HOEVE, KAREN BLACKBURN. Ph.D. Building Validity Evidence for the Use of Aggregate Scores in Accountability. (2021)
Directed by Drs. Micheline Chalhoub-Deville and Robert Henson. 111 pp.

High stakes test-based accountability systems primarily rely on aggregates and derivatives of scores from tests that were originally developed to measure individual student mastery of content specifications. Current validity models do not explicitly address this use of aggregate scores to measure the performance of teachers, administrators, and schools. Empirical methodologies that allow evaluation of test-based accountability systems need to be identified and developed. One empirical method that lends itself to the comparison of individual and group-level outcomes is hierarchical generalized linear modeling (HGLM). This research explores the validation of aggregate scores used in accountability.

BUILDING VALIDITY EVIDENCE FOR THE USE OF AGGREGATE SCORES IN ACCOUNTABILITY

by

Karen Blackburn Hoeve

A Dissertation

Submitted to

the Faculty of The Graduate School at

The University of North Carolina at Greensboro

in Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

Greensboro

2021

Approved by

Dr. Micheline Chalhoub-Deville
Committee Co-Chair

Dr. Robert Henson
Committee Co-Chair

DEDICATION

Dedicated to my daughters, Anne Marie and Indy, and their father who was taken from us much too soon. Anne Marie and Indy selflessly supported me – and each other – so that I could fulfill my educational goals and dreams. Throughout her high school years, Anne Marie picked up her sister after school, cooked meals when I worked late or was in evening classes, and never failed to tell me she believed in me. Anne Marie, your perseverance, and resilience are extraordinary, and your endless love and sense of adventure are irresistible. During a pandemic, you dedicated your life in service to others as a first responder and earned your certification as a paramedic. I am endlessly proud of you. Indy always welcomed me home with a smile at the end of the day and never failed to make me feel like she was truly happy to see me. Indy, your gifts of artistic expression, quick-witted humor, emotional intelligence, and passion for equality and justice lift and inspire me. You have made me a better person. You leave an indelible and beautiful mark on the lives of everyone you meet. I am honored and blessed to be your mother.

APPROVAL PAGE

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

Committee Co-Chair

Dr. Micheline-Chalhoub-Deville

Committee Co-Chair

Dr. Robert Henson

Committee Members

Dr. Richard Luecht

Dr. Tammy Howard

October 1, 2021

Date of Acceptance by Committee

October 1, 2021

Date of Final Oral Examination

ACKNOWLEDGEMENTS

Throughout the writing of this dissertation, I have received a great deal of support and guidance. It would be impossible to acknowledge everyone.

I would first like to thank my dissertation committee members; co-chairs, Drs. Micheline Chalhoub-Deville, and Bob Henson; and committee members, Drs. Ric Luecht, and Tammy Howard. Dr. Chalhoub-Deville, your mentorship was invaluable in developing my critical thinking and research interests. Your insightful feedback challenged me and brought my work to a higher level. Your encouragement buoyed me in difficult times. Dr. Henson, your expertise was invaluable in formulating and executing the methodology. Your unlimited patience and skillful teaching built my knowledge and my confidence. Dr. Luecht, you taught me more doctoral classes than any other faculty member and you always balanced theory with practice. I am immensely thankful for your guidance. Dr. Howard, you encouraged me to pursue a PhD and gave me the wise counsel that time would pass whether I pursued the degree or not.

I would also like to acknowledge my colleagues and classmates in the ERM department. I would particularly like to single out Dr. Aileen Reid and Myrah Stockdale for their friendship and support. Aileen, thank you for your kindness and prayers when I needed them most. Myrah, thank you for your friendship and the many hours of studying together. I cannot forget to acknowledge Jeremy Acree and Dr. J.B. Weir for their collaboration in creating the Parallel IUA which was fundamental in developing the topic for this dissertation.

