriculum, the test has six distinct forms each with their own unique item pools: Kindergarten, Grade 1, Grades 2-3, Grades 4-5, Grades 6-8, and Grades 9-12. There are four sections to the test, each measuring one of the language skills: Reading, Listening, Speaking, and Writing. A variety of item types appear in the sections, including fixed-response, multiple-response, and free-response items. The current score report format provided to educators, schools, and parents (Figure 1) includes a student's performance on each test section, reported as one of five performance levels. In addition, an overall score classifies the student as an EL, or if they scored high enough to no longer be in need of language support services, are reclassified as English Language Proficient (ELP).

An issue of practicality when fitting a DCM to any test concerns the complexity of test items and the overall length of the test. The ELPA21 sections that measure the two receptive skills, listening and reading, are largely composed of a high number of dichotomously scored multiple-choice items. In contrast, the sections measuring the productive skills, speaking and writing, have a fewer number of items in total, many of which are open-response items. As fitting a DCM model to either the speaking or writing section would pose a significantly more complex challenge than with either receptive skill, the current study focuses exclusively on doing an in-depth investigation of the ELPA21 reading section as a preliminary proof-of-concept approach to using a DCM with the test. Naturally, extending the validity argument to the test as a whole would require separate investigation of the other test sections in a similar fashion to the current study. To keep the amount of data to a manageable level for an exploratory study such as the current one, three out of the six forms of the test were chosen to be analyzed in regard to Research Questions 2 & 3: Kindergarten, Grades 4-5, and Grades 9-12, representing a range of test forms designed and targeted at the youngest, middle, and oldest EL students respectively.

7. Research Question 1: Teacher's Perspectives on Current ELPA21 Score Reports and Potential Diagnostic Reports

7.1 Methods

Before an attempt is made to implement a DCM scoring approach to the ELPA21, it is important to start by investigating educators' perspectives about diagnostic score reports in general. As pointed out by Goodman and Huff (2007), educators themselves have little stake in *which* scoring approaches or models are behind the information they receive. What educators need are for assessments to provide information that is instructionally relevant and aligned enough with their classroom practices to be of value. In other words, a key issue to be investigated is whether educators feel that ELPA21 diagnostic score reports have the potential to address these needs. If educators' feel diagnostic information has potential to be useful for these purposes, it would support an assumption that providing such information would lead to it being used in instructional practice. If however if the potential usefulness was felt to be low, especially in comparison to the information that educators already receive, the opposite assumption must be made that educators are unlikely to use diagnostic information without some kind of intervention to change such beliefs.

An online survey for teachers was developed to investigate this issue (Appendix A). The survey was designed so that it could be completed in 20 minutes or less. After a few background questions, the survey proceeds to a series of questions asking about teachers' opinions for the current ELPA21 score report. This information provides important context and backstory to how teachers might feel about receiving a diagnostic score report in place of or in addition to the reports they are already accustomed to seeing. After being asked the degree to which the current report is useful in informing their instruction, a sample score report similar to current score report is shown as a reference, and teachers are asked to comment specifically on the understandability and usefulness of

the three sections that make up the report: the overall proficiency status of the student, the student's performance on the test subsections, and performance summaries of the school, district, and state. An open-ended question gives teachers the opportunity to clarify any of their responses or comment further on the current score report format.

The second section of the survey investigates teachers opinions about the potential usefulness of being provided with diagnostic information about students' reading proficiency. One challenge is that teachers might not be familiar beforehand with diagnostic score reports and how they typically report information. To help teachers imagine how such a report is likely to look, they first see a sample diagnostic report like that shown in an earlier chapter (Figure 5), showing a student's hypothetical diagnostic performance on a science skills test. Accompanying the sample report is a short explanation that diagnostic score reports typically show information about student's strengths and weaknesses in particular skills, rather than an overall or composite evaluation of their overall performance on a test or performance by test section.

As the ELPA21 was not intentionally designed as a diagnostic assessment, a second challenge is the lack of a definitive list of skills that the test measures, and therefore no single answer to the question of what specific skills an ELPA21 diagnostic report should show. State proficiency standards are a readily available source that comes close to providing an answer to this question, as ELPA21 items are theoretically aligned to them. State standards have the added benefit of being aligned to state curriculums as well, and therefore have a presumed connection to instructional practice. However, one complication is that there is no single set of standards at the state level. Instead, multiple sets of state standards define the language learning expectations of ELs in slightly different, but often overlapping ways. In theory, any of these sets of standards could be used as a basis for designing an ELPA21 diagnostic report, but it is likely teachers' opinions would be greatly influenced by the specific skills a report included. Therefore, rather than make this decision

independently from teachers, it was decided that the survey should show a variety of possible sets of standards to teachers, and have them identify the ones with the most potential to provide useful diagnostic information.

Teachers were presented with three sets of reading skills and asked to rate each set for the appropriateness and usefulness of the diagnostic information it would provide if they were to appear on a diagnostic score report. Each set represents a different set of state ELA or ELP standards (Table 7). Sets were selected to present teachers with variation across sets in the relative level of detail, total number of skills, and whether they are grade level and domain specific. For two of the sets that are not specific to the domain of reading (Sets 1 and 2), some skills that are not relevant to reading were not shown to teachers. A catalogue of all skill sets as they appeared to teachers on the survey is included in Appendix B.

Table 7.

Characteristics of the Reading Standard Sets Provided to Teachers

Skill set		Level of detail	Number of skills	Grade level specific	Domain specific
1	The ELA Practices[1]	Low	4	N	N
2	The ELP Standards[2]	Medium	5	Y	N
3	The ELA Standards for Reading[3]	High	9-10	Y	Y

[1] Council of Chief State School Officers (CCSSO). (2012). Framework for English language proficiency development standards corresponding to the Common Core State Standards and the Next Generation Science Standards. Washington, DC

[2] Council of Chief State School Officers (CCSSO). (2014). English Language Proficiency Standards with Correspondences to the K-12 Practices and Common Core State Standards. Washington, DC

[3] Common Core State Standards Initiative. (2010). Common core state standards for English language arts & literacy in history/social studies, science, and technical subjects. Common Core State Standards Initiative.

Distribution of the survey was done online over the period of June to September 2019. For logistical simplicity and to minimize the inconvenience caused by the survey to the teacher population, distribution was limited to the state of Oregon, one of eight member states of the ELPA21 consortium and the current acting lead state agency. Although this was a convenience sample, Oregon's EL demographics resemble those of the nation as a whole. In the 2016-17 school year ELs made up approximately 11 percent of Oregon's K-12 student population, or 60,676 students (Sugarman & Geary, 2018). Among reported home languages among ELs, Spanish was the most common (75.3%), followed by Russian (3.2%), Vietnamese (2.5%), Arabic (2.0%), and Chinese (1.9%). As is the case in many states, a high degree of variation in EL representation can be seen at the district level, with the EL share by district ranging from as low as 8.5 percent to as high as 34.7 percent.

The Oregon Department of Education provided email addresses of the Title III and District Test Coordinators who are currently serving as ELPA21's primary contacts in each of the state's educational districts. An email was sent to 388 of these contacts, and contained an explanation of the study, an information sheet describing the survey, and an internet link for taking the survey. Coordinators were requested to forward the email to K-12 public school teachers in their district who are likely to be familiar with both the ELPA21 and who have received ELPA21 score reports in the past. This group would include teachers in English Language Development (ELD) programs who teach ELs, content area teachers who teach former ELs, and teachers in immersion or bilingual education programs. The email and information sheet made it clear that participation in the survey was completely voluntary and responses would be anonymous.

7.2 Findings

Of 388 survey invitations distributed, 61 responses were collected[4]. The lower than expected response rate is perhaps due to the overall complexity of the survey, the time of year it was distributed (coinciding with summer vacation and the start of the school term), and the lack of compensation provided for participation. The background information provided by those who did respond, however, suggests a respondent population representative of a highly experienced and diverse group of EL educators and administrators (Table 8).

Table 8.

Summary of Respondent Characteristics (n=61)

Years teaching	n (%)
1-2	3 (4.9)
3-5	6 (9.8)
6-9	7 (11.5)
10+	45 (73.8)
Relationship to ELs	
Teach ESL in an ELD program	43 (60.6)
Teach content to former ELs	8 (11.3)
Teach in an immersion/bilingual program	6 (8.5)
Other	14 (19.7)
EL or multicultural program coordinator/administrator, Title III director, co-teacher	
Training in ESOL	
Yes	58 (96.7)
ESOL teaching license/certificate/endorsement/MA degree	
No	2 (3.3)

The large majority of respondents reported 10 or more years of teaching experience (73.8%), firsthand experience instructing ELs in the classroom (60.6%), and formal training in ESOL or EL instruction by means of a ESOL license, certification, or endorsement, or an MA degree in ESOL (96.7%). While limited, it could be hypothesized that this group of respondents represents a

[4] In some cases the reported *n* may be less than 61 due to items being skipped by some respondents

motivated and informed group of educators and other EL professionals whose perspectives concerning ELPA21 score reporting practices would be of value.

Figures 7 and 8 show how respondents feel about the understandability and usefulness of the three sections making up the current ELPA21 score report: the overall score, the aggregate score, and the subdomain scores.

Figure 7.

Understandability of the Score Report by Section

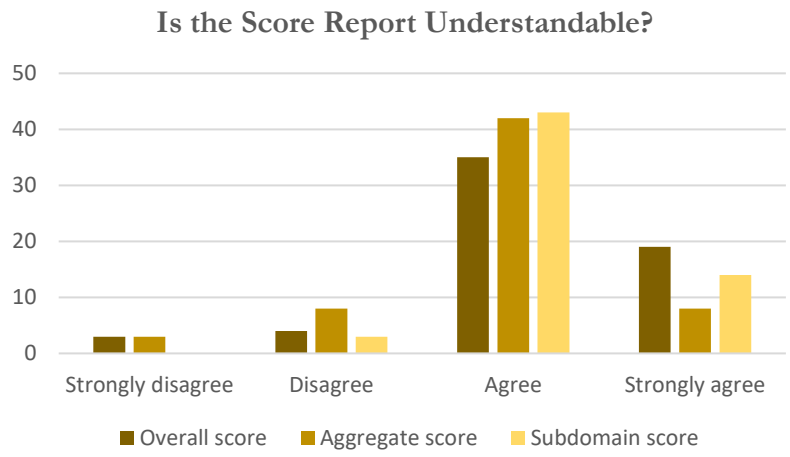
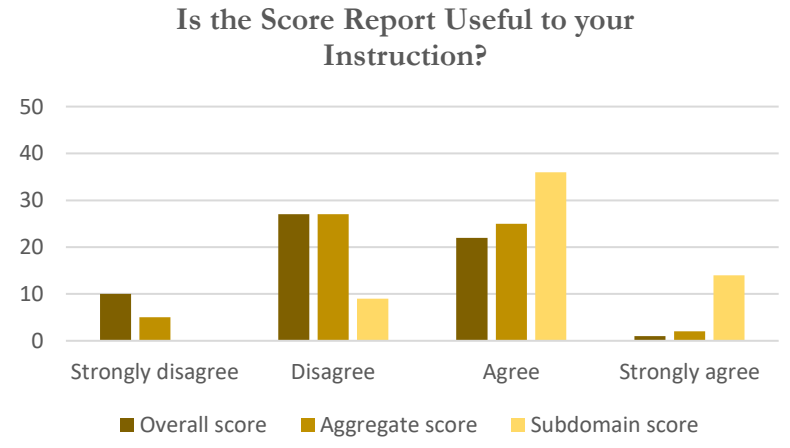


Figure 8.

Usefulness of the Score Report by Section



The understandability of the sections was rated highly, with more than 81% to 93.4% of respondents agreeing or strongly agreeing for all three sections that the scores are understandable. Respondents were noticeably more divided on the usefulness of the sections to their instruction. For the overall and aggregate score sections, more than half of respondents (60.7% and 52.5% respectively) strongly disagreed or disagreed that they were useful. Only the subdomain score section rated highly, with 85.2% of respondents in agreement or strong agreement that it was useful for their instruction.

As at this point in the survey, respondents had not yet seen a diagnostic score report nor had they been asked about the possibility of being provided with one. However, their open-ended comments about the current score report provide insight to their level of interest in what a diagnostic report could show. Over half of respondents (n=34) provided some kind of comment (Appendix C). Of these, nearly half (15) made direct reference to a dissatisfaction with the level of specificity of scores provided in the current report, calling them too broad, general, not detailed enough, or difficult to understand what it meant for individual students. A selection of some of the more revealing of these comments are listed below:

- The verbiage of Emerging, Progressing, Proficient is SO unhelpful. It does not break down our language learners into ability levels in an informative or meaningful way. Almost all students fall into Progressing, which creates incredible confusion for both ELD Specialists and core/content teachers.

- The label of Progressing is so vague and inclusive that by itself it is worthless for determining targeted lessons or grouping.

- They are not formative at all. The monikers of emerging, progressing, etc. are meaningless given how broad they are.

- Section C descriptions are nice to have especially for parents understanding where their

students are. However, after waiting all year for this report, it needs to show more specifically where students place in specific areas.

- For the report to be used in planning and instruction, it would need to include more details around the specifics of each domain score. Possibly, it could include the types of questions the student missed most often (like how the SAT is broken down by topic within each section). Otherwise, it's just a general number without contextual meaning.

- Overall score is very broad--not helpful. More specific detail on subdomain performance would be helpful. I.e. what areas in writing did students perform well/not well in.

After being shown an example of a diagnostic score report and three sets of ELP standards that could form the basis of such a score report for the ELPA21, Figure 9 and 10 show how respondents feel about the level of specificity and potential usefulness of each set.

Figure 9.

Appropriateness of Skill Specificity by Set

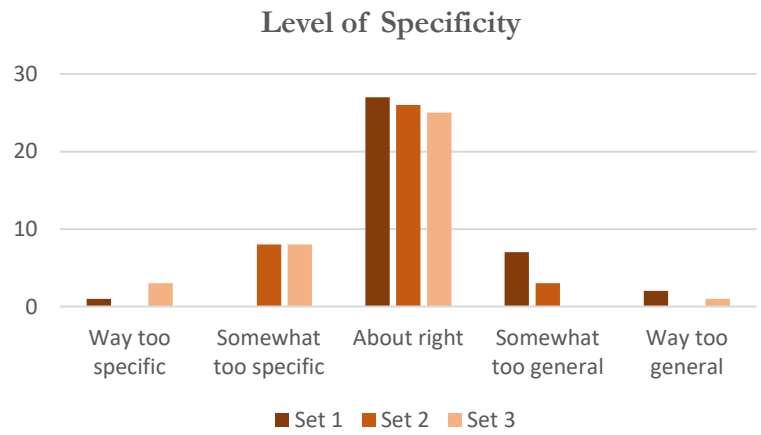
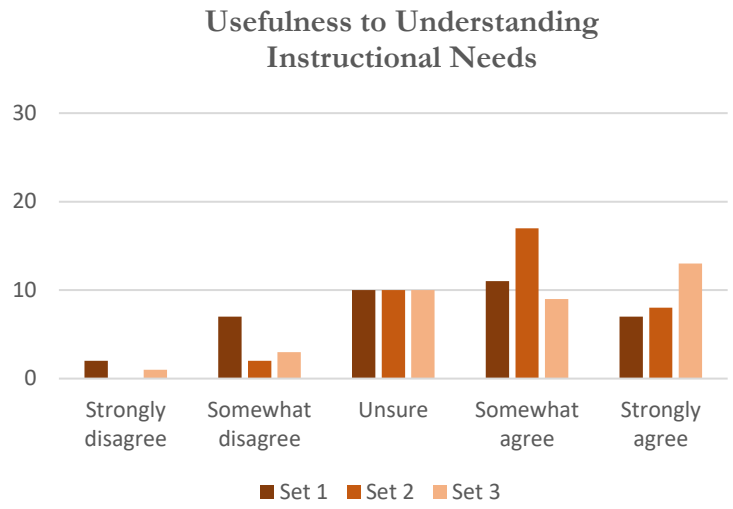


Figure 10.

Usefulness for Informing Instructional Needs by Set



Most respondents (ranging from 67.6% to 72.8%) feel that the skills are at an appropriate level of specificity regardless of set. Although there was a small trend for some respondents to feel the Set 1 skills were too general and the Set 2 and Set 3 skills to be too specific. In regard to potential usefulness, most respondents were in agreement that diagnostic scores based on the sets could be useful for informing instruction, with some respondents who were unsure, and a small group who disagreed. Interest in the skills in Sets 2 and Set 3 were the highest, with 67.6% and 61.1% of respondents somewhat agreeing or strongly agreeing that diagnostic scores would be useful. In comparison, the interest in Set 1 was less, with only 48.7% of respondents feeling the same way. Most of the differences observed across sets is due to differences between respondents themselves rather than individual respondents reacting differently to the sets. In other words, the large majority of respondents gave essentially the same ratings of specificity and usefulness to all three sets. This would suggest that rather than the sets having an effect, it is the respondents themselves that are either interested, unsure, or not interested in getting skill-based feedback regardless of what the skills

are or how specific they are. Table 9 summarizes the findings regarding the specificity and usefulness of the three skill sets.

Table 9.

Summary of Specificity and Potential Usefulness of Diagnostic Scores by Set

	Too specific	About right	Too general
Set 1	1 (2.7)	27 (72.8)	9 (24.3)
Set 2	8 (21.6)	26 (70.3)	3 (8.1)
Set 3	11 (29.7)	25 (67.6)	1 (2.7)
Total	20 (18.0)	78 (70.3)	13 (11.7)
	Not Useful	**Unsure**	**Useful**
Set 1	9 (24.3)	10 (27.0)	18 (48.7)
Set 2	2 (5.4)	10 (27.0)	25 (67.6)
Set 3	4 (11.1)	10 (27.8)	22 (61.1)
Total	15 (13.6)	30 (27.3)	65 (59.1)

8. Research Question 2: Aligning Diagnostic Feedback with State Learning Standards

8.1 Methods

A preliminary step to aligning ELPA21 diagnostic feedback to state learning standards was to have teachers judge which set of reading standards from Table 4 had the most promise for generating useful diagnostic feedback on the survey, as described in the previous section of this report. The next step was to specify the attributes of a *Q*-matrix for three forms of the ELPA21 reading test, Kindergarten, Grade 4-5, and Grade 9-12, in a way that would reflect these standards. Recall that the *Q*-matrix of a DCM is essentially just a two-dimensional matrix showing the skills measured by a test (called 'attributes' in DCM literature) as columns, and individual test items as

rows. Adopting this layout lends itself to a simple shortcut whereby the standards with the most promise can simply be reinterpreted as the attributes of a *Q*-matrix.

However, a critical step in this part of the analysis occurs after attributes that will appear in a *Q*-matrix have been defined, namely the filling out of the matrix itself. This was accomplished by reviewing test items one at a time and determining which attribute (or attributes) they measured. This step was carried out by the author, an experienced ESL teacher who has worked extensively on several standardized language assessments, including the ELPA21. To document this process and make it as systematic as possible, a protocol was followed whereby each test item was reviewed and its type, grade level, task features, and what action or actions students needed to take to arrive at the correct answer was catalogued. Once the catalogue had been completed, similar items were grouped together, and a collective determination for each group was made for which attributes could reasonably be associated with those items.

An additional obstacle in defining *Q*-matrices as described above is that since in the case of ELPA21, attributes and *Q*-matrix are being defined *ex post facto* well after test development, therefore it is possible that the best *Q*-matrix based on item content may not necessarily be viable for purposes of running a DCM model. In cases where a test is planned and developed from the onset with the intention of being a diagnostic assessment, the test's *Q*-matrix attributes are known ahead of time and items can be strategically written so that the test as a whole meets the properties described in Table 5. As such, particularly great care must be taken when designing the retrofitted *Q*-matrices to maximize, to the extent possible, how closely each matrix adheres to properties that would allow the DCM to accurately estimate diagnostic scores.

Given the possibility of encountering difficulties achieving this however, some modification to the attributes might be necessary if issues are found to be occurring to a degree that would likely complicate running the DCM. These modifications could involve changes like combining multiple

attributes into a single one if they appeared frequently on similar test items, or dropping attributes if they appeared on nearly every item in the test. As doing so would represent not only a shift away from the state standards, but from the indicated preferences of teachers, a cautious approach was taken where such changes were avoided if at all possible, but if they must be done then care was taken to document and address them in the discussion section below.

8.2 Findings

This part of the study involved the selection of attributes and defining of three separate *Q*-matrices corresponding to three grade level forms of the ELPA21. Using the feedback provided by Table 9, it was decided to use a combination of the skills represented by Set 2 and Set 3, as the skills in these sets were often rated as being appropriate in specificity and as having the highest potential usefulness by educators. As these two sets alone present a high number of potential skills, far more than the 10 recommended as a maximum for a *Q*-matrix, it was decided that an initial review of the skills within these two sets would be carried out first to eliminate any skills with a high degree of overlap in content, as well as identify only those skills that could be related to the content of items appearing on each test form. While the skills in Set 1 were not completely unpopular, as they were neither grade specific, and thought to be far too general to hold promise for alignment to specific reading test items, they were not used as attributes in the current study. Although a similar procedure was followed in the case of each of the three *Q*-matrices developed, unique issues were encountered for each form, and therefore development of each *Q*-matrix will be discussed separately.

In the case of the Kindergarten ELP standards, a total of 12 were deemed suitable for use as *Q*-matrix attributes, two from Set 2 and 10 from Set 3 (Table 10). In general the description of each attribute stays similar to how they are worded in the standards, however in some cases they have

been simplified in Table 10 to remove references to aspects of reading ability that may go beyond what ELPA21 test items measured.

Table 10.

Proposed Attributes for Kindergarten Reading Test

#	Source	Description of attribute
1	ELP K.1.L1	can identify a few key words
2	ELP K.1.L2	can identify some key words and phrases
3	ELA K.1	can answer questions about key details in a text
4	ELA K.2	can retell familiar stories including key details (Literature)
5	ELA K.3	can identify characters, settings, and major events in a story (Literature)
6	ELA K.4	can answer questions about unknown words in a text (Literature)
7	ELA K.5	can describe the relationship between illustrations and the story in which they appear (Literature)
8	ELA K.6	can identify the main topic and retell key details of a text (Informational)
9	ELA K.7	can describe the connection between two individuals, events, ideas, or pieces of information in a text (Informational)
10	ELA K.8	can answer questions about unknown words in a text (Informational)
11	ELA K.9	can describe the relationship between illustrations and the text in which they appear (Informational)
12	ELA K.10	can identify the reasons an author gives to support points in a text (Informational)

Using this set of attributes, a catalogue was made of the Kindergarten form of the ELPA21 aligning each test item to the attribute or attributes that it could measure (Appendix D1), resulting in a proposed *Q*-matrix design (Appendix D2). While the initial set of attributes exceeded the maximum number by two attributes, this issue is not present in the proposed *Q*-matrix as it was discovered that only five attributes are measured by items appearing on the test (1, 2, 3, 8, and 9).

Note however that the *Q*-matrix is unrefined as it has not been modified to address several design violations: 1) fewer than three items measuring an attribute, 2) more than 3 attributes measured by a single item, and 3) an attribute measured by every single item on the test.

In an attempt to resolve these issues, the first modification made was to combine attribute 1 and 2. Trial DCM runs on the unrefined *Q*-matrix indicated that these two attributes were either highly correlated or not distinguishable from one another based on the current ELPA21 test. As the original attributes were intended to distinguish the ability to recognize individual words from the ability to recognize short phrases, the "new" Attribute 1 could be defined as a student having the ability "to recognize key words and phrases".

A second modification involved a further combining of attributes, with attributes 8 and 9 being combined. Attribute 8 relates to a student being able to identify the main topic of an informational text and retell key details. Attribute 9 on the other hand relates to a student being able to connect two pieces of information across an informational text. While both attributes arguably represent distinct reading abilities, only two items on the test measure Attribute 9 while only a single item measures Attribute 8. As both attributes could be considered as having the ability to answer complex comprehension questions about informative texts (going beyond simple recall of details), they were combined into a single attribute.

A third and final modification is related to the presumption of a hierarchical structure to some of the attributes in the unrefined *Q*-matrix. A close inspection of the attributes defined in Table 10, and the "new" combined attributes described above, suggests that an inherent hierarchy between attributes is possible. For example, a student could not be reasonably expected to answer a question about a key detail in a text (Attribute 3) without having first mastered the ability to identify key words and phrases (Attributes 1 & 2). Similarly, a student could not reasonably be expected to answer a complex question about a text (Attributes 8 & 9) without having first mastered Attributes 1

& 2. To complicate matters, such an attribute hierarchy also imposes a within-item hierarchy in the case of a reading comprehension item. Not only would a student be expected to have mastered Attributes 1 & 2 before mastering Attribute 3, they also likely could not answer a test item measuring Attribute 3 without also applying Attribute 1 & 2. This is because it is difficult to conceive of how a test item could be written to measure whether a student can answer a question about a key detail in a text without having them read words and phrases. Building these hierarchies into a *Q*-matrix however resulted in two major design problems, a very high number of items measuring more than 3 attributes, and no items measuring Attributes 3 or the newly combined Attributes 8 & 9 in isolation. To address the problem with this *Q*-matrix and all others in this study, the decision was made for all final *Q*-matrix designs to follow a structure whereby only the highest order attribute (highest presumed hierarchy) would be associated with an item. This would create a *Q*-matrix whereby all items are only associated with a single attribute, also known as a *factorially simple item* (FSI) structure.

To justify this decision, it was reasoned that while the attribute hierarchies described may and probably do exist in theory, given the relatively short length of the ELPA21 test forms the number of items wouldn't typically be enough to distinguish a true attribute from the attribute specific to the items a student saw, which would not necessarily be equivalent measures. In other words, a test having a couple of vocabulary items measuring a student's ability to recognize the words "bee", "plant", and "map" for example, wouldn't necessarily reveal as much about that student's *general ability to recognize words* as opposed to their ability to recognize *those three words specifically*. While knowing a student's general ability to recognize words may very well help to predict their performance at a higher order reading attribute, their ability to recognize just a handful of specific words would be expected to be far less useful. Appendix D3 shows the refined *Q*-matrix for the

Kindergarten form that incorporates the modifications described above and based on the attributes listed in Table 11.

Table 11.

Final *Q*-matrix attributes for Kindergarten Reading Test

	Description of attribute
$\boldsymbol{a_1}$	can identify key words and phrases
$\boldsymbol{a_2}$	can answer questions about key details in a text
$\boldsymbol{a_3}$	can answer questions requiring them to connect multiple pieces of information across an informative text

In Grades 4 & 5, a total of 11 ELP standards were deemed suitable for use as *Q*-matrix attributes, one from Set 2 and 10 from Set 3 (Table 12). Similar to the case with the Kindergarten form, a catalogue was made of the Grade 4 & 5 items (Appendix E1) and a *Q*-matrix design was proposed (Appendix E2).

With the Grade 4 & 5 form, a longer test combined with greater diversity in item type allowed for the inclusion of more attributes than was possible in the Kindergarten form. A total of nine of the 11 proposed attributes could be included (1, 2, 3, 4, 5, 7, 8, 9, and 10). As with the Kindergarten form however, design violations are present, including: 1) fewer than three items measuring an attribute, 2) more than 3 attributes measured by a single item, and 3) an attribute measured by every single item on the test.

As was done with the Kindergarten form, several of the originally proposed attributes were combined with other attributes in an attempt to increase the number of items per attribute to be at least three items in all cases. In the case of the Grade 4 & 5 form, a logical approach to achieving this was to combine similar reading skills across attributes for literary and informational texts. For

example, the attribute related to determining the theme or main idea of a literature text, Attribute 3, was combined with the attribute related to determine the main idea of an informational text, Attribute 7. Similarly, the attribute related to describing an element of a literary text in depth, Attribute 4, was combined with the attribute related to describing an element of an informational text in depth, Attribute 8.

Table 12.

Proposed Attributes for Grade 4 & 5 Reading Test

	Source	Description of attribute
1	ELP 4-5.1.L1	can identify a few key words and phrases
2	ELA 4-5.1	can refer to details and examples in a text when explaining what the text says explicitly and when drawing inferences
3	ELA 4-5.2	can determine the theme of or summarize a story from details in a text (Literature)
4	ELA 4-5.3	can describe in depth a character, setting, or event in a story (Literature)
5	ELA 4-5.4	can describe the meaning of words and phrases as they are used in a text (Literature)
6	ELA 4-5.5	can make connections between the text of a story and a visual or oral presentation of the text (Literature)
7	ELA 4-5.6	can determine the main idea of or summarize a text and explain how it is supported by key details (Informational)
8	ELA 4-5.7	can explain events, procedures, ideas, or concepts in a historical, scientific, or technical text, including what happened and why based on specific information (Informational)
9	ELA 4-5.8	can determine the meaning of general academic and domain-specific words or phrases in a text (Informational)
10	ELA 4-5.9	can interpret information presented visually, orally, or quantitatively (e.g., in charts or graphs) and explain how the information contributes to an understanding of the text (Informational)
11	ELA 4-5.10	can explain how an author uses reasons and evidence to support particular points in a text (Informational)

Finally, the attribute related to understanding new vocabulary or phrases in context in a literary text, Attribute 5, was combined with the analogous attribute for informational texts, Attribute 9.

This approach resolved all issues with low item counts per attribute except for one, Attribute 10, interpreting information visually in a graph. There was only one such item on the test, and no items measuring its analogue for literature, Attribute 6. In this case, rather than remove the item completely from the *Q*-matrix, it was decided to run the item as is and evaluate its reliability in the next phase of analysis. The final modification made to the unrefined *Q*-matrix echoes that of the Kindergarten *Q*-matrix as well, and an FSI structure was adopted whereas all items were solely associated with their highest order attribute. Appendix E3 shows the refined *Q*-matrix for the Grade 4 & 5 form, based on the attributes listed in Table 13.

Table 13.

Final *Q*-matrix attributes for Grade 4 & 5 Reading Test

	Description of attribute
$\boldsymbol{a_1}$	can identify key words and phrases
$\boldsymbol{a_2}$	can refer to details and examples in a text when explaining what the text says explicitly and when drawing inferences
$\boldsymbol{a_3}$	can determine the theme or main idea of a literary or informational text
$\boldsymbol{a_4}$	can describe an element of a literary text (a character, setting, or event) or informational text (an event, procedure, idea, or concept) in depth
$\boldsymbol{a_5}$	can describe or determine the meaning of words or phrases in a literary or informational text
$\boldsymbol{a_6}$	can interpret information presented visually, orally, or quantitatively (e.g., in charts or graphs) and explain how the information contributes to an understanding of an informational text

With Grades 9-12, a total of 10 ELP standards were deemed suitable for use as *Q*-matrix attributes. Due to high overlap between the standards in Set 2 and Set 3, only standards from Set 3 were used (Table 14). For Set 3, two grade specific sets of standards were applicable, those for Grades 9 & 10 and those for Grades 11 & 12, however in the case of every standard except 1

(Standard 9), the Grade 9 & 10 standards were found to be a better match for the test items. A catalogue was made of the Grade 9-12 items (Appendix F1) and a *Q*-matrix design was proposed (Appendix F2). A total of nine of the 10 proposed attributes were possible in the Grade 9-12 form (1, 2, 3, 4, 6, 7, 8, 9, and 10), with the only attribute missing being Attribute 5.

Table 14.

Proposed Attributes for Grade 9-12 Reading Test

	Source	Description of attribute
1	ELP 9-10.1	can cite textual evidence to support analysis of what the text says explicitly as well as draw inferences
2	ELP 9-10.2	can determine the theme or central idea of a text and analyze in detail its development over the course of the text (Literature)
3	ELP 9-10.3	can analyze how complex characters develop over the course of the text, interact, and advance the plot or theme (Literature)
4	ELP 9-10.4	can determine the meaning of words and phrases as they are used in a text (Literature)
5	ELP 9-10.5	can analyze the representation of a subject or key scene in a text in two different artistic mediums (Literature)
6	ELP 9-10.6	can determine the central idea of a text and analyze its development over the course of a text (Informational)
7	ELP 9-10.7	can analyze how the author unfolds an analysis or series of ideas or events (Informational)
8	ELP 9-10.8	can determine the meaning of words and phrases as they are used in a text (Informational)
9	ELP 11-12.8	can integrate and evaluate multiple sources of information presented in different media or formats (e.g. visually, quantitatively) (Informational)
10	ELP 9-10.10	can delineate and evaluate the argument and specific claims in a text, assessing whether the reasoning sound and the evidence relevant and sufficient (Informational)

As with the both previous forms however, design violations are present, including: 1) fewer than three items measuring an attribute, 2) more than 3 attributes measured by a single item, and 3) an attribute measured by every single item on the test.

As was the case with previous forms, attributes were combined and an FSI structure was used to modify the *Q*-matrix to address design violations. Attributes 7 and 10 were combined into a single attribute, as it was felt they both dealt with an ability related to analyzing or evaluating how an analysis or argument is made over the course of an informational text.

In the case of the remaining two attributes with low item issues, Attribute 2 and Attribute 3, it was decided that there were no viable options for combining them with other attributes that would not result in significant negative impact on other attributes. In the case of Attribute 2, related to the determining of the theme or central idea of a literary text, a combination with the analogous attribute for informational texts (Attribute 6) is possible, as was done with the Grade 4 & 5 form. Attribute 6 however already exceeds the minimum item requirement by itself, and combination with Attribute 2 would result in the attribute losing its ability to provide diagnostic information about students' abilities specific to informational texts. A similar situation existed for Attribute 3, whereby combination with Attribute 7 would result in loss of the latter's ability to provide feedback specific to informational texts.

Given the potential value that this sort of text-specific feedback might have on a diagnostic report, and the current ability of the test to provide this feedback for several literary-text attributes and all informational-text attributes, it was decided not to sacrifice this feature and to include Attribute 2 and 3 in the model as they are despite their low item counts. Appendix F3 shows the refined *Q*-matrix for the Grade 9-12 form, based on the attributes listed in Table 15.

Table 15.

Final *Q*-matrix attributes for Grade 9-12 Reading Test

	Description of attribute
$\boldsymbol{a_1}$	can cite textual evidence to support analysis of what the text says explicitly as well as draw inferences
$\boldsymbol{a_2}$	can determine the theme or central idea of a text and analyze in detail its development over the course of the text (Literature)
$\boldsymbol{a_3}$	can analyze how complex characters develop over the course of the text, interact, and advance the plot or theme (Literature)
$\boldsymbol{a_4}$	can determine the meaning of words and phrases as they are used in a text (Literature)
$\boldsymbol{a_5}$	can determine the central idea of a text and analyze its development over the course of a text (Informational)
$\boldsymbol{a_6}$	can determine the meaning of words and phrases as they are used in a text (Informational)
$\boldsymbol{a_7}$	can integrate and evaluate multiple sources of information presented in different media or formats (e.g. visually, quantitatively) (Informational)
$\boldsymbol{a_8}$	can analyze or evaluate how an author unfolds an analysis or argument over the course of a text (Informational)

9. Research Question 3: Investigating DCM Fit and the Reliability of Diagnostic Scores

9.1 Methods

To investigate the capacity for a DCM to capture diagnostic information from the ELPA21 reading test, DCMs were retrofitted to ELPA21 reading test forms from three grade levels: Kindergarten, Grade 4 & 5, and Grade 9-12. The capacity of the DCM at each grade level was evaluated using the fit indices shown in Table 16, categorized by model fit, item fit, and person fit.

Model fit indices evaluate how well the overall model explains the observed pattern of scores, including the attributes selected and the *Q*-matrix design. As model fit indices are context-dependent, and largely uninformative in isolation, in order to evaluate them they need to be contrasted with indices obtained from comparison models.

Table 16.

Summary of Evaluated DCM Fit Indices

Model Fit	AIC, BIC
Item Fit	RMSEA
Person Fit	ρ (Probability of responding aberrantly)

These models typically use the same or a similar set of data as the model of interest, but under different model assumptions. For the current study, a "baseline" model was generated for each ELPA21 grade form by simply having the input reading text for each item represent its own attribute (Appendix G). For example, if a set of three items were all based on the same short reading passage, they were associated with the same attribute. These "text type" Q-matrices differ in design from the standards-based Q-matrices obtained from the previous section in that they completely ignore item content and the ELP standards, but have the benefit of being structurally very simple. They do not have any design violations and therefore require no modification before running. These matrices are intended to serve as a baseline reference to which the model fit indices of the hopefully more meaningful standards-based Q-matrices can be compared to. If there were no difference in fit between the two models, or if the text type Q-matrix performed better, this would be an indication that using a DCM to report diagnostic information may not be any more diagnostically informative than reporting students reading abilities based on the type of reading task they tended to get correct.

Item and person fit evaluate misfit of individual items and examinees. Generalizable scales can be used to approximate the quality of fit of individual items using the RMSEA index, and can help identify specific items on a test that do not seem to fit the model being applied to the overall test form. Person fit was estimated using the ρ index, or the probability of a person having an aberrant response pattern (Liu, Douglas, & Henson, 2009). Typically this index is used to identify students

whose test score patterns do not reflect the attribute abilities in the expected ways. This can happen for example in cases where a student might become bored with a test and start selecting answer choices randomly, have problems specific to a particular section of the test, such as a group of technology-enhanced items, or in cases of cheating. While the person fit index is not necessarily an indication of a problem with either the DCM or the test form, high numbers of misfitting persons may be a way of diagnosing a possible problem with the Q-matrix design.