In addition, I would like to thank my colleagues at the American Board of Pediatrics for their support and encouragement. Finally, I could not have completed this dissertation without the support of my dear friends, Christine Black, and Kristen Trolenberg, who provided encouragement and happy distractions to rest my mind outside of work and studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Education reform is a global initiative characterized by a trend toward high stakes, test-based accountability systems that reward and sanction teachers and schools (Chalhoub-Deville, 2009, 2016, 2020; Sahlberg, 2012, 2014). Test scores are used as a metric against which to measure the success of these reforms. The claim is that "[a]ssessing schools against the common metric of standardized student test scores provides…information regarding how well schools and school districts (and potentially teachers) are doing in comparison to their peers or to outside performance standards" (Figlio & Loeb, 2011, p. 386).

In the U.S., education reform and the design of accountability systems is driven by government policies such as the No Child Left Behind Act (2002), Race to the Top (U.S. Department of Education, 2009), and the Every Student Succeeds Act (ESSA) (U.S. Department of Education, 2015). These policies mandate testing and attach consequences (e.g., rewards and sanctions) to schools and teachers based on "aggregates of test scores such as school-wide averages, percentages of students scoring above a certain level, or growth or value-added modeling results…" (Standards for Educational and Psychological Testing, AERA, APA, NCME, 2014, referred to hereafter as the *Standards*, p. 203).

Accountability systems in the U.S. are administered at the state level and state statute often regulates the definition of accountability systems (Education Commission of the States, 2018). Test scores "…are used both to measure student achievement on state educational standards and to evaluate the degree to which teachers and schools are effective in educating students" (Bandalos, Ferster, Davis & Samuelsen., 2011, p. 155). As a result, test scores that were traditionally used to make decisions about individual students (e.g., mastery, placement, promotion) are now also aggregated for use in accountability systems. Historical and

contemporary theories of validity and validation were designed with individual test scores in mind, but in accountability, these scores are aggregated to create a score or index at a school or other testing system level. These aggregated scores or indexes are then interpreted in much the same way as an individual score, but at the school level.

Sireci and Soto (2016, p. 149) assert that, "Using tests for educational accountability often entails employing the test for purposes beyond which it was originally developed. Like the originally intended purposes, using test scores for accountability purposes also requires evidence and theory to justify their use." Validity is the cornerstone for the use and interpretation of test scores. According to the *Standards* (2014), validity is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" and is "the most fundamental consideration in developing tests and evaluating tests" (p. 11).

The 2014 *Standards*

The *Standards* (2014) identify five sources of validity evidence: test content, response processes, internal structure, relations to other variables, and testing consequences. According to the *Standards* (2014), "Content-oriented evidence of validation is at the heart of the process in the educational arena known as alignment, which involves evaluating the correspondence between student learning standards and test content" (p. 15). Sireci and Faulkner-Bond (2014) describe one aspect of content validity as the "appropriateness of the test development process" which "refers to all processes used when constructing a test to ensure that test content faithfully and fully represents the construct intended to be measured and does not measure irrelevant material" (p. 101).

A method for test development that facilitates test content evidence for a validity argument is evidence-centered design (ECD). ECD takes interpretation and use claims into

account during the test development phase (Mislevy, Steinberg, & Almond, 2003; Plake, Huff, Reshetar, Kaliski & Chajewski, 2015). It is a principled assessment design approach that is engineered towards intended interpretations and uses with explicit design decisions and rationales (Ferrara, Lai, Reilly, & Nichols, 2017). In other words, the building of the validity argument explicitly begins at the design phase of the test development process (Ferrara, Lai, Reilly, & Nichols, 2017; Im, Shin, & Cheng, 2019; Kane, 2015, 2020). “A hallmark of ECD is thus to commence the assessment design process by articulating a chain of reasoning that links evidence to claims about target constructs" (Riconscente, Mislevy & Corrigan 2016 p. 41).