Also of interest is the question of the reliability of diagnostic information that can be achieved through a DCM scoring approach. There is a question as to what reliability standard the diagnostic classifications have to meet in order to be useful to teachers in the classroom, however, there can be no doubt that highly unreliable classifications are unlikely to be of any practical value for instructional purposes. In order to investigate this question of reliability, the attribute properties obtained through the DCM will be thoroughly evaluated in the case of each ELPA21 test form. One source of evidence for attribute reliability when using DCMs are the classification probabilities generated for students on each attribute. Probabilities close to 50% represent the poorest classification reliability, as all that can be said about a student with such a score is they are equally likely to be a master as a non-master. In contrast, ranges of less than 40% or greater 60% have been used as a benchmark for reliable classification. A second source of evidence for reliability is the reliability index provided by Templin & Bradshaw (2013), a classical test theory-like reliability index calculated by measuring the stability of attribute classifications across a simulated a test-retest. The reliability of the attribute classifications obtained by the DCM at each of the three grade levels will be investigated and compared.

9.2 Findings

The model, item, and person fit statistics for the DCMs run on each ELPA21 test form are shown in Table 17, along with the fit statistics of the baseline models for comparison. With the model fit statistics, lower AIC and BIC values indicate better fit. In the case of the Kindergarten and Grade 4 & 5 test forms, the DCM was a better fit and outperformed the baseline model. In the case of Grade 9-12 however, the baseline model outperformed the DCM.

Table 17.

Fit Statistics for DCMs Retrofit to the ELPA21 Reading Test

	Kindergarten		Grade 4 & 5		Grade 9-12	
	DCM	**Baseline**	**DCM**	**Baseline**	**DCM**	**Baseline**
Model Fit						
AIC	217768	217968	416312	425628	807854	806428
BIC	218153	218381	416902	426219	808712	807262
Item Fit (RMSEA)						
Poor > 0.1	n = 0	-	n = 1	-	n = 0	-
OK 0.05 – 0.1	n = 5	-	n = 7	-	n = 11	-
Good < 0.05	n = 18	-	n = 18	-	n = 25	-
Average	0.031	-	0.049	-	0.042	-
Person Fit (p<.05)						
Spurious high	1032 (9.9)	-	1382 (9.7)	-	2165 (11.1)	-
Spurious low	1553 (14.9)	-	2064 (14.4)	-	2667 (13.7)	-

In terms of item fit, lower values of RMSEA also indicate better fit. Across all forms there was only one item on the Grade 4 & 5 test from that would be considered a poorly fitting item, and the majority of items could be considered as having good fit. However, there was a high proportion of items with only average fit, especially in the Grade 4 & 5 and Grade 9-12 test forms.

The person fit indices suggest that between 9.7% and 11.1% of students would be flagged as spurious high scorers (answering more items correctly than the model suggests) and between 13.7% and 14.9% of students would be flagged as spurious low scorers (answering fewer items correctly

than the model suggests). Unfortunately, no metric for comparing these percentages to in the literature could be found, but they do fall within the ranges of those reported in other language assessments (Liu, Douglas, & Henson, 2009). A summary of the major attribute properties for the DCMs for each test form are shown in Table 18.

Table 18.

Attribute Mastery and Correlations by Test Form

Kindergarten	**Mastery (%)**	**A1^{K}**	**A2^{K}**	**A3^{K}**
Attribute 1	80.7	1.000	-	-
Attribute 2	47.7	0.634	1.000	-
Attribute 3	49.3	0.648	0.999	1.000

Grade 4 & 5		**A1^{G45}**	**A2^{G45}**	**A3^{G45}**	**A4^{G45}**	**A5^{G45}**	**A6^{G45}**
Attribute 1	71.1	1.000	-	-	-	-	-
Attribute 2	26.3	0.725	1.000	-	-	-	-
Attribute 3	66.7	0.936	0.777	1.000	-	-	-
Attribute 4	62.1	0.931	0.815	0.984	1.000	-	-
Attribute 5	56.6	0.968	0.758	0.956	0.965	1.000	-
Attribute 6	62.1	0.779	0.620	0.764	0.832	0.783	1.000

Grade 9-12		**A1^{G912}**	**A2^{G912}**	**A3^{G912}**	**A4^{G912}**	**A5^{G912}**	**A6^{G912}**	**A7^{G912}**	**A8^{G912}**
Attribute 1	44.7	1.000	-	-	-	-	-	-	-
Attribute 2	37.7	0.631	1.000	-	-	-	-	-	-
Attribute 3	33.7	0.973	0.716	1.000	-	-	-	-	-
Attribute 4	46.8	0.945	0.685	0.941	1.000	-	-	-	-
Attribute 5	47.9	0.965	0.646	0.979	0.940	1.000	-	-	-
Attribute 6	55.7	0.966	0.623	0.882	0.993	0.972	1.000	-	-
Attribute 7	48.9	0.956	0.614	0.978	0.952	0.969	0.964	1.000	-
Attribute 8	31.0	0.996	0.674	0.954	0.944	0.990	0.949	0.950	1.000

The mastery probability indicates the overall proportion of students who were classified as masters of each attribute. These ranged from as low as 26.3%, indicating only around 1 in 4 students were classified as masters for the attribute, to as high as 80.7%. Generally speaking the attributes were harder to master for the higher grade level forms, indicating that the tests in general tended to increase in difficulty.

The correlation tables show the tetrachoric correlations between pairs of attributes (Templin & Henson, 2006). Generally speaking, the expectation is for these to be rather high, as the various reading attributes are likely to be associated with one another. The correlations ranged from 0.614 to 0.999. A case of a pair of attributes having an extremely high correlation that approaches 1.0 in value is of concern as it would suggest that the DCM may not able to distinguish the attributes based on the test items. In other words, nearly all students would either be masters of both attributes or non-masters, with few cases where one attribute had been mastered but not the other.

The reliability estimates for each attribute are shown in Table 19.

Table 19.

Attribute Reliability by Test Form

	# of Items	High Certainty (%)	α
Kindergarten			
Attribute 1	13	96.0	0.971
Attribute 2	7	83.4	0.815
Attribute 3	3	82.4	0.818
Grade 4 & 5			
Attribute 1	6	92.6	0.943
Attribute 2	5	99.6	0.997
Attribute 3	4	91.9	0.952
Attribute 4	7	93.6	0.972
Attribute 5	5	91.1	0.933
Attribute 6	1	90.8	0.906
Grade 9-12			
Attribute 1	4	91.2	0.960
Attribute 2	1	72.2	0.601
Attribute 3	2	88.4	0.864
Attribute 4	6	92.1	0.941
Attribute 5	4	90.9	0.931
Attribute 6	9	94.3	0.967
Attribute 7	6	92.6	0.967
Attribute 8	4	90.9	0.947

The leftmost column indicates the total number of items measuring each attribute. Reliability would be expected to increase with a greater number of items. The middle column shows the percentage of total classifications considered to be high certainty classifications, a greater than 60% chance of being either a master or non-master. A high percentage would indicate a greater degree of certainty in the attribute classifications, and these values range from 72.2% to 99.6%. The rightmost column shows the Templin & Bradshaw reliability index. Generally speaking the agreement between the two reliability indices is quite high, with a correlation of 0.967.

There also appears to be an overall trend for attributes measured with more items to have higher reliability, however there are notable exceptions. Attribute 2 for example on the Kindergarten test form is measured by seven items but has a lower than expected reliability. Attribute 6 on the Grade 4 & 5 test form is only measured by 1 item but has a relatively high reliability. This contrasts quite starkly with Attribute 2 on the Grade 9-12 test form, which is also measured by a single item but has a reliability that is much lower.

10. Discussion

10.1 General Discussion

The final sections of this report will attempt to tie these findings and their takeaway messages into a cohesive narrative, one that hopefully could serve as a guide for future work in the area of applying DCMs to large-scale assessments. This chapter is divided into three sections, each devoted to a discussion of the findings related to a specific research question. As a general overview, the points that will be covered are as follows:

- EL Educators felt that parts of the current score report are useful for informing their instruction, but are interested in receiving the kind of finer-grain feedback that a DCM could provide.
- Educators' opinions about the usefulness of the feedback do not change based on the specificity or scope it would be provided at. 'More information is better' seems to be the general mindset.
- Design constraints on the ELPA21 summative forms substantially limit how many items can appear that measure some ELP Standards, and therefore the number of standards that can be reflected in a score report is also limited relative to the total number.
- In order to meet *Q*-matrix requirements, some ELP Standards may need to be merged with others, but often these combinations result in sensible attributes that are still interpretable.
- A DCM approach should allow for score reporting at a finer-grain size than in the current score report.
- When sound *Q*-matrix design is followed, most attributes can be reported at high reliability levels.

10.2 Findings Related to Educator Perspectives

The survey instrument was intended to provide insight into the question of how useful and valuable fine-grain feedback obtained with a DCM might be to educators of EL students, particularly in relation to the current score reports. An interesting pattern in the findings was that the majority of teachers reported that the subdomain scores, the part of the report that presently provides the most detailed information about students' language abilities, are useful for informing their instruction. This is an interesting finding, and a key missing piece of information lies in

identifying how and where teachers are incorporating these subdomain scores into their instruction. Such information could reveal a great deal about how to effectively provide diagnostic scores to have similar instructional tractability. However, this pattern should not necessarily be taken as an endorsement of the currently provided subdomain scores as suitable feedback for instruction. It may simply be a reflection of the subdomain scores being the highest grain-size of information that teachers are currently being provided. Indeed, the fact that the majority of teachers are interested in the potential of even finer-grain feedback, shown in Table 9 and especially their open-ended comments (Appendix C), are strong indicators that suggest that many teachers are dissatisfied enough with the current score report to have a strong interest in the kind of feedback that a DCM could provide as a better way for the ELPA21 to provide information that is relevant to the instructional context.

However, also suggested by the survey is that such a broad generalization may not fully capture the complexity of what is happening with educators' actual feelings about this feedback. The high degree of consistency seen in teachers' feelings regardless of the set of skills they were shown suggests that rather than an over focusing on the form that the feedback should take, as was the purpose of the survey, it may be more worthwhile at this stage to identify those teachers who are most likely to be receptive to feedback that is intended to be useful to them in the first place. It seems like one group of teachers surveyed recognizes a need for the ELPA21 score report to include more detailed information about students, without placing much restrictions on what the scores should look like as long as the change is an improvement. A second group of teachers does not feel there is a need for this information, regardless of what form it comes takes, while a third group is unsure about whether this information is needed. The question of whether providing teachers with fine-grain feedback that would be useful may depend on their membership in one of these subgroups, potential adopters, non-adopters, and undecided, and perhaps the more critical task for

researchers is to identify where the potential adopters are, and what personal or school-based factors contribute to group membership in the first place. Perhaps the non-adopter group represent teachers with significant antagonistic feelings towards large-scale testing in general, and are unlikely to be receptive to feedback no matter what form it takes. As such, rather than a design-first and implement-everywhere approach to implementing a diagnostic feedback system, where the format of feedback is optimized first before rolling out at a large scale, such as a school district or state, a better approach might be to implement-locally and design-second. In the latter case, feedback would be introduced first, perhaps in a rough and prototype form, but within a contained setting such as a classroom or individual school, where most or all teachers could be confirmed ahead of time as potential adopters. The design of the feedback could then be optimized locally by this highly motivated group of users in the context that they will use it.

10.3 Findings Related to Alignment with ELP Standards

The aim of the item catalogues shown in Appendices D1, E1, and F1, and the attribute lists and *Q*-matrices based on them, was to ask the question of how possible it would be to align ELPA21 reading test items to attributes that would likely be relevant to the instructional content of EL teachers, in this case adhering as closely to the ELP standards as possible. There is certainly some positivity to be taken from the findings in this regard. Providing Kindergarten teachers with diagnostic feedback based on even the limited set of three attributes that made up the final DCM model for the Kindergarten ELPA21 form (Table 11), would almost certainly be an improvement over receiving feedback in the form a single reading subdomain score or proficiency level. For instance, teachers could at least be given some indication whether their students were in need of instruction at recognizing isolated words and phrases, or were ready to be reading short texts for basic or complex reading comprehension skills. The feedback looks even more promising at the

higher grade levels, where the tests at Grade 4 & 5 and Grade 9-12 allow for reading ability to be reported across an even larger numbers of skills. However, simply doing better than the current subdomain scores shouldn't be seen as that much of an accomplishment, and there was little reason to doubt that DCMs would not be able to achieve at least some form of improvement. Echoing a point made earlier in this report, it is important to use this as an opportunity to focus on the relationship between ELP test content and the models' capacity to provide this feedback. Revealed in the course of this investigation was that a number of constraints on the ELPA21 reading test directly impact the quality of the feedback that can be provided.

For instance, over the course of developing the item catalogues, it became clear that there are some differences between test forms in how well the ELPA21 items provide coverage of the standards. Both the Grade 4 & 5 and Grade 9-12 test forms show generally good coverage of their respective standards, with only one or two standards not being covered by any items on the test. The most concerning case is with the Kindergarten test, which measures fewer than half of the relevant standards. This can clearly be seen in the obvious mismatch between the rich variety of reading skills defined by the standards in Table 10, and the content of the actual test (Table 11). While it is a possibility that this is the result of a justifiable decision to adjust the content of the test in light of what may be unreasonably high language expectations in the Kindergarten standards, and not an omission by test developers, this mismatch should nonetheless be a topic of concern.

Even when a standard can be shown to be covered by a test form however, in order to contribute to generating finer-grain feedback using a DCM it must be covered by enough items to satisfy the model. As timed, secure, and state-administered assessments, there are a high number of constraints on how the ELPA21 can be administered and an understandably vested interest in making the test as short as possible. Four of these constraints in particular will be addressed here:

- Certain attributes can only be measured once per reading text

- Only a single example of each text type appears on each form
- Overrepresentation of word recognition, vocabulary, and explicit information questions
- Underrepresentation of higher order attributes for literary texts

A consistent finding of the analysis was that certain attributes constitute "expensive" items to administer on the ELPA21 reading test, in the sense that only a single item measuring the attribute can be given per reading text or visual input, such as a graph. As these texts and inputs often require a large amount of time for the student to read, even a single one of these items represents a considerable investment of testing time. Examples of expensive items include items that ask students to identify the main idea of a text or interpret the information that graph shows. While the impact of having these items on a test can be mitigated somewhat by having multiple items associated with each text or input, this does not change the fact that only one main idea item per text can be asked. In the cases of the Kindergarten and Grade 4 & 5 forms, this limitation had the effect of restricting the number of these items to such a degree that it required attributes to be combined across different text types (literature and informational) and text lengths (short and extended).

The ELPA21 reading test employs several different categories of reading texts, including correspondences, literature, informational texts, and argument & support essays, and in the case of literature and informational texts, both short and extended versions. Assuming that there is a need to report diagnostic feedback for at least some of these text categories separately, as the standards would suggest, a further complication is caused by the fact that only a single example of each text category appears on any given test form. This has the effect of limiting, severely at times, the capacity for a diagnostic model to report attributes in this way. Having separate attributes for each text category was not possible in the cases of Kindergarten and Grade 4 & 5, where attributes across test categories had to be combined. Only in the case with Grade 9-12 did the length of the test

permit for some attributes to be reported for specific categories of text, but even so, the attributes for argument & support essays needed to be combined with those for informational texts. It is unknown how modifying the attributes in the ways described in this and the preceding paragraph impact the value of the diagnostic feedback provided (for example reporting a student's ability to identify the main idea in texts in general as opposed to their ability to identify the main idea of a literature text and an informational text separately.)

A final point of concern is that of a lack of a balanced representation of items that measure certain attributes on the test forms, particularly an overrepresentation of word recognition and basic reading comprehension skills, and an underrepresentation of complex reading skills and attributes related to literary texts. In the Kindergarten form, an breakdown of item type shows that 13 items measure word recognition, 7 measure basic reading comprehension, and only 3 measure complex reading skills. This problem is further exacerbated when recognizing that the complex reading skills attribute for this form actually represents a combination of 8 separate attributes. While better balance is achieved with the Grade 4 & 5 form, the Grade 9-12 form also shows a bias towards vocabulary in context items. Out of a test consisting of 36 items total, 15 measure the vocabulary in context attribute. Furthermore, while the test has 23 items measuring attributes specific to informational texts, there are only 8 items measuring attributes specific to literary texts.

The findings suggest that while the ELPA21 reading test forms lend themselves to Q-matrices based on the ELP standards in some regards, at times extensive modification of the attributes are required to meet the standards necessary for running DCMs. It is less clear why the test content does not conform more closely to the standards that it claims to draw inspiration from, however a reasonable guess would be the prioritization of certain item types over others for their discrimination and reliability properties in unidimensional scoring models. Another could be differences in the difficulty of writing items of one type over another. In any case, it is clear that

these summative assessments are in and of themselves not capable of providing fine-grain feedback that paints a complete picture of a students' proficiency as it relates to the ELP standards. Educators will likely need to supplement the information that the summative assessment can provide with formative assessments, or other ELP assessment instruments such as interim assessments that would be less bound by the constraints discussed earlier.

10.4 Findings Related to Reliability

Using the best *Q*-matrix available, model fit, attribute characteristics, and reliability indices were calculated for each test form to evaluate whether the retrofitted DCMs would be capable of producing attribute classifications that were sufficiently consistent. In the case of the Kindergarten and Grade 4 & 5 test forms, the *Q*-matrix had a better fit than the baseline, suggesting that the attributes are capable of explaining students' test score patterns at least better than a model that only considered the types of texts they were reading. In the case of Grade 9-12, the *Q*-matrix was not able to achieve a better fit than the baseline. Likely this is due to the model struggling to explain Attribute 2, which was only measured by a single item on the test and had the lowest reported reliability of any attribute in the study. While eliminating that attribute from the model completely would likely improve model fit, it was decided to keep this attribute in for purposes of this study to observe its properties. Item fit statistics were largely acceptable, with no items across any of the text forms having poor fit, although the proportion of items with RMSEA values over 0.05 is of some concern, especially with the upper test forms. While dropping of items based on fit is an option in retrofitted DCMs in operational contexts (Liu, Huggins-Manley, Bulut, 2018), since no items would be considered poorly fitting, all items were kept for subsequent analysis. Person fit statistics indicate that roughly 9.9% to 11.1% of students responded correctly more than the model would expect, while 13.7% to 14.9% responded incorrectly more than would be expected. These percentages are

close to ranges found in the literature (Liu, Douglas, & Henson, 2009) and the number of potential sources for these patterns in the testing context of the ELPA21 should be taken into consideration, including: having trouble with the testing computer, not immediately taking the test seriously, giving up early, and getting bored with the test. In future research these statistics may be useful for identifying students or contexts where these behaviors are occurring, but for purposes of this analysis person fit was deemed appropriate.

Turning to the attribute mastery proportions, the patterns generally reflect what would be expected aside from a few exceptions. The attributes generally appear harder to master as the grade level increases, and lower numbered (associated with less complex abilities in theory) attributes tend to have slightly higher mastery proportions. Two attributes stand out however, Attribute 2 on Grade 4 & 5 and Attribute 1 on Grade 9-12 have low mastery proportions considering what those attributes are, the relatively lower order reading ability to interpret or infer key details from a text. An inspection of the item catalogues shows that the items associated with these attributes tended to have low rates of correct responses, lower than those of supposedly higher order reading skills like connecting ideas within a text, or analyzing complex characters in a story or facts in an informational text. One explanation for this pattern might be that as the length and complexity of language in the reading texts increases by grade level, this attribute changes from being one of relatively low complexity to one of high complexity. In other words, students may have greater difficulty recalling or recognizing details from reading texts when they are longer and more dense with information. Another explanation is that text developers might inadvertently be writing these items to be more difficult than necessary at the higher grade levels to compensate for them typically being "easier" items, or inversely, writing complex reading items at lower difficulty levels.

The attribute correlations and reliability indices are perhaps the most useful information in the findings regarding which attributes the DCMs are capable of providing diagnostic feedback on, aside

from the item catalogues. It is promising to see the correlations between attributes are generally large and positive, as we would expect reading attributes to be associated with one another. However, some of the largest correlations, especially those in the range of 0.950+ or higher, might suggest that the attributes measure the same reading ability, or, that the test did not have sufficient items to distinguish the abilities from each other. Attributes 2 and Attributes 3 on the Kindergarten form are candidates for being combined into a single attribute for diagnostic reporting, given their high correlation, somewhat low reliability, and sensible grouping into a single "answering questions about key details and information in a text" attribute. Several attributes in the Grade 9-12 test form might be candidates for combining based on their correlations as well, such as Attributes 4 and Attribute 6, which could be combined into a "determining the meaning of words and phrases in context of text" attribute. While Attribute 1 and Attribute 8 share a high correlation, a sensible definition of a shared attribute is not immediately clear. In general, the attribute reliabilities are quite high. Almost every attribute can be reported with a reliability higher than 0.90 (using the Templin & Bradshaw 2013 index) or more than 90% high certainty classification. The exceptions are Attributes 2 and 3 in Kindergarten, which are high candidates for combining, and Attributes 2 & 3 in Grade 9-12, which were suspected to likely have reliability issues due to the low number of items measuring them. In an operational context, it would be recommended than both these attributes should be dropped in the reporting of diagnostic scores. The findings discussed thus far suggest that if a protocol such as that described in this study is followed, including careful selection of attributes and mapping to items, *Q*-matrix optimization, and reliability analysis, sufficient reliability can be achieved for the reporting of diagnostic feedback for the ELPA21 reading test at greater detail than the current reading subdomain score.

11. Conclusion

The findings presented in the preceding chapter shed some light into the research questions proposed in the current study. They paint a picture of the potential for using DCMs in the context of the ELPA21 and other ELP assessments that certainly shows some promise. EL educators are struggling to see the current score reports as interpretable or instructionally useful, and many are receptive to alternative approaches to getting feedback about their students. Their survey comments confirm that for many educators, this feedback should take the form of fine-grained, skill-based information about what students strengths and weaknesses are. In addition, there are clear pathways between the ELPA21 reading items and the ELP Standards, and there is little doubt that creatively applied DCMs could be used to generate finer-grain feedback about students' reading abilities than is currently being done. Furthermore, the feedback provided was found to have surprisingly high reliability given that it was derived from a single administration of the summative ELPA21 reading test.

Taking a moment to step back, we can now contemplate the broader question of whether a DCM should be used with the ELP assessments to provide EL educators with fine-grain score reports, and if so, what is to be gained? A starting place is to explore how a score report generated using a DCM would be different when compared to the current ELPA21 score reports provided to educators. Table 20 compares the attribute profiles obtained by the DCM, with the Reading Level (1 to 5) that was reported to educators for all Grade 4-5 students analyzed in the study. An expected pattern is generally followed in Table 20. Students with more attribute masteries are getting classified into higher reading levels. However, the high degree of variation in profiles within reading levels is particularly noteworthy, especially so at Level 3. Of students currently receiving a Level 3 score in reading, all possible attribute profiles are represented at high frequencies. What this would suggest is

that there is diversity in the reading abilities of these students, and therefore their instructional needs, which the current reported reading level does not capture.

Table 20.

DCM Obtained Attribute Profiles Compared to Reported Reading Level (Grade 4 & 5)

No. of Attributes Mastered	**0**	**1**	**1**	**2**	**2**	**4**	**4**	**5**	**6**
Profile[5]	**000000**	**100000**	**000001**	**101000**	**100001**	**101110**	**101101**	**101111**	**111111**
Level 1	1043	16	345	0	12	0	0	0	0
Level 2	810	88	324	31	87	3	5	5	2
Level 3	779	218	433	176	338	461	305	1784	441
Level 4	8	4	7	6	10	196	61	1402	1025
Level 5	1	0	0	0	1	52	5	908	1899

For example, one of the 779 students with a profile of 000000 (indicating they are non-masters of all six reading attributes) would likely need significantly different instruction to enable them to progress to Level 4 compared with one of the 1784 students with a profile of 101111 (indicating they are masters of all reading attributes except for attribute 2.

Currently however, an educator of those students could expect to get the exact same score report for both students. Providing EL educators with these attribute profiles in addition to or in lieu of the reading levels might therefore support educators in providing necessary targeted language instruction in the classroom. Tables similar to Table 11 for the other grade levels investigated in this study appear in Appendix H. All show a similar trend for a high degree of diversity in ability in Level 3 students.

[5] Note that not every possible profile appears here. Profiles with less than 100 students in them were omitted from this Table.

The fine-grained feedback provided by a DCM is likely to be more interpretable and informative to educators in the classroom context. However, as discussed in Chapter 2, the summative ELPA21 on its own paints a partial picture of the total set of academic reading skills and abilities required for EL students to possess. The current study was found to support this argument. Table 21 shows 11 attributes thought to define the Grade 4 and 5 academic reading construct based on the ELP standards, and their level of coverage on the summative ELPA21 as determined by three factors: whether they are represented on the test, whether there was a sufficient number of items to measure them in isolation, and whether the scores were sufficiently reliable (higher than 0.9).

Table 21.

Coverage of Reading Attributes on the Summative ELPA21 Reading

		Is it represented?	**Is it isolated?**	**Is it reliable?**
1	ELP 4-5.1.L1	✓	✓	✓
2	ELA 4-5.1	✓	✓	✓
3	ELA 4-5.2	✓		✓
4	ELA 4-5.3	✓		✓
5	ELA 4-5.4	✓		✓
6	ELA 4-5.5			
7	ELA 4-5.6	✓		✓
8	ELA 4-5.7	✓		✓
9	ELA 4-5.8	✓		✓
10	ELA 4-5.9	✓		✓
11	ELA 4-5.10			

Noteworthy is that only two attributes satisfy all three conditions, the ability to identify key words and phrases, and the ability to refer to details and examples in a text when making explanations or drawing inferences. Seven other attributes are technically represented on the test, but appear on items with such low frequency that they needed to be combined with similar attributes to satisfy the conditions of the DCM. Two attributes are not currently represented on the assessment at all.

The point to make with Table 21 is not that the summative ELPA21 is failing to meet the needs of educators. The assessment should not be judged by the criteria to which we would expect from a fully diagnostic assessment designed for that purpose. Rather, the takeaway from the current study is that a DCM shows that the ELPA21 is in fact capable of providing high quality, fine-grained feedback on at least some skills, and the potential for likely feedback for other skills that is at least of some usefulness.

Additional assessments will be necessary to fill in the remaining gaps, but this is in fact the whole point of having a fully balanced assessment system, where multiple assessments work in tandem to provide coverage of stakeholders' needs. The benefit provided by the DCM and the "enhanced" score reports is not only that they allow the ELPA21 to generate feedback that is in a shared format to what would be used by assessments in the classroom context, but also that because of this there can be transparency for "seeing" where the gaps in information are. Educators seeing such score reports would have a greater sense of what parts of their students' abilities they have information on because of the summative ELPA21, what areas they should target with their own local assessments to get more information. The application of a DCM hopefully helps the ELPA21 to unlock some of its untapped potential to serve as a platform of collaboration linking the macro- context (i.e. the federal and state level) with the local classroom context.

Also evident from Table 21 is that there might be further untapped potential for providing additional fine-grain information in the ELPA21 item pool themselves, as many attributes are

represented by items but prevented from measuring them in isolation due to constraints unique to the summative assessment context.

Therefore, providing information about students' abilities in these attributes is well within the realm of possibility for the ELPA21 without necessarily requiring any new item development, and could be achieved with a redeployment of these items onto assessment instruments that would be free of the limitations of the summative assessment. For example, a test form drawing upon a much larger item pool than is currently possible with the summative assessment, or smaller, focused interim assessment drawing upon multiple items from a single attribute. The development and piloting of such assessments should be a priority, as analysis of them in a similar fashion to that done in the current study would clarify more about what attributes the ELPA21 item pool is measuring, and what further information it is capable of providing that is yet unrealized.

12. Future Research

There is reason for some optimism for the prospect of DCMs in the ELP assessment context in light of the benefits discussed in the preceding chapter. However, the study also makes it clear that retrofitting of DCM models is not an uncomplicated procedure. There are a number of cautionary flags raised in this study in regard to this class of models and these particular assessments that requires further attention.

A significant limitation of retrofitting of DCMs is the inherent subjectivity of the *Q*-matrix design process (Alderson, 2010). Decisions around defining attributes and aligning them to specific test items are an *ex post facto* process, and while care can be taken to do this in a process-based and analytical manner, are still open to interpretation and likely will be open to legitimate challenge. As an example of this, the *Q*-matrices used in this study have significant differences when compared to the item-to-standard alignments made by the original ELPA21 developers. These differences include

both the standards used and the degree of coverage of them that is achieved by the test. While a difference in making this comparison is that ELPA21 developers did not have DCMs or a *Q*-matrix in mind when doing their alignment, in nonetheless is true that a different set of individuals carrying out the same methodology as was done in this study may arrive at different *Q*-matrix designs. Again however, an advantage of this dilemma is that could stimulate deeper conversations specifically about the content of tests like the ELPA21. By focusing on attributes and *Q*-matrices, which necessitate direct connections between the skills students are learning in the classroom to the test, such conversations could expand to include educators as engaged stakeholders in the test development process to a greater degree than is currently achieved. Different *Q*-matrices could be compared based both on the informational value of their attributes, as well as empirically on their statistical qualities, making item-to-standard alignment less of a formality in the test development process and more of a comprehensive and contestable validity exercise. In short, the specific *Q*-matrices used in this study are fallible, but this is the point. They are only a launching point for what should be an ongoing debate about what skills the ELP assessments measure and how well they do it, and the argument is that doing this under a DCM framework represents an improvement to the test development process.

Another priority topic for future research related to generating this kind of fine-grain feedback from a summative assessment are questions related to the high correlations found in this study between attributes (see Table 18), and what they may imply about the dimensionality of test items. On one hand, the observed reliability found in this study is quite high considering the relatively few items used to measure individual attributes (4-13 items) compared to the dozens or more items are typically used to measure an individual attribute in diagnostic assessment.

This study focused on the reading section of the ELPA21. Further research needs to be done to investigate whether similar findings hold true for the other test sections. Fitting a similar DCM

model to the productive sections of the test would be especially complicated, as the summative ELPA21 contains only a few speaking and writing tasks that generate a limited amount of data about students abilities. In these cases it is likely that alternatives to the current test administration may have to be explored, such as additional test tasks or scoring of these tasks analytically as opposed to holistically.

An important next step is to ascertain the true usefulness of diagnostic feedback in practice, as opposed to its potential to be useful. In other words, making a strong case for the collaborative argument laid out in Figure 4 only raises the credibility that providing diagnostic feedback should result in the desired outcomes. A key next step is actually implementing a feedback system and observing it having these outcomes. If this is a next step for future research, several important takeaways can be gathered from the current study. Firstly, given that teachers seem to fall into relatively distinct groups of potential adopters and non-adopters, it is recommended that diagnostic feedback be trialed in smaller contexts with high proportions of potential adopters first. Secondly, a significant proportion of educators responded being unsure about receiving diagnostic feedback, suggesting that as a part of providing any diagnostic feedback an important consideration should be providing educators with resources and support to clarify and help them become familiar with what the information shows. Finally, despite its flaws, educators reported using the current ELPA21 score report to inform their instruction. This suggests they have adapted their practices autonomously to make the best use of the information they were being provided. Rather than dictating how diagnostic feedback should or is intended to be used therefore, it may be more beneficial to simply provide the information and explain what it shows, and allowing educators to define and refine their own systems of use around it.

Appendix A. ELPA21 Score Report Teacher Survey

The ELPA21 Evaluation Team is conducting work to understand how score reports can better support educators and English language programs at a local level. We'd very much like to hear your feedback on this topic through completing this survey.

Your responses will be treated with strict privacy and confidentiality. You will never be identified individually in any reporting of results. If you have any questions about the survey or technical issues, please contact Eric Setoguchi at esetoguchi@ucla.edu. The survey should take about 20 minutes to complete.

First, let us know a little about your background and your relation to ELPA21.

Q1 How long have you been teaching?
o 1-2 years
o 3-5 years
o 6-9 years
o 10 or more years

Q2 Which statement(s) best describes your connection to English Learners (ELs)
▢ I teach classes in an ELD program specific to ELs
▢ I teach content area classes to ELs or former ELs
▢ I teach ELs in an immersion or bilingual program
▢ Other (Please briefly describe)

Q3 Do you have formal training in ESOL or ESL/ELD? (e.g., coursework, certificate, authorization, and/or credential)
o Yes (please briefly describe your training)
o No

Q4 Which Education Service District are you primarily involved with?
o Northwest Regional
o Southern Oregon
o Willamette
o Other (or prefer not to answer)

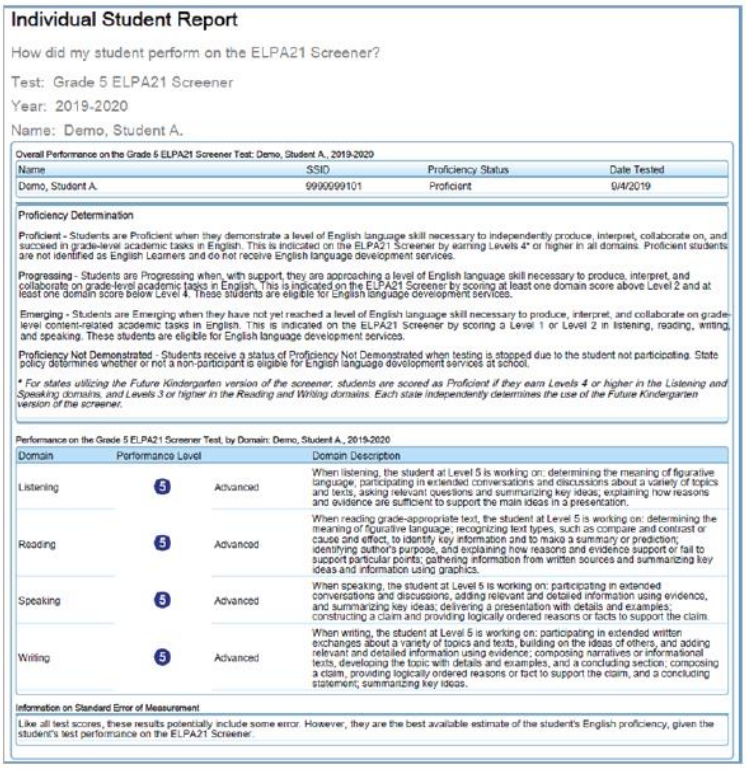
Q5 Do you believe ELPA21 does a good job of measuring student ability?
o Strongly disagree
o Somewhat disagree
o Unsure
o Somewhat agree
o Strongly agree

Q6 Do test results for individual students help inform your instruction?
o Never
o Rarely
o Sometimes
o Often

Q7 Do aggregate test results (for a class or school) help inform your instruction?
o Never
o Rarely
o Sometimes
o Often

This is a sample ELPA21 Individual Student Report, hopefully similar to ones you have seen before. The information shown includes:

A The student's individual proficiency status (Emerging, Progressing, or Proficient)
B State, District, & School proficiency summaries with proficiency level descriptions
C The student's performance on the test's subsections with written descriptions



Individual Student Report

How did my student perform on the ELPA21 Screener?

Test: Grade 5 ELPA21 Screener

Year: 2019-2020

Name: Demo, Student A.

Overall Performance on the Grade 5 ELPA21 Screener Test: Demo, Student A., 2019-2020

Name	SSID	Proficiency Status	Date Tested
Demo, Student A.	9990099101	Proficient	9/4/2019

Proficiency Determination

Proficient - Students are Proficient when they demonstrate a level of English language skill necessary to independently produce, interpret, collaborate on, and succeed in grade-level academic tasks in English. This is indicated on the ELPA21 Screener by earning Levels 4* or higher in all domains. Proficient students are not identified as English Learners and do not receive English language development services.

Progressing - Students are Progressing when, with support, they are approaching a level of English language skill necessary to produce, interpret, and collaborate on grade-level academic tasks in English. This is indicated on the ELPA21 Screener by scoring at least one domain score above Level 2 and at least one domain score below Level 4. These students are eligible for English language development services.

Emerging - Students are Emerging when they have not yet reached a level of English language skill necessary to produce, interpret, and collaborate on grade-level content-related academic tasks in English. This is indicated on the ELPA21 Screener by scoring a Level 1 or Level 2 in listening, reading, writing, and speaking. These students are eligible for English language development services.

Proficiency Not Demonstrated - Students receive a status of Proficiency Not Demonstrated when testing is stopped due to the student not participating. State policy determines whether or not a non-participant is eligible for English language development services at school.