The accountability chapter of the *Standards* (2014) states that intended uses and consequences should be clearly outlined and “evidence to support their validity should be provided when available” (p. 212). The *Standards* recognize that there can also be unintended consequences such as group differences (fairness issues) and washback. Unintended consequences can be anything that increases test scores without truly improving performance on the construct measured by the test. “Potential negative consequences represent hypotheses to be studied — and those studies should be included in the validation framework” (Sireci, 2020, p. 7).

Despite recognizing the need for validity evidence for the use of test scores in accountability systems, the 2014 *Standards* limit test developers’ responsibility for validity research in accountability. Chalhoub-Deville (2020) infers this may be self-serving on the part of the authors of the *Standards* who are professionals in the field of measurement, writing to an audience of other professionals in the field, and not engaging in conversation with users such as teachers, school administrators, and program evaluators. This limited responsibility on test developers puts more responsibility for researching the consequences of accountability systems on test users (p. 248).

The 2014 *Standards* make a distinction between test use at the individual student level and aggregating scores for use in accountability systems. An accountability index is defined as "a number or label that reflects a set of rules for combining scores and other information to arrive at conclusions and inform decision making" (*Standards*, p. 206). Examples of accountability indexes include school performance grades, and value-added measures of teacher effectiveness. According to the *Standards*, a validity argument aids users in understanding the extent to which the model supports causal inferences by providing evidence related to the validity of the interpretations for each use of a test.

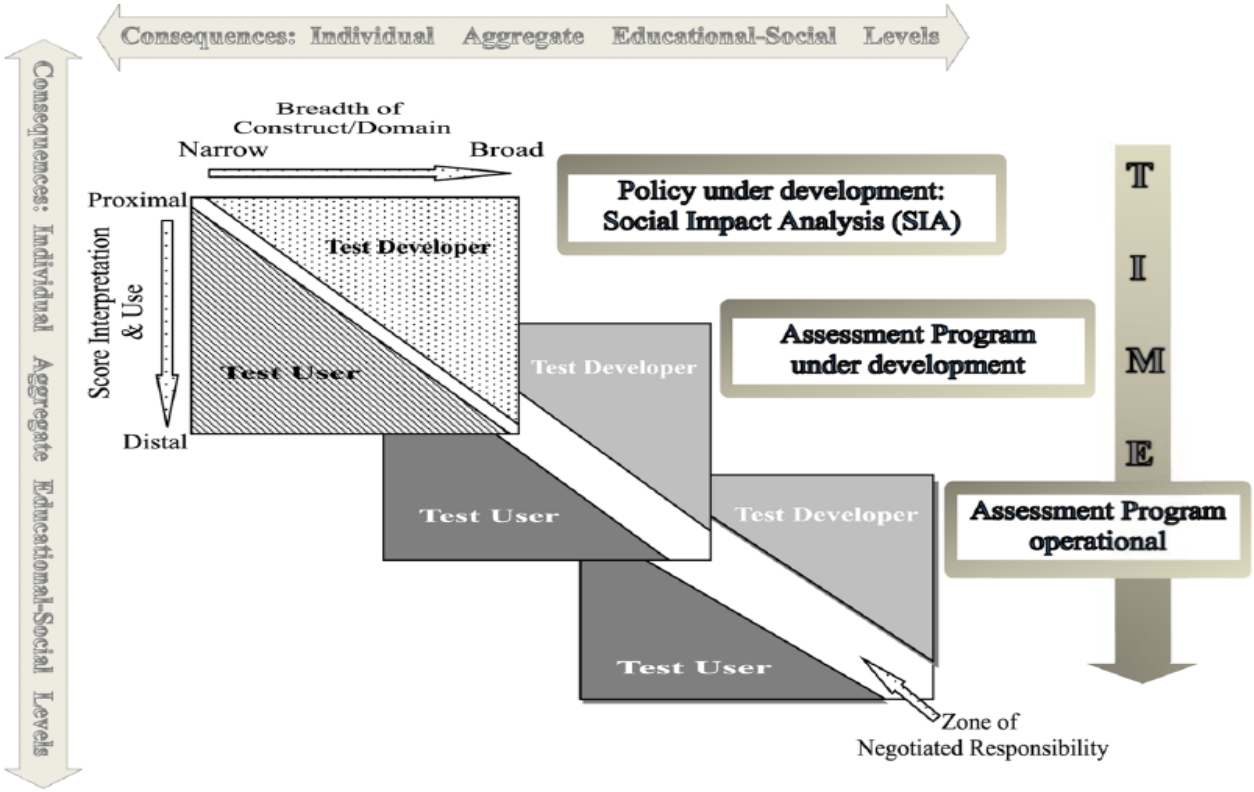
Unit of Analysis and Consequences in Accountability Systems