** For states utilizing the Future Kindergarten version of the screener, students are scored as Proficient if they earn Levels 4 or higher in the Listening and Speaking domains, and Levels 3 or higher in the Reading and Writing domains. Each state independently determines the use of the Future Kindergarten version of the screener.*

Performance on the Grade 5 ELPA21 Screener Test, by Domain: Demo, Student A., 2019-2020

Domain	Performance Level		Domain Description
Listening	5	Advanced	When listening, the student at Level 5 is working on: determining the meaning of figurative language, participating in extended conversations and discussions about a variety of topics and texts, asking relevant questions and summarizing key ideas; explaining how reasons and evidence are sufficient to support the main ideas in a presentation.
Reading	5	Advanced	When reading grade-appropriate text, the student at Level 5 is working on: determining the meaning of figurative language, recognizing text types, such as compare and contrast or cause and effect, to identify key information and to make a summary or prediction; identifying author's purpose, and explaining how reasons and evidence support or fail to support particular points; gathering information from written sources and summarizing key ideas and information using graphics.
Speaking	5	Advanced	When speaking, the student at Level 5 is working on: participating in extended conversations and discussions, adding relevant and detailed information using evidence, and summarizing key ideas; delivering a presentation with details and examples; constructing a claim and providing logically ordered reasons or facts to support the claim.
Writing	5	Advanced	When writing, the student at Level 5 is working on: participating in extended written exchanges about a variety of topics and texts, building on the ideas of others, and adding relevant and detailed information using evidence; composing narratives or informational texts, developing the topic with details and examples, and a concluding section; composing a claim, providing logically ordered reasons or fact to support the claim, and a concluding statement; summarizing key ideas.

Information on Standard Error of Measurement

Like all test scores, these results potentially include some error. However, they are the best available estimate of the student's English proficiency, given the student's test performance on the ELPA21 Screener.

Q7 Based on the current report format, in your opinion...

Is the information in Part A of the report

	Strongly disagree	Disagree	Agree	Strongly agree
clearly presented in a way that is easy to understand?	o	o	o	o
useful in your curriculum planning or teaching practice?	o	o	o	o

Q8 Is the information in Part B of the report

	Strongly disagree	Disagree	Agree	Strongly agree
clearly presented in a way that is easy to understand?	o	o	o	o
useful in your curriculum planning or teaching practice?	o	o	o	o

Q9 Is the information in Part C of the report

	Strongly disagree	Disagree	Agree	Strongly agree
clearly presented in a way that is easy to understand?	o	o	o	o
useful in your curriculum planning or teaching practice?	o	o	o	o

Q10 Overall for the report, how strongly would you agree with the following statements?

	Strongly disagree	Disagree	Agree	Strongly agree
I feel confident that I am able to interpret the information in the report accurately.	o	o	o	o
I feel that the report has been designed with educators as an intended user.	o	o	o	o

Q11 Any additional feedback or comments to elaborate on these responses?

This is the final part of the survey.

The overall and subtest scores in the current report may be too broad or vague for making instructional decisions at the classroom or curriculum level. An alternative is a diagnostic report that rather than focusing on test scores, instead shows something like a profile of a student's strengths and weaknesses at specific language skills.

We'd like to collect some feedback from you about which skills you think are important and relatable to your classroom. In addition to helping in the revision of the score reports, this feedback will also be used to conduct a review of the current ELPA21 test questions.

This part of the survey is grade specific. Which grade level do you most frequently teach at?
o Kindergarten
o Grade 1
o Grade 2
o Grade 3
o Grade 4
o Grade 5
o Grade 6
o Grade 7
o Grade 8
o Grades 9 to 10
o Grades 11 to 12

You will see 3 sets of skills that have been adapted from existing proficiency standards for English Learners at the **Kindergarten** level. While the sets are similar, they have differences in which skills they include and their choice of wording. For simplicity the sets are limited to skills related to reading. [6]

Rate each set for its specificity, relevance to your classroom, and whether being provided a diagnostic report showing a student's strengths and weaknesses at the skills in a set would be useful.

Here is an example of what a diagnostic report could look like (this one is based on 3 skills related to fractions.) Proficiency levels for diagnostic reports vary, but common ones include: Master, Transitioning, and Non-Master.

Diagnostic Report for: John Doe

Fractions	
Understand two fractions as equivalent (equal) if they are the same size, or the same point on a number line.	Master
Recognize and generate simple equivalent fractions, e.g., 1/2 = 2/4, 4/6 = 2/3). Explain why the fractions are equivalent, e.g., by using a visual fraction model.	Transitioning
Express whole numbers as fractions, and recognize fractions that are equivalent to whole numbers. *Examples: Express 3 in the form 3 = 3/1; recognize that 6/1 = 6; locate 4/4 and 1 at the same point of a number line diagram.*	Non-Master

[6] The questions from this point until the end of the survey are identical for teachers at all grade levels. The only difference is the survey displays grade specific standards for Set 2 and Set 3 (as shown in Appendix B).

Set 1: The English Language Arts Practices
Council of Chief State School Officers (CCSSO). (2012). Framework for English language proficiency development standards corresponding to the Common Core State Standards and the Next Generation Science Standards. Washington, DC

A student can…

- Support analyses of a range of grade-level complex texts with evidence.
- Construct valid arguments from evidence and critique the reasoning of others.
- Build and present knowledge through research by integrating, comparing, and synthesizing ideas from texts.
- Use English structures to communicate context-specific messages.

Suppose a student's score report were to provide information about the extent to which a student can do the things listed above.

Q12 What do you think of this level of detail/specificity?
o Way too specific
o Somewhat too specific
o About the right level of specificity
o Somewhat too general
o Way too general

Q13 Would this sort of feedback provide you with information you don't already have access to (through teaching or the current score report) that would help you to better understand this student's instructional needs?
o Strongly disagree
o Somewhat disagree
o Unsure
o Somewhat agree
o Strongly agree

Set 2: The English Language Proficiency Standards: Level 4 Descriptors
Council of Chief State School Officers (CCSSO). (2014). English Language Proficiency Standards with Correspondences to the K-12 Practices and Common Core State Standards. Washington, DC

A student can…

- With prompting and support (including context and visual aids), use an increasing range of strategies to: identify main topics, answer questions about key details or parts of stories, and retell events from read-alouds, picture books, and oral presentations.
- Participate in conversations and discussions, ask and answer simple questions, and follow increasing number of rules for discussion about a variety of topics.
- With prompting and support from adults, recall information from experience or use information from provided sources to answer a question showing increasing control.
- With prompting and support, identify a reason an author or speaker gives to support a point.
- With prompting and support (including context and visual aids), answer and sometimes ask questions about the meaning of words and phrases in simple oral presentations and read-alouds about a variety of topics, experiences, or events.

Suppose a student's score report were to provide information about the extent to which a student can do the things listed above.

Q14 What do you think of this level of detail/specificity?
o Way too specific
o Somewhat too specific
o About the right level of specificity
o Somewhat too general
o Way too general

Q15 Would this sort of feedback provide you with information you don't already have access to (through teaching or the current score report) that would help you to better understand this student's instructional needs?
o Strongly disagree
o Somewhat disagree
o Unsure
o Somewhat agree
o Strongly agree

Set 3: The English Language Arts Standards: Reading for Literature & Informational Texts Common Core State Standards Initiative. (2010). Common core state standards for English language arts & literacy in history/social studies, science, and technical subjects. Common Core State Standards Initiative.

A student can…

- With prompting and support, ask and answer questions about key details in a text.

For Literature	For Informational Texts
• With prompting and support, retell familiar stories, including key details.	• With prompting and support, identify the main topic and retell key details of a text.
• With prompting and support, identify characters, settings, and major events in a story.	• With prompting and support, describe the connection between two individuals, events, ideas, or pieces of information in a text.
• Ask and answer questions about unknown words in a text.	• With prompting and support, ask and answer questions about unknown words in a text.
• With prompting and support, describe the relationship between illustrations and the story in which they appear (e.g., what moment in a story an illustration depicts.)	• With prompting and support, describe the relationship between illustrations and the text in which they appear (e.g. what person, place, thing, or idea in the text an illustration depicts).
	• With prompting and support, identify the reasons an author gives to support points in a text.

Suppose a student's score report were to provide information about the extent to which a student can do the things listed above.

Q16 What do you think of this level of detail/specificity?
o Way too specific
o Somewhat too specific
o About the right level of specificity
o Somewhat too general
o Way too general

Q17 Would this sort of feedback provide you with information you don't already have access to (through teaching or the current score report) that would help you to better understand this student's instructional needs?
o Strongly disagree
o Somewhat disagree
o Unsure
o Somewhat agree
o Strongly agree

Q18 Are there reading skills that you are especially interested in getting feedback on but didn't see in any set?

Q19 Any additional feedback for us about these sets?

Appendix B1. Set 1 The ELA Practices

A student can…

- Support analyses of a range of grade-level complex texts with evidence.
- Construct valid arguments from evidence and critique the reasoning of others.
- Build and present knowledge through research by integrating, comparing, and synthesizing ideas from texts.
- Use English structures to communicate context-specific messages.

Appendix B2. Set 2 The ELP Standards

Kindergarten

A student can…

- With prompting and support (including context and visual aids), use an increasing range of strategies to: identify main topics, answer questions about key details or parts of stories, and retell events from read-alouds, picture books, and oral presentations.
- Participate in conversations and discussions, ask and answer simple questions, and follow increasing number of rules for discussion about a variety of topics.
- With prompting and support from adults, recall information from experience or use information from provided sources to answer a question showing increasing control.
- With prompting and support, identify a reason an author or speaker gives to support a point.
- With prompting and support (including context and visual aids), answer and sometimes ask questions about the meaning of words and phrases in simple oral presentations and read-alouds about a variety of topics, experiences, or events.

Grade 1

A student can…

- Use an increasing range of strategies to: identify main topics, ask and answer questions about an increasing number of key details, and retell familiar stories or episodes of stories
- Participate in discussions, conversations, and written exchanges, follow rules for discussion, ask and answer questions, respond to the comments of others, and make comments of his or her own about a variety of topics and texts.
- With prompting and support from adults, participate in shared research projects, gather information, summarize information, and answer a question from provided sources showing increasingly independent control.
- Identify reasons an author or speaker gives to support the main point
- Using sentence context, visual aids, and some knowledge of frequently occurring root words and their inflectional forms, answer and ask questions to help determine the meaning of less common words, phrases, and simple idiomatic expressions in oral presentations and written texts about a variety of topics, experiences, or events.

Grade 2-3

A student can…

- Use an increasing range of strategies to: determine the main idea or message, identify or answer questions about some key details that support the main idea/message, and retell a variety of stories from read-alouds, written texts, and oral presentations
- Participate in discussions, conversations, and written exchanges, follow rules for discussion, ask and answer questions, build on the ideas of others, and contribute his or her own ideas about a variety of topics and texts.
- With prompting and support, carry out short individual or shared research projects, recall information from experience, gather information from multiple sources, and sort evidence into provided categories.
- Tell how one or two reasons support the specific points an author or speaker makes.
- Using context, some visual aids, reference materials, and an increasing knowledge of morphology (root words, some prefixes) determine the meaning of less-frequently occurring words and phrases and some idiomatic expressions, and (at Grade 3) some general academic and content-specific vocabulary in oral discourse, read-alouds, and written texts about a variety of topics, experiences, or events.

Grade 4-5

A student can…

- Use an increasing range of strategies to: determine the main idea or theme, explain how some key details support the main idea or theme, and summarize part of a text from read-alouds, written texts, and oral presentations
- Participate in conversations, discussions, and written exchanges, build on the ideas of others, express his or her own ideas, ask and answer relevant questions, and add relevant information and evidence about a variety of topics and texts.
- Recall information from experience, gather information from print and digital sources to answer a question, record information in organized notes, with charts, tables, or other graphics, as appropriate, and provide a list of sources.
- Describe how reasons support the specific points an author or speaker makes or fails to make.
- Using context, reference materials, and an increasing knowledge of English morphology, determine the meaning of general academic and content-specific words and phrases, and determine the meaning of a growing number of idiomatic expressions in texts about a variety of topics, experiences, or events.

Grade 6-8

A student can…

- Use an increasing range of strategies to: determine two or more central ideas or themes in oral presentations or written text, explain how the central ideas/themes are supported by specific textual details, and summarize a simple text.
- Participate in conversations, discussions, and written exchanges on a variety of topics, texts, and issues, build on the ideas of others, express his or her own ideas, ask and answer relevant questions, add relevant information and evidence, and paraphrase the key ideas expressed.
- Gather information from multiple provided print and digital sources, summarize or paraphrase observations, ideas, and information with labeled illustrations, diagrams, or other graphics, as appropriate, and cite sources.
- Analyze the argument and specific claims made in texts or speech, determine whether the evidence is sufficient to support the claims, and cite textual evidence to support the analysis.
- Using context, reference materials, and an increasing knowledge of English morphology, determine the meaning of general academic and content-specific words and phrases, and a growing number of idiomatic expressions in texts about a variety of topics, experiences, or events.

Grade 9-12

A student can…

- Use an increasing range of strategies to: determine two central ideas or themes in oral presentations and written texts, analyze the development of the themes/ideas, cite specific details and evidence from the texts to support the analysis, and summarize a simple text.
- Participate in conversations, discussions, and written exchanges on a range of topics, texts, and issues, build on the ideas of others, express his or her own ideas, support points with specific and relevant evidence, ask and answer relevant questions to clarify ideas and conclusions, and summarize the key points expressed.
- Carry out both short and more sustained research projects to answer a question, gather and synthesize information from multiple print and digital sources, use search terms effectively, evaluate the reliability of each source, integrate information into an organized oral or written report, and cite sources appropriately.
- Analyze the reasoning and use of rhetoric in persuasive texts or speeches, including documents of historical and literary significance, determine whether the evidence is sufficient to support the claim, and cite textual evidence to support the analysis.
- Using context, increasingly complex visual aids, reference materials, and an increasing knowledge of English morphology, determine the meaning of general academic and content-specific words and phrases, figurative and connotative language, and a growing number of idiomatic expressions in texts about a variety of topics, experiences, or events.

Appendix B3. Set 3 The ELA Standards for Reading

Kindergarten

A student can…

- With prompting and support, ask and answer questions about key details in a text.

For Literature	For Informational Texts
• With prompting and support, retell familiar stories, including key details.	• With prompting and support, identify the main topic and retell key details of a text.
• With prompting and support, identify characters, settings, and major events in a story.	• With prompting and support, describe the connection between two individuals, events, ideas, or pieces of information in a text.
• Ask and answer questions about unknown words in a text.	• With prompting and support, ask and answer questions about unknown words in a text.
• With prompting and support, describe the relationship between illustrations and the story in which they appear (e.g., what moment in a story an illustration depicts.)	• With prompting and support, describe the relationship between illustrations and the text in which they appear (e.g. what person, place, thing, or idea in the text an illustration depicts).
	• With prompting and support, identify the reasons an author gives to support points in a text.

Grade 1

A student can…

- Ask and answer questions about key details in a text.

For Literature	For Informational Texts
• Retell stories, including key details, and demonstrate understanding of their central message or lesson	• Identify the main topic and retell key details of a text.
• Describe characters, settings, and major events in a story, using key details.	• Describe the connection between two individuals, events, ideas, or pieces of information in a text.
• Identify words and phrases in stories or poems that suggest feelings or appeal to the senses.	• Ask and answer questions to help determine or clarify the meaning of words and phrases in a text.
	• Use the illustrations and details in a text to describe it's key ideas.
	• Identify the reasons an author gives to support points in a text.

Grade 2

A student can…

- Ask and answer such questions as *who, what, where, when, why,* and *how* to demonstrate understanding of key details in a text.

For Literature

- Recount stories, including fables and folktales from diverse cultures, and determine their central message, lesson, or moral.
- Describe how characters in a story respond to major events and challenges.
- Describe how words and phrases (e.g. regular beats, alliteration, rhymes, repeated lines) supply rhythm and meaning in a story, poem, or song.
- Use information gained from the illustrations and words in a print or digital text to demonstrate understanding of its characters, setting, or plot.

For Informational Texts

- Identify the main topic of a multiparagraph text as well as the focus of specific paragraphs within the text.
- Describe the connection between a series of historical events, scientific ideas or concepts, or steps in technical procedures in a text.
- Determine the meaning of words and phrases in a text relevant to a grade 2 topic or subject area.
- Explain how specific images (e.g. a diagram showing how a machine works) contribute to and clarify a text.
- Describe how reasons support specific points the author makes in a text.

Grade 3

A student can…

- Ask and answer such questions to demonstrate understanding of a text, referring explicitly to the text as the basis for the answers

For Literature

- Recount stories, including fables, folktales, and myths from diverse cultures; determine the central message, lesson, or moral and explain how it is conveyed through key details in the text.
- Describe characters in a story (e.g., their traits, motivations, or feelings) and explain how their actions contribute to the sequence of events.
- Describe the meaning of words and phrases as they are used in a text, distinguishing literal from nonliteral language.
- Explain how specific aspects of a text's illustrations contribute to what is conveyed by words in a story (e.g., create mood, emphasize aspects of a character or setting).

For Informational Texts

- Determine the main idea of a text; recount the key details and explain how they support the main idea.
- Describe the relationship between a series of historical events, scientific ideas or concepts, or steps in technical procedures in a text, using language that pertains to time, sequence, and cause/effect.
- Determine the meaning of general-academic and domain-specific words and phrases in a text relevant to a grade 3 topic or subject area.
- Use information gained from illustrations (e.g., maps, photographs) and the words in a text to demonstrate understanding of the text (e.g., where, when, why, and how key events occur).
- Describe the logical connection between particular sentences and paragraphs in a text (e.g., comparison, cause/effect, first/second/third in a sequence).

Grade 4

A student can…

- Refer to details and examples in a text when explaining what the text says explicitly and when drawing inferences from the text.

For Literature

- Determine the theme of a story, drama, or poem from details in the text; summarize the text.
- Describe in depth a character, setting, or event in a story or drama, drawing on specific details in the text (e.g., a character's thoughts, words, or actions).
- Describe the meaning of words and phrases as they are used in a text, including those that allude to significant characters found in mythology (e.g., Herculean).
- Make connections between the text of a story or drama and a visual or oral presentation of the text, identifying where each version reflects specific descriptions and directions in the text.

For Informational Texts

- Determine the main idea of a text and explain how it is supported by key details; summarize the text.
- Explain events, procedures, ideas, or concepts in a historical, scientific, or technical text, including what happened and why, based on specific information in the text.
- Determine the meaning of general academic and domain-specific words or phrases in a text relevant to a grade 4 topic or subject area.
- Interpret information presented visually, orally, or quantitatively (e.g., in charts, graphs, diagrams, time lines, animations, or interactive elements on Web pages) and explain how the information contributes to an understanding of the text in which it appears.
- Explain how an author uses reasons and evidence to support particular points in a text.