Education reform shifts responsibility away from students for their performance to holding teachers and schools accountable, and in turn, this shifts the unit of measurement from individual to aggregate scores. With this shift in unit of measure, validity evidence in the traditional individual score unit of measurement now requires consideration of "aggregate and socio-educational consequences" (Chalhoub-Deville, 2020, p. 247). Consequences, also referred to as impact, backwash, and washback, are a subject of discussion and debate among measurement theorists and researchers. The debate is not whether there are consequences in test score interpretation and use, but rather in whether they fall under the purview of validity, in identifying who is responsible for evaluating them, which ones, and when (Chalhoub-Deville, 2016).

With regards to responsibility for the analysis of consequences, Chalhoub-Deville (2016) identifies 'role conflation' between test-taker groups and test-user groups such as federal and state policy makers and the assessment programs or systems that use the tests. Government agencies often dictate "the development, interpretation and use of accountability systems"

(Chalhoub-Deville, 2016, p. 466). The allocation of roles and responsibilities needs to address the "fluid roles test developers and users play in reform-driven assessments" (p. 467). Chalhoub-Deville (2016, p. 467) offers a theoretical guide for structuring role allocation in researching the consequences of test score interpretation (Figure 1).

Figure 1. From "Validity theory: Reform policies, accountability testing, and consequences," by M. Chalhoub-Deville, 2016, *Language Testing, 33*, p. 467. Copyright 2015 by Micheline Chalhoub-Deville. Reprinted with permission.



Consequences are emphasized along both the top and left sides of the figure. Chalhoub-Deville calls out three levels or units of analysis that are relevant in the interpretation and use of assessments for accountability: individual, aggregate, and educational-social. This figure draws attention to the need for validations studies at both the individual and aggregate levels, however there is no empirical work that has looked at methods for doing validation studies at the aggregate level. Responsibility is allocated along a continuum of the breadth of the construct (or

domain) across the top and on score interpretations and uses down the left. The test developer holds greater responsibility in the upper right, i.e., when score interpretation and use is in line with what the test developer intended. Test users take on greater responsibility in the lower left, as score interpretations and uses expand further away from what the test developer intended.

Going from the top left down towards the bottom right, in between the more clearly defined responsibilities, Chalhoub-Deville (2016) identifies a zone of negotiated responsibility (ZNR). As the construct broadens and interpretation and use expand, responsibility for evaluation of consequences begins to encroach into the fuzzier areas of role-conflation. Here the test developer and the test users have shared responsibility and need to discuss or negotiate possibilities for researching the consequences of test interpretation and use. The ZNR grows wider (or more conflated) as it moves from the top left quadrant to the bottom right quadrant of the figure, i.e., as the breadth of the construct increases and the score interpretations and uses move further away from original specifications.

Validating the Consequences of Accountability Systems

Consequences of an accountability system are the rewards, sanctions and interventions imposed on teachers, schools, and districts. Emergent consequences precede rewards and sanctions in anticipation of the possibility that they may be imposed, or they follow the imposed sanctions (CCSSO, 2004). Consideration of emergent consequences requires anticipating not only the consequences that may occur after the implementation of the accountability system, but also the consequences that may occur in anticipation of the implementation. Emergent consequences of accountability systems include activities or conditions in the school that may be positive or negative. Examples of positive emergent consequences are improved teaching and learning. Washback, such as narrowing of the curriculum and focusing on test strategies rather

than on the knowledge and skills the test intends to measure or decreases in morale because of being identified as a low performing school are examples of negative emergent consequences.