Grade 5

A student can…

- Quote accurately from a text when explaining what the text says explicitly and when drawing inferences from the text.

For Literature

- Determine the theme of a story, drama, or poem from details in the text, including how characters in a story or drama respond to challenges or how the speaker in a poem reflects upon a topic; summarize the text.
- Compare and contrast two or more characters, settings, or events in a story or drama, drawing on specific details in the text (e.g. how characters interact).
- Describe the meaning of words and phrases as they are used in a text, including figurative language such as metaphors and similes.
- Analyze how visual and multimedia elements contribute to the meaning, tone, or beauty of a text (e.g., graphic novel, multimedia presentation of fiction, folktale, myth, poem).

For Informational Texts

- Determine two or more main ideas of a text and explain how they are supported by key details; summarize the text.
- Explain the relationships or interactions between two or more individuals, events, ideas, or concepts in a historical, scientific, or technical text based on specific information in the text.
- Determine the meaning of general academic and domain-specific words and phrases in a text relevant to a grade 5 topic or subject area.
- Draw on information from multiple print or digital sources, demonstrating the ability to locate an answer to a question quickly or to solve a problem efficiently.
- Explain how an author uses reasons and evidence to support particular points in a text, identifying which reasons and evidence support which point(s).

Grade 6

A student can…

- Cite textual evidence to support analysis of what the text says explicitly as well as inferences drawn from the text.

For Literature

- Determine a theme or central idea of a text and how it is conveyed through particular details; provide a summary of the text distinct from personal opinions or judgements.
- Describe how a particular story's or drama's plot unfolds in a series of episodes as well as how the characters respond or change as the plot moves towards a resolution.
- Determine the meaning of words and phrases as they are used in a text, including figurative and connotative meanings; analyze the impact of a specific word choice on meaning and tone.
- Compare and contrast the experience of reading a story, drama, or poem to listening to or viewing an audio, video, or live version of the text, including contrasting what they "see" and "hear" when reading the text to what they perceive when they listen or watch.

For Informational Texts

- Determine a central idea of a text and how it is conveyed through particular details; provide a summary of the text distinct from personal opinions or judgements.
- Analyze in detail how a key individual, event, or idea is introduced, illustrated, and elaborated in a text (e.g., through examples or anecdotes).
- Determine the meaning of words and phrases as they are used in a text, including figurative, connotative, and technical meanings.
- Integrate information presented in different media or formats (e.g., visually, quantitatively) as well as in words to develop a coherent understanding of a topic or issue.
- Trace and evaluate the argument and specific claims in a text, distinguishing claims that are supported by reasons and evidence from claims that are not.

Grade 7

A student can…

- Cite several pieces of textual evidence to support analysis of what the text says explicitly as well as inferences drawn from the text.

For Literature

- Determine a theme or central idea of a text and analyze its development over the course of the text; provide an objective summary of the text.
- Analyze how particular elements of a story or drama interact (e.g., how setting shapes the characters or plot).
- Determine the meaning of words and phrases as they are used in a text, including figurative and connotative meanings; analyze the impact of rhymes and other repetitions of sounds (e.g., alliteration) on a specific verse or stanza of a poem or section of a story or drama.
- Compare and contrast a written story, drama, or poem to its audio, filmed, staged, or multimedia version, analyzing the effects of techniques unique o each medium (e.g. lighting, sound, color, or camera focus and angles in a film).

For Informational Texts

- Determine two or more central ideas in a text and analyze their development over the course of the text; provide an objective summary of the text.
- Analyze the interactions between individuals, events, and ideas in a text (e.g., how ideas influence individuals or events, or how individuals influence ideas or events).
- Determine the meaning of words and phrases as they are used in a text, including figurative, connotative, and technical meanings; analyze the impact of a specific word choice on meaning and tone.
- Compare and contrast a text to an audio, video, or multimedia version of the text, analyzing each medium's portrayal of the subject (e.g., how the delivery of a speech affects the impact of the words).
- Trace and evaluate the argument and specific claims in a text, assessing whether the reasoning is sound and the evidence is relevant and sufficient to support the claims.

Grade 8

A student can…

- Cite the textual evidence that most strongly supports an analysis of what the text says explicitly as well as inferences drawn from the text.

For Literature

- Determine a theme or central idea of a text and analyze its development over the course of the text, including its relationship to the characters, setting, and plot; provide an objective summary of the text.
- Analyze how particular lines of dialogue or incidents in a story or drama propel the action, reveal aspects of a character, or provide a decision.
- Determine the meaning of words and phrases as they are used in a text, including figurative and connotative meanings; analyze the impact of specific word choices on meaning and tone, including analogies or allusions to other texts.
- Analyze the extent to which a filmed or live production of a story or drama stays faithful to or departs from the text or script, evaluating the choices made by the director or actors.

For Informational Texts

- Determine a central idea of a text and analyze its development over the course of the text, including its relationship to supporting ideas; provide an example summary of the text.
- Analyze how a text makes connections among and distinctions between individuals, ideas, or events (e.g., through comparisons, analogies, or categories).
- Determine the meaning of words and phrases as they are used in a text, including figurative, connotative, and technical meanings; analyze the impact of a specific word choice on meaning and tone, including analogies or allusions to other texts.
- Evaluate the advantages and disadvantages of using different mediums (e.g., print or digital text, video, multimedia) to present a particular topic or idea.
- Delineate and evaluate the argument and specific claims in a text, assessing whether the reasoning is sound and the evidence is relevant and sufficient; recognize when irrelevant evidence is introduced.

Grade 9-10

A student can…

- Cite strong and thorough textual evidence to support analysis of what the text says explicitly as well as inferences drawn from the text.

For Literature

- Determine a theme or central idea of a text and analyze in detail its development over the course of the text, including how it emerges and is shaped and refined by specific details; provide an objective summary of the text.
- Analyze how complex characters (e.g. those with multiple or conflicting motivations) develop over the course of a text, interact with other characters, and advance he plot or develop the theme.
- Determine the meaning of words and phrases as they are used in a text, including figurative and connotative meanings; analyze the cumulative impact of specific word choices on meaning and tone (e.g., how the language evokes a sense of time and place; how it sets a formal or informal tone).
- Analyze the representation of a subject or a key scene in two different artistic mediums, including what is emphasized or absent in each treatment (e.g. Auden's "Musée des Beaux Arts" and Breughel's *Landscape with the Fall of Icarus*).

For Informational Texts

- Determine a central idea of a text and analyze its development over the course of the text, including how it emerges and is shaped and refined by specific details; provide an objective summary of the text.
- Analyze how the author unfolds an analysis or series of ideas or events, including the order in which the points are made, how they are introduced and developed, and the connections that are drawn between them.
- Determine the meaning of words and phrases as they are used in a text, including figurative, connotative, and technical meanings; analyze the cumulative impact of specific word choices on meaning and tone (e.g., how the language of a court opinion differs from that of a newspaper).
- Analyze various accounts of a subject told in different mediums (e.g., a person's life story in both print and multimedia), determining which details are emphasized in each account.
- Delineate and evaluate the argument and specific claims in a text, assessing whether the reasoning is sound and the evidence is relevant and sufficient; identify false statements and fallacious reasoning.

Grade 11-12

A student can…

- Cite strong and thorough textual evidence to support analysis of what the text says explicitly as well as inferences drawn from the text, including determining where the text leaves matters uncertain.

- Determine two or more themes or central ideas of a text and analyze their development over the course of the text, including how they interact and build on one another to produce a complex account; provide an objective summary of the text.

For Literature

- Analyze the impact of the author's choices regarding how to develop and relate elements of a story or drama (e.g. where a story is set, how the action is ordered, how the characters are introduced and developed).

- Determine the meaning of words and phrases as they are used in a text, including figurative and connotative meanings; analyze the impact of specific word choices on meaning and tone, including words with multiple meanings o language that is particularly fresh, engaging, or beautiful. (Include Shakespeare as well as other authors.)

- Analyze multiple interpretations of a story, drama, or poem (e.g. recorded or live production of a play or recorded novel or poetry), evaluating how each version interprets the source text, (Include at least one play by Shakespeare and one play by an American dramatist.)

For Informational Texts

- Analyze a complex set of ideas or sequence of events and explain how specific individuals, ideas, or events interact and develop over the course of the text.

- Determine the meaning of words and phrases as they are used in a text, including figurative, connotative, and technical meanings; analyze how an author uses and refines the meaning of a key term or terms over the course of a text (e.g., how Madison defines *faction* in *Federalist* No. 10).

- Integrate and evaluate multiple sources of information presented in different media or formats (e.g., visually, quantitatively) as well as words in order to address a question or solve a problem.

- Delineate and evaluate the reasoning in seminal U.S. texts, including the application of constitutional principles and use of legal reasoning (e.g. in US Supreme Court majority opinions and dissents) and the premises, purposes, and arguments in works of public advocacy (e.g., *The Federalist*, presidential addresses).

Appendix C. Open-ended Comments on the Current ELPA21 Score Report

It would be helpful to have the cut scores included on the score report so we know how close a student was to reaching the next level of proficiency.

It is difficult for parents to understand the report. A separate report designed for parents would be helpful.

I don't know of a way to improve the report's usefulness for curriculum planning or teaching practice. It is just one of many data points that teachers should be collecting. The domain descriptions are long lists, and not all items apply to all students at that achievement level.

In my very large elementary school, teachers very rarely use the report shown on this page. We have between 400 and 450 English learners each year, so they are mostly using the domain score data that I provide for them in a spreadsheet format. As the person responsible for making sure students are placed appropriately, I do feel like the domain scores are very well presented with the bold circles and colors. What I don't find helpful are the "emerging," "progressing," and "proficient" labels. We end up having to do a lot of data acrobatics to figure out what level of ESOL to put students in. I would like the state to consider going back to reporting an overall 1, 2, 3, 4, or 5.

The verbiage of Emerging, Progressing, Proficient is SO unhelpful. It does not break down our language learners into ability levels in an informative or meaningful way. Almost all students fall into Progressing, which creates incredible confusion for both ELD Specialists and core/content teachers.

Progressing is such a broad range of abilities; it makes it difficult to group students by ability level for small group instruction.

As an ELD teacher, I don't need this level of explanation. If the intended user (educator) is a classroom teacher without as much background, the extra explanation is sufficient.

Section B is the most problematic. The label of Progressing is so vague and inclusive that by itself it is worthless for determining targeted lessons or grouping. The written definition of Progressing is very confusing. "Students are progressing when, with support, they Approach a level ..... to produce, interpret ... grade-level content..." Was it written this way to avoid a lawsuit?
May I suggest something easier to understand such as - Students who are Progressing in English are not yet able to independently produce, collaborate .... grade-level content without supports.
I understand that the current definition is trying not to included the use of any negatives. However, the resulting sentence is too indirect.
Also, in working with classroom teachers, the ambiguous descriptions of whether or not a student will or will not be qualified for ELD, I predict many struggles pro and con regarding reclassification.

Section C is useful and clear. I can take those descriptors to meetings with colleagues and discuss students' progress with a clear eye to English language development.

You left the student's name in a part of the example report. Also, I have never seen anything like this form from my district. The narrative piece seems helpful, especially for explaining specific elements of growth to parents.

I feel only having 2 levels of active EL's can be confusing since progressing covers so much. We also use numbers for ELPA and we haven't come up with a way to average the scores for 1 number, so advanced kids (3's and 4's combined) with kids that scored 2's and 3's need very different things. I am glad the parents get this though! I would love for us to have a copy to put in their CUMs like other districts though. :)

I don't know how helpful the percentage is.... I am not used to seeing it this way.....
If it is possible, parent communication would be super helpful to have listed on here as well. It will save time.

I said disagree for part C just because I can never read the numbers next to their score.

I don't feel the score report always accurately reflects the success potential of the student. There are many students who we have been surprised have exited due to language gaps in grammar, sentence structure etc, and who could definitely benefit from longer support in the ELD program.

The overall proficiency determination is basically useless for us. This is because the "progressing" status is so wide as to be nearly useless, and there are many cases (ie, in database records) where this status is not broken down into domains. I would like to see the status returned to early intermediate, intermediate, early advanced as per a few years ago.

The report itself us useful. However, I do not feel that ELPA gives us an accurate picture of a student's day-to-day classroom skills, how they perform in their homeroom classrooms or in the community as a whole. I have very serious concerns that ELPA is releasing our students from ELD programs too soon and that this is the sole determination of whether or not a student exits the program. As funding gets cut, we have no way to provide support for kids who are struggling with academic language in their content subjects. It dismays me, that the professional judgement of teachers has been taken away when we are the people who know the most about our students' language skills.

I'm concerned about the ELPA21 scoring scales. The proficiency status scores students as a 1, 2, or 3 (Emergent, Progressing, Proficient), but the domain scores are levels 1-5. This is quite confusing to explain to parents. Also, the scale scores don't appear to be aligned. For example, an 8th grader who scores a 480 in Listening but scores a 490 the next year actually went down a level. This makes no sense.

They are not formative at all. The monikers of emerging, progressing, etc are meaningless given how broad they are.

I think the information is good for classroom teachers who don't understand the proficiency levels or why their particular student is not at a higher (or lower) level of proficiency.

Part A is too broad. The "progressing" rating only tells the reader that the student is past emerging but not yet proficient. Most teachers with English Language Learners could tell you that without the test.
More specific overall placement would be useful, especially when using the ELPA21 Individual Student Report with mainstream teachers, parents, and students themselves.

Section A & B are not specific enough to be useful data to inform teaching.
`Section C descriptions are nice to have especially for parents understanding where their students are. However, after waiting all year for this report, it needs to show more specifically where students place in specific areas.

I feel like the report is intended for parents to get an overall understanding of their child's level. For the report to be used in planning and instruction, it would need to include more details around the specifics of each domain score. Possibly, it could include the types of questions the student missed most often (like how the SAT is broken down by topic within each section). Otherwise, it's just a general number without contextual meaning.

I feel this form is intended more for people outside of the EL world (Classroom teachers, parents, Admin) who may have some knowledge, but rely of the given descriptors to understand where a student is at. As someone who is in the EL world, I know what each level (emerging, progressing proficient) entails in general.

I do think the scale score should be better represented. Showing the score band and where the student fits.

The report is not specific enough to pinpoint areas of growth for individual students

I don't feel this is a true assessment of student's abilities, especially in the domain of Writing. We don't teach kids to fill in the blanks of sentences or to write about unfamiliar topics. This test does both. It does not represent the way kids are taught to write. It is more like the old worksheet teaching and does not relate to the current standards of writing. I am not sure how to remedy this issue, but I do know it is not true to what a student can really do, more so for our kinder and 1st grade students.

It would be helpful to have a parent friendly report translated for their home language to send home.

Overall score is very broad--not helpful. More specific detail on subdomain performance would be helpful. I.e. what areas in writing did students perform well/not well in.

It's confusing to teachers and parents to have two different measurement scales. If the domains are scored on a scale of 1-5, then so should a child's overall proficiency. Parents wonder why a child who has scored 3's or above in all domains winds up with an overall score of 2.

I would love to see a growth chart that also contains previous years of ELPA results. I feel that this data is far more valuable when determining the student's growth rate.

I noticed that, depending on a year, my students score lower or higher than expected. I'd speculate that this is happening because different individuals grade the speaking or writing parts each year so there can be room for some subjectivity.

Regarding a previous question. I feel that the cut scores for some grades are accurate to what I see in student performance. For some grades it seems that the cut scores are passing students out of the EL program too early. I also believe that if students score all 4s that does not always indicate proficiency. Example: An EL 2nd grade student scored all 4s last year and was scored proficient. He is still in our Title reading program. That is a problem for me.

It would be helpful to know where the individual student lies in the continuum of progressing. A scale of some sort would be useful especially since the scores do not track across grade bands.

Moving to the three broad proficiency levels is not helpful at all. the progressing level is way to general. To a degree teachers and parents do not see growth over the years when the student is progressing every year. the 1 - 5 scores for language level was much more helpful like in the individual domains.

ELPA 21 seems to be so rigorous that it would even identify native English speakers at LEP.

Appendix D1. Item Catalogue for ELPA21 Kindergarten Reading Test

Item	Type	Input	Description	Points	% Correct	Attributes
1	Read and Match	1-word (written and spoken)	Student reads and hears the word and selects a picture from 3 choices	1	0.877	1
2	Read and Match	1-word (written and spoken)	Student reads and hears the word(s) and selects a picture from 3 choices	1	0.828	1
3	Read and Match	1-word (written and spoken)	Student reads and hears the word(s) and selects a picture from 3 choices	1	0.898	1
4	Read and Match	2-word phrase (written and spoken)	Student reads and hears the word(s) and selects a picture from 3 choices	1	0.900	1
5	Read and Match	2-word phrase (written and spoken)	Student reads and hears the word(s) and selects a picture from 3 choices	1	0.901	1
6	Read and Match	3-word phrase (written and spoken)	Student reads and hears the word(s) and selects a picture from 3 choices	1	0.912	1,2
7	Read and Match	2-word sentence (written and spoken)	Student reads and hears the input and selects a picture from 3 choices	1	0.767	1,2
8	Read and Match	4-word sentence (written and spoken)	Student reads and hears the input and selects a picture from 3 choices	1	0.717	1,2
9	Read and Match	5-word sentence (written and spoken)	Student reads and hears the input and selects a picture from 3 choices	1	0.852	1,2
10	Word Wall	everyday object (written and spoken)	Student reads and hears "a word" and drags the word to the correct picture from 4 choices	1	0.885	1
11	Word Wall	everyday object (written and spoken)	Student reads and hears "a word" and drags the word to the correct picture from 4 choices	1	0.921	1
12	Word Wall	everyday object (written and spoken)	Student reads and hears "a word" and drags the word to the correct picture from 4 choices	1	0.853	1
13	Word Wall	everyday object (written and spoken)	Student reads and hears "a word" and drags the word to the correct picture from 4 choices	1	0.886	1
14	Word Wall	"What are the pictures about?" (spoken)	Student hears the question and looking at 4 pictures of a common theme, and selects the theme from 3 choices	1	0.808	1,9
15	Short Correspondence	short correspondence (34 words)	Student reads and hears a short correspondence (with pictures for key vocabulary words) and hears a key detail question about it. They select the correct picture from 3 options.	1	0.695	1,2,3
16	Short Correspondence	short correspondence (34 words)	Student reads and hears a short correspondence (with pictures for key vocabulary words) and hears a key detail question about it. They select the correct answer from 3 options.	1	0.404	1,2,3
17	Short Correspondence	short correspondence (34 words)	Student reads and hears a short correspondence (with pictures for key vocabulary words) and reads and hears a logical structure question about it. They select the correct answer from 3 options.	1	0.454	1,2,8
18	Read-Along Story	short story (45 words)	Student reads and hears a short story and reads and hears a key detail question about it. They select the correct answer from 3 options.	1	0.672	1,2,3
19	Read-Along Story	short story (45 words)	Student reads and hears a short story and reads and hears a key detail question about it. They select the correct answer from 3 options.	1	0.587	1,2,3
20	Read-Along Story	short story (45 words)	Student hears a short story and hears a key detail question about it. They select the correct picture from 3 options.	1	0.480	1,2,3
21	Informational Set	short informational text (68 words)	Student hears a short informational text and hears a key detail question about it. They select the correct picture from 3 options.	1	0.607	1,2,3

22	Informational Set	short informational text (68 words)	Student reads and hears a short informational text and reads and hears an inference question about it. They select the correct answer from 3 options.	1	0.297	1,2,3
23	Informational Set	short informational text (68 words)	Student hears a short informational text and reads and hears a key detail question about it. They select the correct picture from 3 options.	1	0.775	1,2,9
24	Read and Match	1 word (written and spoken)	Student reads and hears a word and selects a picture from 3 choices that matches it	1	0.877	1,2,3

Appendix D2. Unrefined Q-matrix for Kindergarten Reading Test

	a_1	a_2	a_3	a_4	a_5	Attributes per item
Attribute	1	2	3	8	9	
Item 1	1	0	0	0	0	1
Item 2	1	0	0	0	0	1
Item 3	1	0	0	0	0	1
Item 4	1	0	0	0	0	1
Item 5	1	0	0	0	0	1
Item 6	1	1	0	0	0	2
Item 7	1	1	0	0	0	2
Item 8	1	1	0	0	0	2
Item 9	1	1	0	0	0	2
Item 10	1	0	0	0	0	1
Item 11	1	0	0	0	0	1
Item 12	1	0	0	0	0	1
Item 13	1	0	0	0	0	1
Item 14	1	0	0	0	1	2
Item 15	1	1	1	0	0	3
Item 16	1	1	1	0	0	3
Item 17	1	1	0	1	0	3
Item 18	1	1	1	0	0	3
Item 19	1	1	1	0	0	3
Item 20	1	1	1	0	0	3
Item 21	1	1	1	0	0	3
Item 22	1	1	0	0	1	3
Item 23	1	1	1	0	0	3
Items per attribute	23	13	7	1	2	

Appendix D3. Refined Q-Matrix for Kindergarten Reading Test

	a_1	a_2	a_3
Item 1	1	0	0
Item 2	1	0	0
Item 3	1	0	0
Item 4	1	0	0
Item 5	1	0	0
Item 6	1	0	0
Item 7	1	0	0
Item 8	1	0	0
Item 9	1	0	0
Item 10	1	0	0
Item 11	1	0	0
Item 12	1	0	0
Item 13	1	0	0
Item 14	0	0	1
Item 15	0	1	0
Item 16	0	1	0
Item 17	0	0	1
Item 18	0	1	0
Item 19	0	1	0
Item 20	0	1	0
Item 21	0	1	0
Item 22	0	0	1
Item 23	0	1	0

Appendix E1. Item Catalogue for ELPA21 Grade 4 & 5 Reading Test

Item	Type	Input	Description	Points	% Correct	Attributes
1	Match Picture to Word and Sentence	Picture of an everyday object	Student sees the picture and chooses the word that best matches it from 4 options	1	0.771	1
2	Match Picture to Word and Sentence	Picture of an everyday object	Student sees the picture and chooses the word that best matches it from 4 options	1	0.802	1
3	Match Picture to Word and Sentence	Picture of a geometric shape	Student sees the picture and chooses the word that best matches it from 4 options	1	0.710	1
4	Match Picture to Word and Sentence	Picture of an everyday scene	Student sees the picture and chooses the word that best matches it from 4 options	1	0.406	1
5	Match Picture to Word and Sentence	Picture of an everyday object	Student sees the picture and chooses the sentence (6-10 words) that best matches it from 4 options.	1	0.777	1
6	Match Picture to Word and Sentence	Picture of an everyday scene	Student sees the picture and chooses the sentence (6-8 words) that best matches it from 4 options.	1	0.870	1
7	Short Correspondence Set	Short correspondence (129 words)	Student reads a short correspondence and a main idea question about it and selects the sentence that answers it from 4 choices.	1	0.819	1,2,7
8	Short Correspondence Set	Short correspondence (129 words)	Student reads a short correspondence and a vocabulary question about it and selects the phrase that answers it from 4 choices.	1	0.622	1,2,9
9	Short Correspondence Set	Short correspondence (129 words)	Student reads a short correspondence and a logical structure question about it and selects the sentence that answers it from 4 choices.	1	0.392	1,2,8
10	Short Correspondence Set	Short correspondence (129 words)	Student reads a short correspondence and an inference question about it and selects the answer from 4 choices.	1	0.767	1,2
11	Short Literary Set	Short story (187 words)	Student reads a short story with a picture and completes a vocabulary question using a drop-down menu with 4 options.	1	0.715	1,2,5
12	Short Literary Set	Short story (187 words)	Student reads a short story with a picture and completes a main idea question by selecting the sentence that answers it from 4 choices.	1	0.734	1,2,3
13	Short Literary Set	Short story (187 words)	Student reads a short story with a picture and a vocabulary question about an event in the story and chooses the sentence that answers it from 4 options.	1	0.758	1,2,5
14	Short Literary Set	Short story (187 words)	Student reads a short story with a picture and an inference question and selects the answer from 4 options.	1	0.824	1,2,4
15	Short Informational Set	Short experiment description & graph (132 words)	Student reads a short description of a science experiment with a graph and chooses which sentence describes the graph from 4 options.	1	0.685	1,2,10
16	Short Informational Set	Short experiment description & graph (132 words)	Student reads a short description of a science experiment with a graph and chooses which sentence from it is incorrect from 4 options.	1	0.558	1,2,8
17a	Short Informational Set	Short experiment description & graph (132 words)	Student reads a short description of an experiment with a graph and completes a sentence describing a key detail from the experiment using 2 drop-down menus with 4 options each.	1	0.311	1,2
17b	Short Informational Set	Short experiment description & graph (132 words)	Student reads a short description of an experiment with a graph and completes a sentence describing a key detail from the experiment using 2 drop-down menus with 4 options each.	1	0.273	1,2
18	Short Informational Set	Short experiment description & graph (132 words)	Student reads a short description of a science experiment with a graph and inference question about it and selects the answer from 4 options.	1	0.812	1,2,8

19	Extended Literary Set	Long passage describing a historical event (194 words)	Student reads a long passage describing a historical event and a main idea question and chooses the sentence that answers it from 4 options.	1	0.545	1,2,3
20	Extended Literary Set	Long passage describing a historical event (194 words)	Student reads a long passage describing a historical event and a key detail question and chooses the sentence that answers it from 4 options.	1	0.157	1,2
21	Extended Literary Set	Long passage describing a historical event (194 words)	Student reads a long passage describing a historical event and a logical structure question and chooses the sentence that answers it from 4 options.	1	0.420	1,2,4
22	Extended Informational Set	Long descriptive passage (280 words)	Student reads a long passage describing and a main idea question and chooses the sentence that answers it from 4 options.	1	0.594	1,2,7
23	Extended Informational Set	Long descriptive passage (280 words)	Student reads a long passage and a vocabulary question and chooses the sentence that answers it from 4 options.	1	0.407	1,2,9
24a	Extended Informational Set	Long descriptive passage (280 words)	Student reads a long passage and chooses two sentences that are true about it from 6 options.	1	0.460	1,2
24b	Extended Informational Set	Long descriptive passage (280 words)	Student reads a long passage and chooses two sentences that are true about it from 6 options.	1	0.625	1,2
25	Extended Informational Set	Long descriptive passage (280 words)	Student reads a long passage and an inference question and chooses the sentence that answers it from 4 options.	1	0.123	1,2,8
26	Extended Informational Set	Long descriptive passage (280 words)	Student reads a long passage and a vocabulary question and chooses the word that answers it from 4 options.	1	0.293	1,2,9

Appendix E2. Unrefined Q-Matrix for Grade 4 & 5 Reading Test

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	**Attributes per item**
Attribute	1	2	3	4	5	7	8	9	10	
Item 1	1	0	0	0	0	0	0	0	0	1
Item 2	1	0	0	0	0	0	0	0	0	1
Item 3	1	0	0	0	0	0	0	0	0	1
Item 4	1	0	0	0	0	0	0	0	0	1
Item 5	1	0	0	0	0	0	0	0	0	1
Item 6	1	0	0	0	0	0	0	0	0	1
Item 7	1	1	0	0	0	1	0	0	0	3
Item 8	1	1	0	0	0	0	0	1	0	3
Item 9	1	1	0	0	0	0	1	0	0	3
Item 10	1	1	0	0	0	0	0	0	0	2
Item 11	1	1	0	0	1	0	0	0	0	3
Item 12	1	1	1	0	0	0	0	0	0	3
Item 13	1	1	0	0	1	0	0	0	0	3
Item 14	1	1	0	1	0	0	0	0	0	3
Item 15	1	1	0	0	0	0	0	0	1	3
Item 16	1	1	0	0	0	0	1	0	0	3
Item 17a	1	1	0	0	0	0	0	0	0	2
Item 17b	1	1	0	0	0	0	0	0	0	2
Item 18	1	1	0	0	0	0	1	0	0	3
Item 19	1	1	1	0	0	0	0	0	0	3
Item 20	1	1	0	0	0	0	0	0	0	2
Item 21	1	1	0	1	0	0	0	0	0	3
Item 22	1	1	0	0	0	1	0	0	0	3
Item 23	1	1	0	0	0	0	0	1	0	3
Item 24a	1	1	0	0	0	0	0	0	0	2
Item 24b	1	1	0	0	0	0	0	0	0	2
Item 25	1	1	0	0	0	0	1	0	0	3
Item 26	1	1	0	0	0	0	0	1	0	3
Items per attribute	28	22	2	2	2	2	3	1	1	

Appendix E3. Refined Q-Matrix for Grade 4 & 5 Reading Test

	Skill Type					
	a_1	a_2	a_3	a_4	a_5	a_6
Item 1	1	0	0	0	0	0
Item 2	1	0	0	0	0	0
Item 3	1	0	0	0	0	0
Item 4	1	0	0	0	0	0
Item 5	1	0	0	0	0	0
Item 6	1	0	0	0	0	0
Item 7	0	0	1	0	0	0
Item 8	0	0	0	0	1	0
Item 9	0	0	0	1	0	0
Item 10	0	0	0	1	0	0
Item 11	0	0	0	0	1	0
Item 12	0	0	1	0	0	0
Item 13	0	0	0	0	1	0
Item 14	0	0	0	1	0	0
Item 15	0	0	0	0	0	1
Item 16	0	0	0	1	0	0
Item 17a	0	1	0	0	0	0
Item 17b	0	1	0	0	0	0
Item 18	0	0	0	1	0	0
Item 19	0	0	1	0	0	0
Item 20	0	1	0	0	0	0
Item 21	0	0	0	1	0	0
Item 22	0	0	1	0	0	0
Item 23	0	0	0	0	1	0
Item 24a	0	1	0	0	0	0
Item 24b	0	1	0	0	0	0
Item 25	0	0	0	1	0	0
Item 26	0	0	0	0	1	0

Appendix F1. Item Catalogue for ELPA21 Grade 9-12 Reading Test

Item	Type	Input	Description	Points	% Correct	Attributes
1	Discrete Items	Short informational text about a type of plant (61 words)	Student reads a short informational text and answers a vocabulary question about it	1	0.694	1,8
2	Discrete Items	Short informational text about a type of writing (64 words)	Student reads a short informational text and answers a vocabulary question about it	1	0.756	1,8
3	Discrete Items	Short informational text about a writer (69 words)	Student reads a short informational text and answers a explicit/inference question about it	1	0.740	1
4	Discrete Items	Short informational text about a writer (69 words)	Student reads a short informational text and answers a vocabulary question about it	1	0.524	1,8
5	Discrete Items	Short informational text about a writer (69 words)	Student reads a short informational text and answers a vocabulary question about it	1	0.551	1,8
6	Discrete Items	Short informational text about animals (66 words)	Student reads a short informational text and answers a main idea question about it	1	0.821	1,6
7	Discrete Items	Short informational text about animals (66 words)	Student reads a short informational text and answers a vocabulary question about it	1	0.663	1,8
8	Discrete Items	Short informational text about art (51 words)	Student reads a short informational text and answers a main idea question about it	1	0.617	1,6
9	Discrete Items	Short informational text about art (51 words)	Student reads a short informational text and answers a vocabulary question about it	1	0.810	1,8
10	Short Literary Set	Short excerpt from a novel (344 words)	Student reads a short excerpt from a novel and answers a main idea question about it.	1	0.349	1,2
11	Short Literary Set	Short expert from a novel (344 words)	Student reads a short excerpt from a novel and answers a vocabulary question about it.	1	0.637	1,4
12	Short Literary Set	Short expert from a novel (344 words)	Student reads a short excerpt from a novel and answers a vocabulary question about it.	1	0.562	1,4
13	Short Literary Set	Short expert from a novel (344 words)	Student reads a short expert from a novel and answers a words and phrases question about it	1	0.655	1,4
14	Short Informational Set	Short description of an experiment with a graph (336 words)	Student reads a description of an experiment and answers a vocabulary question about it.	1	0.378	1,8
15	Short Informational Set	Short description of an experiment with a graph (336 words)	Student reads a description of an experiment and answers a graph interpretation question about it.	1	0.698	1,9
16a	Short Informational Set	Short description of an experiment with a graph (336 words)	Student reads a description of an experiment and answers a graph interpretation question about it.	1	0.590	1,9
16b	Short Informational Set	Short description of an experiment with a graph (336 words)	Student reads a description of an experiment and answers a graph interpretation question about it.	1	0.724	1,9
17	Short Informational Set	Short description of an experiment with a graph (336 words)	Student reads a description of an experiment and answers a graph interpretation question about it.	1	0.488	1,9
18	Extended Literary Set	Long excerpt from a novel (589 words)	Student reads a long expert from a novel and answers a explicit/inference question about it	1	0.501	1

19	Extended Literary Set	Long excerpt from a novel (589 words)	Student reads a long expert from a novel and answers a complex character question about it	1	0.502	1,3
20	Extended Literary Set	Long excerpt from a novel (589 words)	Student reads a long expert from a novel and answers a explicit/inference question about it	1	0.267	1
21	Extended Literary Set	Long excerpt from a novel (589 words)	Student reads a long expert from a novel and answers a vocabulary question about it	1	0.489	1,4
22	Extended Literary Set	Long excerpt from a novel (589 words)	Student reads a long expert from a novel and answers a vocabulary question about it	1	0.173	1,4
23	Extended Literary Set	Long excerpt from a novel (589 words)	Student reads a long expert from a novel and answers a vocabulary question about it	1	0.374	1,4
24	Extended Literary Set	Long excerpt from a novel (589 words)	Student reads a long expert from a novel and answers a complex character question about it	1	0.518	1,3
25	Extended Informational Set	Long description of an experiment with 4 figures (649 words)	Student reads a description of an experiment and answers a main idea question about it	1	0.485	1,6
26	Extended Informational Set	Long description of an experiment with 4 figures (649 words)	Student reads a description of an experiment and answers a graph interpretation question about it	1	0.403	1,9
27	Extended Informational Set	Long description of an experiment with 4 figures (649 words)	Student reads a description of an experiment and answers an unfolding idea question about it	1	0.333	1,7
28	Extended Informational Set	Long description of an experiment with 4 figures (649 words)	Student reads a description of an experiment and answers a graph interpretation question about it	1	0.685	1,9
29	Extended Informational Set	Long description of an experiment with 4 figures (649 words)	Student reads a description of an experiment and answers a vocabulary question about it	1	0.644	1,8
30	Argument and Support Essay Set	Argument and support essay (307 words)	Student reads an essay and answers a main idea question about it.	1	0.482	1,6
31	Argument and Support Essay Set	Argument and support essay (307 words)	Student reads an essay and answers an unfolding idea question about it.	1	0.354	1,7
32	Argument and Support Essay Set	Argument and support essay (307 words)	Student reads an essay and answers an argument evaluation question about it.	1	0.296	1,10
33	Argument and Support Essay Set	Argument and support essay (307 words)	Student reads an essay and answers an argument evaluation question about it.	1	0.227	1,10
34	Argument and Support Essay Set	Argument and support essay (307 words)	Student reads an essay and answers a vocabulary question about it.	1	0.266	1,8
35	Argument and Support Essay Set	Argument and support essay (307 words)	Student reads an essay and answers an explicit/inference question about it	1	0.522	1

Appendix F2. Unrefined Q-Matrix for Grade 9-12 Reading Test

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	Attributes per item
Attribute	1	2	3	4	6	7	8	9	10	
Item 1	1	0	0	0	0	0	1	0	0	2
Item 2	1	0	0	0	0	0	1	0	0	2
Item 3	1	0	0	0	0	0	0	0	0	1
Item 4	1	0	0	0	0	0	1	0	0	2
Item 5	1	0	0	0	0	0	1	0	0	2
Item 6	1	0	0	0	1	0	0	0	0	2
Item 7	1	0	0	0	0	0	1	0	0	2
Item 8	1	0	0	0	1	0	0	0	0	2
Item 9	1	0	0	0	0	0	1	0	0	2
Item 10	1	1	0	0	0	0	0	0	0	2
Item 11	1	0	0	1	0	0	0	0	0	2
Item 12	1	0	0	1	0	0	0	0	0	2
Item 13	1	0	0	1	0	0	0	0	0	2
Item 14	1	0	0	0	0	0	1	0	0	2
Item 15	1	0	0	0	0	0	0	1	0	2
Item 16a	1	0	0	0	0	0	0	1	0	2
Item 16b	1	0	0	0	0	0	0	1	0	2
Item 17	1	0	0	0	0	0	0	1	0	2
Item 18	1	0	0	0	0	0	0	0	0	1
Item 19	1	0	1	0	0	0	0	0	0	2
Item 20	1	0	0	0	0	0	0	0	0	1
Item 21	1	0	0	1	0	0	0	0	0	2
Item 22	1	0	0	1	0	0	0	0	0	2
Item 23	1	0	0	1	0	0	0	0	0	2
Item 24	1	0	1	0	0	0	0	0	0	2
Item 25	1	0	0	0	1	0	0	0	0	2
Item 26	1	0	0	0	0	0	0	1	0	2
Item 27	1	0	0	0	0	1	0	0	0	2
Item 28	1	0	0	0	0	0	0	1	0	2
Item 29	1	0	0	0	0	0	1	0	0	2
Item 30	1	0	0	0	1	0	0	0	0	2
Item 31	1	0	0	0	0	1	0	0	0	2
Item 32	1	0	0	0	0	0	0	0	1	2
Item 33	1	0	0	0	0	0	0	0	1	2
Item 34	1	0	0	0	0	0	1	0	0	2
Item 35	1	0	0	0	0	0	0	0	0	1
Items per attribute	36	1	2	6	4	2	9	6	2	