Policy interacts directly and indirectly with the consequences of an accountability system. For example, many states are including so-called indicators of college and career readiness in their accountability system for ESSA (CCSSO, 2016) around which there may be policies that offer rewards. In North Carolina, for instance, the Appropriations Act (2016) offers monetary bonuses to teachers through a pilot program to reward "…teacher performance and to encourage student learning and improvement (p. 24, 48)." The potential consequences of a policy implementing teacher rewards for indicators in the accountability model must be considered in the validation plan.

Similarly, consequences of sanctions imposed by policy must also be considered. For example, ESSA (2015) requires that the lowest performing 5% of schools in the state and high schools that graduate less than two-thirds of their students must be sanctioned as Comprehensive Support and Improvement (CSI) schools. Schools for which any subgroup performs in the same manner as a school under the lowest 5% category must be sanctioned as Targeted Support and Improvement (TSI) schools. These schools must also receive support and interventions. In addition to listing the rewards, sanctions and interventions, the validation plan should conceptualize and operationalize how these consequences are supposed to work. It is also necessary to identify potential challenges in implementation and negative consequences (CCSSO, 2004).

States have the responsibility to evaluate their accountability systems to ensure that they are achieving intended goals and outcomes while avoiding potentially negative consequences. The validation plan for an accountability system must analyze both intended and unintended

consequences (Kane, 2006; 2013). The high stakes in identifying teachers and schools for rewards, sanctions, and interventions obligates states to validate that the "right" teachers and schools are being identified. States need to clearly define the goals of the accountability system. Questions that need to be answered include: what kinds of schools are intended to be identified, how trustworthy are the data used in the model, how are the data aggregated to make decisions, and are the interventions appropriate and effective?

The effectiveness of the accountability system in achieving the intended goals can be evaluated through a Theory of Action (TOA) (Chalhoub-Deville, 2016). A TOA explicitly states the intended outcomes as well as the action mechanisms through which they will occur conceptually and operationally. Furthermore, potential implementation problems and negative consequences are identified. A clearly defined TOA allows for a meaningful evaluation of the accountability system (Bennett, 2015). As part of the validation plan, it is necessary to identify and map key intended or imposed consequences including rewards, sanctions, and interventions (CCSSO, 2004).

Purpose

To meet the demands of policies imposed under the auspices of education reform, test scores are being aggregated for use in accountability systems as a measure of the performance of teachers, administrators, and schools. Validity frameworks need to consider the use of aggregate scores and the consequences of their use in accountability systems. Research has not addressed empirical methodologies for building validity evidence to support the use of aggregate or derivative scores at the school level. Chalhoub-Deville (2016, 2020) has rightly pointed to the need to validate the use of aggregate scores and offers valuable guidelines for determining the burden of responsibility in researching consequences for the use and interpretation of aggregate

scores, but this work is theoretical and largely inaccessible to those who design and operationalize accountability systems. Empirical methods need to be identified to operationalize the validation of aggregate scores. This research helps bridge the gap between theory and operationalization of the use of aggregate scores in accountability systems

Study

The aim of this research is to operationalize Chalhoub-Deville's (2016, 2020) work on validity and test-based accountability. A literature review evaluates the applicability of current validity models to the needs of accountability testing. Hierarchical generalized linear models (HGLM) are explored as a method for investigating the legitimacy of a potential divide between individual and aggregate scores. HGLM is particularly appealing in that it allows evaluation of the relationship between test scores and outcome variables at both the student and school level simultaneously. In this study, HGLM is used to analyze the predictive ability of high-school end-of-grade test scores on three outcome variables: high school graduation, dropout, and graduating senior intention survey responses at both the individual and aggregate levels. Graduation rates are calculated at the high school level and represent the proportion of students who graduate from a particular school. Dropout rates are also calculated at the high school level. Graduating senior intention surveys are administered to all graduating seniors. Compositional effects are analyzed to compare the predictive ability of high school end-of-grade test scores for graduation, dropout, and graduating senior intention survey responses at the student and school level. If there are no compositional effects, then there is a degree of validity evidence that the scores have similar meaning at the individual and aggregate levels for the outcome variable.

Research Questions

1. To what extent do key validity models consider accountability testing purposes where the focus is less on individual test scores, and more on aggregate scores?
2. Is there validity evidence to support the assumption that aggregate high school end-of-grade math and English language arts (ELA) scores at the school level are analogous to individual scores at the student level for predicting graduation?
3. Is there validity evidence to support the assumption that aggregate high school end-of-grade math and English language arts (ELA) scores at the school level are analogous to individual scores at the student level for predicting dropout status?
4. Is there validity evidence to support the assumption that aggregate high school end-of-grade math and English language arts (ELA) scores at the school level are analogous to individual scores at the student level for predicting graduating senior intention survey responses?

Assumptions

For the purposes of this HGLM analysis, an assumption is that proficiency on math and ELA test scores are valid and reliable predictors of high school graduation, high school dropout, and graduating senior intention survey results at the individual student level. Even though the magnitude of this association is of interest in choosing indicators for an accountability system, the suitability of the indicators and their degree of association to the outcome variables is not the focus of this research. Instead, the outcome of interest in this study is the compositional effect, or the difference in the between effect (school level) and the within effect (student level) results. The aim of this research is to determine if there is a difference in the within (student) and between (school) effects for the independent variables (math and ELA proficiency) and the

outcome variables of graduation, dropout, and senior intention. In other words, is the association between the independent variables (math and ELA proficiency status) and the outcome variables (graduation, dropout, and senior intention survey responses) the same at the individual student level as it is at the school level for those same independent and outcome variables?

Organization of the Dissertation

Chapter II reviews literature on validity and hierarchical generalized linear modeling (HGLM) as these relate to test-based accountability. The literature review offers an overview of historical and contemporary validity theory and describes considerations for validating the use of aggregate test scores in test-based accountability systems. The final section of the literature review introduces HGLM and compositional effects. Chapter III outlines the study design, modeling approach, and criteria by which the compositional effects are evaluated. Chapter IV presents the results and Chapter V offers a discussion of the implications, limitations, and future directions for research.

CHAPTER II: LITERATURE REVIEW

This chapter is organized into three main sections. The first section offers an overview of the history of validity theory. The second section reviews current validity models and discusses their applicability to accountability systems. The third and final section provides an overview of hierarchical generalized linear modeling (HGLM) and how it can be applied to build validity evidence for aggregate scores used in accountability systems.

An Overview of the History of Validity Theory

Validity theory has been debated in educational and psychological testing for over a century. Early 20th century definitions of validity were primarily empirical, and criterion related. Validity was evaluated in terms of how well a score predicted the criterion (trait or attribute) of interest. The emphasis was on the test itself and the degree to which it correlated with another objective measure of the same attribute (Shaw & Crisp, 2011).

In the 1940's researchers were increasingly concerned that validity was defined and measured only by correlational and factor analysis studies (Sireci, 2009). Researchers argued that criterion-related validity may be inherently flawed if the criterion used lacks validity and reliability in and of itself. These concerns with correlational or criterion-related validity and the concerns regarding the availability of valid comparison criteria paved the way for the concept of content validity.