Appendix F3. Refined Q-Matrix Candidates for Grade 9-12 Reading Test

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
Attribute	1	2	3	4	6	8	9	7,10
Item 1	0	0	0	0	0	1	0	0
Item 2	0	0	0	0	0	1	0	0
Item 3	1	0	0	0	0	0	0	0
Item 4	0	0	0	0	0	1	0	0
Item 5	0	0	0	0	0	1	0	0
Item 6	0	0	0	0	1	0	0	0
Item 7	0	0	0	0	0	1	0	0
Item 8	0	0	0	0	1	0	0	0
Item 9	0	0	0	0	0	1	0	0
Item 10	0	1	0	0	0	0	0	0
Item 11	0	0	0	1	0	0	0	0
Item 12	0	0	0	1	0	0	0	0
Item 13	0	0	0	1	0	0	0	0
Item 14	0	0	0	0	0	1	0	0
Item 15	0	0	0	0	0	0	1	0
Item 16a	0	0	0	0	0	0	1	0
Item 16b	0	0	0	0	0	0	1	0
Item 17	0	0	0	0	0	0	1	0
Item 18	1	0	0	0	0	0	0	0
Item 19	0	0	1	0	0	0	0	0
Item 20	1	0	0	0	0	0	0	0
Item 21	0	0	0	1	0	0	0	0
Item 22	0	0	0	1	0	0	0	0
Item 23	0	0	0	1	0	0	0	0
Item 24	0	0	1	0	0	0	0	0
Item 25	0	0	0	0	1	0	0	0
Item 26	0	0	0	0	0	0	1	0
Item 27	0	0	0	0	0	0	0	1
Item 28	0	0	0	0	0	0	1	0
Item 29	0	0	0	0	0	1	0	0
Item 30	0	0	0	0	1	0	0	0
Item 31	0	0	0	0	0	0	0	1
Item 32	0	0	0	0	0	0	0	1
Item 33	0	0	0	0	0	0	0	1
Item 34	0	0	0	0	0	1	0	0
Item 35	1	0	0	0	0	0	0	0

Appendix G. Baseline (Text Type) Q-matrices for the ELPA21 Reading Test

Text Type Q-Matrix for Kindergarten Reading Test

	a_1	a_2	a_3	a_4
Item 1	1	0	0	0
Item 2	1	0	0	0
Item 3	1	0	0	0
Item 4	1	0	0	0
Item 5	1	0	0	0
Item 6	1	0	0	0
Item 7	1	0	0	0
Item 8	1	0	0	0
Item 9	1	0	0	0
Item 10	1	0	0	0
Item 11	1	0	0	0
Item 12	1	0	0	0
Item 13	1	0	0	0
Item 14	1	0	0	0
Item 15	0	1	0	0
Item 16	0	1	0	0
Item 17	0	1	0	0
Item 18	0	0	1	0
Item 19	0	0	1	0
Item 20	0	0	1	0
Item 21	0	0	0	1
Item 22	0	0	0	1
Item 23	0	0	0	1

Text Type Q-Matrix for Grade 4 & 5 Reading Test

	a_1	a_2	a_3	a_4	a_5	a_6
Item 1	1	0	0	0	0	0
Item 2	1	0	0	0	0	0
Item 3	1	0	0	0	0	0
Item 4	1	0	0	0	0	0
Item 5	1	0	0	0	0	0
Item 6	1	0	0	0	0	0
Item 7	0	1	0	0	0	0
Item 8	0	1	0	0	0	0
Item 9	0	1	0	0	0	0
Item 10	0	1	0	0	0	0
Item 11	0	0	1	0	0	0
Item 12	0	0	1	0	0	0
Item 13	0	0	1	0	0	0
Item 14	0	0	1	0	0	0
Item 15	0	0	0	1	0	0
Item 16	0	0	0	1	0	0
Item 17a	0	0	0	1	0	0
Item 17b	0	0	0	1	0	0
Item 18	0	0	0	1	0	0
Item 19	0	0	0	0	1	0
Item 20	0	0	0	0	1	0
Item 21	0	0	0	0	1	0
Item 22	0	0	0	0	0	1
Item 23	0	0	0	0	0	1
Item 24a	0	0	0	0	0	1
Item 24b	0	0	0	0	0	1
Item 25	0	0	0	0	0	1
Item 26	0	0	0	0	0	1

Text Type Q-matrix for Grade 9-12 Reading Test

	a_1	a_2	a_3	a_4	a_5	a_6
Item 1	1	0	0	0	0	0
Item 2	1	0	0	0	0	0
Item 3	1	0	0	0	0	0
Item 4	1	0	0	0	0	0
Item 5	1	0	0	0	0	0
Item 6	1	0	0	0	0	0
Item 7	1	0	0	0	0	0
Item 8	1	0	0	0	0	0
Item 9	1	0	0	0	0	0
Item 10	0	1	0	0	0	0
Item 11	0	1	0	0	0	0
Item 12	0	1	0	0	0	0
Item 13	0	1	0	0	0	0
Item 14	0	0	1	0	0	0
Item 15	0	0	1	0	0	0
Item 16a	0	0	1	0	0	0
Item 16b	0	0	1	0	0	0
Item 17	0	0	1	0	0	0
Item 18	0	0	0	1	0	0
Item 19	0	0	0	1	0	0
Item 20	0	0	0	1	0	0
Item 21	0	0	0	1	0	0
Item 22	0	0	0	1	0	0
Item 23	0	0	0	1	0	0
Item 24	0	0	0	1	0	0
Item 25	0	0	0	0	1	0
Item 26	0	0	0	0	1	0
Item 27	0	0	0	0	1	0
Item 28	0	0	0	0	1	0
Item 29	0	0	0	0	1	0
Item 30	0	0	0	0	0	1
Item 31	0	0	0	0	0	1
Item 32	0	0	0	0	0	1
Item 33	0	0	0	0	0	1
Item 34	0	0	0	0	0	1
Item 35	0	0	0	0	0	1

Appendix H. DCM Obtained Attribute Profiles Compared to Reported Reading Level

Kindergarten

No. of Attributes Mastered	0	1	2	3
Profile	**000**	**100**	**011**	**111**
Level 1	1089	390	96	158
Level 2	582	1027	54	541
Level 3	94	1911	45	2039
Level 4	2	173	2	1004
Level 5	0	33	1	1192

Grade 4 & 5

No. of Attributes Mastered	0	1	1	2	2	4	4	5	6
Profile	**000000**	**100000**	**000001**	**101000**	**100001**	**101110**	**101101**	**101111**	**111111**
Level 1	1043	16	345	0	12	0	0	0	0
Level 2	810	88	324	31	87	3	5	5	2
Level 3	779	218	433	176	338	461	305	1784	441
Level 4	8	4	7	6	10	196	61	1402	1025
Level 5	1	0	0	0	1	52	5	908	1899

Grade 9-12

No. of Attributes Mastered	0	1	2	3	3	4	4
Profile	**00000000**	**00000100**	**00010100**	**10110000**	**11100000**	**11110000**	**10001110**
Level 1	3916	1	4	1	0	0	1
Level 2	3577	117	28	17	14	8	14
Level 3	1533	134	149	168	172	211	230
Level 4	10	1	1	2	0	2	5
Level 5	1	0	0	1	0	1	1

	5	6	6	7	8
	10011110	**10111110**	**10011111**	**11111110**	**11111111**
Level 1	0	0	0	0	1
Level 2	13	10	1	8	37
Level 3	563	319	169	449	3405
Level 4	39	35	32	60	1608
Level 5	5	4	6	9	935

13. Bibliography

Act, E. S. S. (2015). of 2015, Pub. *L*, 114-95.

Alderson, J. C. (2010). "Cognitive Diagnosis and *Q*-Matrices in Language Assessment": A Commentary.

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

Au, W. (2011). Teaching under the new Taylorism: High-stakes testing and the standardization of the 21st century curriculum. *Journal of Curriculum Studies*, *43*(1), 25-45.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

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.

Bailey, A. L., & Durán, R. (2019). A Mediator of Valid Interpretations of Information Generated by Classroom Assessments among Linguistically and Culturally Diverse Students. *Classroom Assessment and Educational Measurement*, 46.

Bailey, A. L., & Heritage, M. (Eds.). (2008). *Formative assessment for literacy, grades K-6: Building reading and academic language skills across the curriculum*. Corwin Press.

Berliner, D. (2011). Rational responses to high stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education*, *41*(3), 287-302.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)*, *21*(1), 5.

Blinder, A. (2015). Atlanta educators convicted in school cheating scandal. *New York Times*.

The Center on Standards and Assessment Implementation. (2018). *Setting the Stage for Formative Assessment.* Haubner, J.P. & Chang, S. Retrieved from https://www.csai-online.org/sites/default/files/CSAIwebinar_FA_School_UserGuide_01.pdf

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple…. *Language Testing*, *29*(1), 19-27.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2011). *Building a validity argument for the Test of English as a Foreign LanguageTM*. Routledge.

Chung, M. T. (2014). Estimating the Q-matrix for cognitive diagnosis models in a Bayesian framework (Doctoral dissertation, Teachers College).

Cronbach, L. J. (1988). Internal consistency of tests: Analyses old and new. *Psychometrika*, *53*(1), 63-70.

Cohen, D. K. (1995). What is the system in systemic reform?. *Educational researcher*, *24*(9), 11-31.

Cole, N. S. (1986). Future directions for educational achievement and ability testing. In B. S. Plake & J. C. Witt (Eds.), *Buros-Nebraska symposium on measurement and testing: Vol. 2. The future of testing.* Hillsdale, NJ: Erlbaum.

Congress, U. S. (1994). Goals 2000: Educate America Act. *Public Law*, *103227*.

Council of Chief State School Officers. (2014). *English language proficiency (ELP) standards: with correspondences to K–12 English language arts (ELA), mathematics, and science practices, K–12 ELA standards, and 6–12 literacy standards.* Washington, DC: Author. Retrieved from http://www.elpa21.org/sites/default/files/Final%204_30%20ELPA21%20Standards_1.pdf

Darling-Hammond, L., Haertel, E., & Pellegrino, J. (2015). Making Good Use of New Assessments: Interpreting and Using Scores From the Smarter Balanced Assessment Consortium.

Davidson, F. (2010). Why is cognitive diagnosis necessary? A reaction.

Di Carlo, M. (2011). The Evidence on Charter Schools and Test Scores. Policy Brief. *Albert Shanker Institute*.

Elementary and Secondary Education Act (ESEA). (1965). Public Law No. 89-10.

ETS. (2018). *Understanding Balanced Assessment Systems.* Author. Retrieved from https://www.ets.org/s/k12/pdf/ets-k-12-understanding-measurement-white-paper.pdf

Felch, J., Song, J., & Smith, D. (2010). Who's teaching LA's kids? A Times analysis, using data largely ignored by LAUSD, looks at which educators help students learn, and which hold them back. *Los Angeles Times*.

Figlio, D., & Loeb, S. (2011). School accountability. In *Handbook of the Economics of Education* (Vol. 3, pp. 383-421). Elsevier.

Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment.

Gierl, M. J., & Leighton, J. P. (2007). Directions for future research in cognitive diagnostic assessment. *Cognitive diagnostic assessment for education: Theory and applications*, 341-352.

Goodman, D.P, & Huff, K. (2007). The demand for cognitive diagnostic assessment In Leighton, J., & Gierl, M. (Eds.). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Hamilton, L. S., Stecher, B. M., & Yuan, K. (2012). Standards-based accountability in the United States: Lessons learned and future directions. *Education Inquiry*, 3(2), 149-170.

Hammond, J. W., & Moss, P. A. (2016). Validity Theory in Measurement. *Encyclopedia of Educational Philosophy and Theory*, 1-5.

Hart, R., Casserly, M., Uzzell, R., Palacios, M., Corcoran, A., & Spurgeon, L. (2015). Student Testing in America's Great City Schools: An Inventory and Preliminary Analysis. *Council of the Great City Schools*.

Hauck, M. C., Pooler, E., and Anderson, D. P. (2015). *ELPA21 item development process report*. Report submitted by Educational Testing Service (ETS), May 15, 2015.

Henson, R., & Templin, J. (2008, March). *Implementation of standards setting for a geometry end-of-course exam*. Paper presented at the 2008 American Educational Research Association conference in New York, New York

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.

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: Theory and applications* (pp. 19-60). New York, NY, US: Cambridge University Press.

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL* (Doctoral dissertation, University of Illinois at Urbana-Champaign).

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.

Kane, M. T. (2006). Validation. *Educational measurement*, *4*(2), 17-64.

Kappa, P. D. Gallup.(2015). The 47th PDK/Gallup Poll of the public's attitudes toward the public schools: Testing doesn't measure up for Americans. *Phi Delta Kappan*, *97*(1).

LAUSD (2016). *Using SBA Summative Results for Long Term Planning* [PowerPoint slides]. Retrieved from https://slideplayer.com/slide/12623085/

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.

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

LeMAHIEU, P. G., & Reilly, E. C. (2004). Systems of coherence and resonance: Assessment for education and assessment of education. *YEARBOOK-NATIONAL SOCIETY FOR THE STUDY OF EDUCATION*, (2), 189-202.

Layton, L. (2013). Bush, Obama focus on standardized testing leads to "opt-out" parents" movement. *The Washington Post*.

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.

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. (2017). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement*, 0013164416685599.

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.

Messick, S. (1995). Standards of validity and the validity of standards in performance asessment. *Educational measurement: Issues and practice*, *14*(4), 5-8.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts. *education policy analysis archives*, *18*(23), n23.

No Child Left Behind Act of 2001. (2002). Pub. L. No. 107-110, 115 Stat 1425.

Onosko, J. (2011). Race to the Top leaves children and future citizens behind: The devastating effects of centralization, standardization, and high stakes accountability. *Democracy and Education*, *19*(2), 1.

Partnership for Assessment of Readiness for College and Careers. (2018, May 10). Retrieved from https://parcc-assessment.org/assessments/

Partnership for Assessment of Readiness for College and Careers. (2014). *PARCC Model Content Frameworks: A companion to the common core state systems*. Retrieved from https://parcc-assessment.org/content/uploads/2017/11/PARCC-K2-MCF-for-Mathematics-9-24-14-2.pdf

Popham, W. J. (2003). The seductive allure of data. *Educational Leadership*, *60*(5), 48-51.

Resnick, D. P. (1980). Chapter 1: Minimum Competency Testing Historically Considered. *Review of research in education*, *8*(1), 3-29.

Rentner, D. S., Kober, N., Frizzell, M., & Ferguson, M. (2016). Listen to Us: Teacher Views and Voices. *Center on Education Policy*.

Robinson, K. (2010). Changing education paradigms. *RSA Animate, The Royal Society of Arts, London, http://www. youtube. com/watch*.

Rothstein, R. (2009). Getting accountability right. *Education Week*, *28*(19), 36-26.

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

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

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

Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. *Handbook of test development*, 677-710.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5-24.

The Smarter Balanced Assessment System. (2018, April 20). Retrieved from http://www.smarterbalanced.org/assessments/

Snyder, T. D., de Brey, C., & Dillow, S. A. (2019). Digest of Education Statistics 2017, NCES 2018-070. *National Center for Education Statistics*.

Sugarman, J., & Geary, C. (2018). English Learners in Select States: Demographics, Outcomes, and State Accountability Policies. Fact Sheet. *Migration Policy Institute*.

Supovitz, J. (2009). Can high stakes testing leverage educational improvement? Prospects from the last decade of testing and accountability reform. *Journal of Educational Change*, *10*(2-3), 211-227.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal*, *41*(4), 901-926.

Templin, J. (2016). Diagnostic assessment: Methods for the reliable measurement of multidimensional abilities. *Technology and testing*, 285-304.

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., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*(2), 251-275.

Templin, J., Cohen, A., & Henson, R. (2008). Constructing tests for optimal classification in standard setting. In *annual meeting of the National Council on Measurement in Education, New York, NY*.

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

United States Department of Education. (2018, October 4). Secretary DeVos Unveils Parents' Guide to ESSA Flexibilities [Press release]. Retrieved from https://www.ed.gov/news/press-releases/secretary-devos-unveils-parents-guide-essa-flexibilities

United States Department of Education. (2017). *Revised state template for the consolidated state plan*. Washington, DC: US Department of Education. https://www2.ed.gov/admins/lead/account/stateplan17/revisedessastateplanguidance.docx.

United States Department of Education. (2016). *Examples of leveraging ESEA funds to support fewer, smarter, high-quality assessments*. Washington, DC: US Department of Education. https://www2.ed.gov/policy/elsec/leg/essa/essaassessmentdcltr1207.pdf

United States Department of Education. (2015). *Fact sheet: Testing action plan*. Washington, DC: US Department of Education. https://www.ed.gov/news/press-releases/fact-sheet-testing-action-plan.

Ussher, B. & Earl, K., (2010), 'Summative' and 'Formative': Confused by the Assessment Terms?, New Zealand Journal of Teachers' Work, Volume 7, Issue 1, 53-63.

von Davier, M. (2014). The Log-Linear Cognitive Diagnostic Model (LCDM) as a Special Case of the General Diagnostic Model (GDM). *ETS Research Report Series*, *2014*(2), 1-13.

Wilson, M., & Draney, K. (2004). Some Links Between Large-Scale and Classroom Assessments: The Case of the BEAR Assessment System. *Yearbook of the National Society for the Study of Education*, *103*(2), 132-154.

Wolf, M. K., Herman, J. L., Bachman, L. F., Bailey, A. L., & Griffin, N. (2008). Recommendations for Assessing English Language Learners: English Language Proficiency Measures and Accommodation Uses. Recommendations Report (Part 3 of 3). CRESST Report 737. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST).*

Xin, T., Xu, Z., & Tatsuoka, K. (2004). Linkage between teacher quality, student achievement, and cognitive skills: A rule-space model. *Studies in Educational Evaluation*, *30*(3), 205-223.

Xu, X., & von Davier, M. (2008). Fitting the structured general diagnostic model to NAEP data. *ETS Research Report Series*, *2008*(1), i-18.

Yen, W. M., Fitzpatrick, A. R., & Brennan, R. L. (2006). Educational measurement.

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice*, *31*(2), 21-26.

Zhao, Y. (2014). *Who's afraid of the big bad dragon?: Why China has the best (and worst) education system in the world*. John Wiley & Sons.