Content validity shifts the emphasis from a criterion comparison to an operational definition of what the test is intended to measure, which is compared with an analysis of test content (e.g., Rulon, 1946 as cited in Sireci, 2009; Kane, 2006). Content validity has the advantage of providing a validation method that does not depend on an external criterion (Kane, 2006, p. 19). To establish content validity, it is necessary to demonstrate that the test items are a

sample from the domain of interest. A test blueprint is built defining the content areas for the domain of interest and applying weights to each content area. Items are sampled from the universe, or pool of items in that content domain, to build a test with the appropriate weights. Performance on this sample of items is used to estimate overall level of skill or ability in the content domain. Despite the usefulness of methods for establishing content validity for more observable attributes, these methods have been considered weak and less useful in supporting validity claims regarding theoretical constructs such as cognitive processes. Furthermore, it has been argued that methods for content validity are prone to confirmatory bias on the part of test developers (Kane, 2006; Sireci, 2009).

Early approaches to content validity remain influential in educational testing to this day (Kane 2006; Sireci, 2009). Lissitz and Samuelsen (2007) view content validity as key in educational testing. In this view, "...the test definition and development process (what is currently known as content validity) and test stability (what is currently known as reliability, or sometimes generalizability [Brennan, 1983]) become the critical descriptors of the test" (Lissitz & Samuelsen, 2007, p. 446).

The lack of a validation method for theoretical attributes was of concern to psychologists, so the American Psychological Association (APA) Committee on Psychological Tests began searching for types of evidence to support psychological interpretations of theoretical constructs for which there is no established criterion or content domain from which to sample. The committee's consensus recommendations were published in the Technical Recommendations (American Psychological Association, 1954). A key point in the recommendations was the concept of construct validity which focused on the latent trait or attribute that a test intends to measure (Cronbach & Meehl, 1955). The Technical Recommendations were the precursor to the

Standards (AERA, APA, & NCME, 1974, 1985, 1999, 2014), which is also a consensus document.

Construct validity embraces theory and the relationship of test scores to theory. Cronbach and Meehl (1955) define a construct as "some postulated attribute of people, assumed to be reflected in test performance (p. 283)." Because constructs cannot be directly measured with a standardized gauge in the same way as physical characteristics such as height and weight, Cronbach and Meehl (1955) promoted the idea that establishing construct validity involves basic theory or hypothesis testing techniques and deductive reasoning. Their argument was that if a proven theory or precise measurement for a characteristic does not exist, then it is necessary to conduct a series of studies examining different theoretical possibilities. It required that there be a well-defined theory that can be used to make empirical predictions. Like hypothesis testing, if the predictions are not reflected in the measurements, then there are three possibilities: the theory is incorrect, the measurements are inaccurate (or inadequate), or some other assumption was violated. If the predictions are supported by the measurements, then the theory and the score interpretations are also supported (Kane, 2006). Cronbach and Meehl (1955) considered construct validity as an alternative for criterion or content validity in cases where there is not an established criterion against which to compare the measurement.

When the concept of construct validity emerged, it was widely considered that there are distinct types of validity, i.e., content, criterion, and construct validity. According to Messick (1989), content validity involves a professional judgment regarding the relevance of test content to the content domain. Criterion validity uses correlation and regression techniques to compare scores with external variables that are believed to measure the same characteristic. Predictive and concurrent are types of criterion validity that either predict an individual's future level from a

test score or describes their current level on the criterion. Construct validity evaluates the degree of fit between the underlying theory of the characteristic being measured and test performance. One criticism was that researchers could pick and choose from the various types of evidence that supported their argument, potentially ignoring other types of validity evidence that weakened their argument (e.g., Messick, 1989).

Later in the 20th century, theorists moved away from these distinct types of validity and instead proposed construct validity as a unitary concept with several aspects (e.g., Messick, 1989; Moss 2007). The type of validity evidence required depends on score use. If a test score is being used to describe an individual, then content or construct validity evidence is necessary. If a test score is being used to make decisions about a person, then evidence of criterion validity is needed. Given that both content and criterion validity evidence contribute to score meaning, Messick (1989) concluded that both content and criterion validity evidence are aspects of construct validity and therefore there is only one category of validity related evidence – construct.

The different aspects of validity are viewed as pieces of evidence that support the overarching concept of construct validity. Messick (1989) proposed that "[v]alidity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment" (p. 13).

Messick included social consequences of test use as part of the definition of the construct validity of score interpretations sparking a controversy that still divides the measurement community (Cizek, 2016; Newton & Shaw, 2014). Researchers such as Lane (1999), Moss (2013), Shepard (2016), Kane (2006, 2013), and Chalhoub-Deville (2016, 2020) argue that

consequences should be considered in validity evidence, while others, like Lissitz and Samuelsen (2009), argue that consequences should not be considered as part of validity evidence. Messick referenced social consequences, but Shepard (2016) further expanded the concept of test consequences to include positive, negative, intended, and unintended consequences as part of score-based inferences.

Consensus in Contemporary Validity Theory

Professional standards emerged in parallel with the academic developments and changes in validity theory. As mentioned previously, in 1954, the APA, AERA, and NCME developed the Technical Recommendations (American Psychological Association, 1954) which promoted the concept of different types of validity with a preference for presenting multiple types of evidence. The situation dictated what types of evidence were preferable. In the 1966 and 1974 *Standards*, the focus was on criterion-related validity (concurrent and predictive), construct, and content validity (Sireci, 2009). A shift towards the idea of a unitary theory and validating score-based inferences was evident in the 1985 *Standards*.

As discussed in the introduction to this dissertation, the 1999 and 2014 *Standards* describe five sources of validity evidence: test content, response processes, internal structure, relations to other variables, and testing consequences. Notably, the 2014 *Standards* indicate that not all five sources of evidence are required, instead test developers and researchers should gather evidence that they deem appropriate for their validity argument (*Standards*, 2014; Chalhoub-Deville, 2020). This change from the 1999 *Standards* may renew concerns that hearken back to the discussion on content validity, i.e., that researchers can pick and choose from the various types of evidence that support their argument, potentially ignoring other types of validity evidence that weaken their argument (e.g., Messick, 1989).

Cizek (2016) summarizes five areas of general agreement in contemporary validity theory: validity concerns the intended inferences or interpretations made from test scores; the unitary concept focuses on the evidence supporting interpretations of scores with respect to specific constructs; validity judgements are described along a continuum of evidentiary support for the intended score inferences; the validation process is not a one-time activity as there are many factors that can alter original judgements and require new validity conclusions; and, the process of validation involves the application of values (p. 213).

Validity and Accountability

A review of the literature on validity models addresses the first research question in this study: To what extent do validity models consider accountability testing purposes where the focus is less on individual test scores, and more on aggregate scores?

Chalhoub-Deville (2016, 2020) observes that traditionally tests have focused on individuals and as such, validity theory has evolved around score use at the individual level. Chalhoub-Deville also observes that accountability testing has moved beyond individual scores to the use of aggregate scores to evaluate teachers and schools. “This aggregated data is a centerpiece of educational reform policies. Aggregated scores are the unit of accountability; this is where validation needs to be anchored” Chalhoub-Deville (2020 p. 253). Chalhoub-Deville (2016, 2020) also argues for the inclusion of consequences in validating accountability testing and further argues that test developers and users have a shared responsibility in addressing consequences. A validity model that takes accountability systems into account will need to consider the use of aggregate scores during the test development process and the consequences of their use for education reform.

This review of validity models focuses on Kane's (2006, 2013) Interpretation and Use Argument (IA/IUA), a Parallel IUA proposed by Acree, Hoeve, & Weir (2016), and Embretson's (2007, 2008, 2017) Universal Design or Unified Framework. The extent to which these models consider accountability-testing purposes with regards to aggregate scores, consequences, and consideration of validity in the test development process is analyzed. Finally, a comprehensive model that expands on Acree et al.'s modifications of Kane's model is proposed.

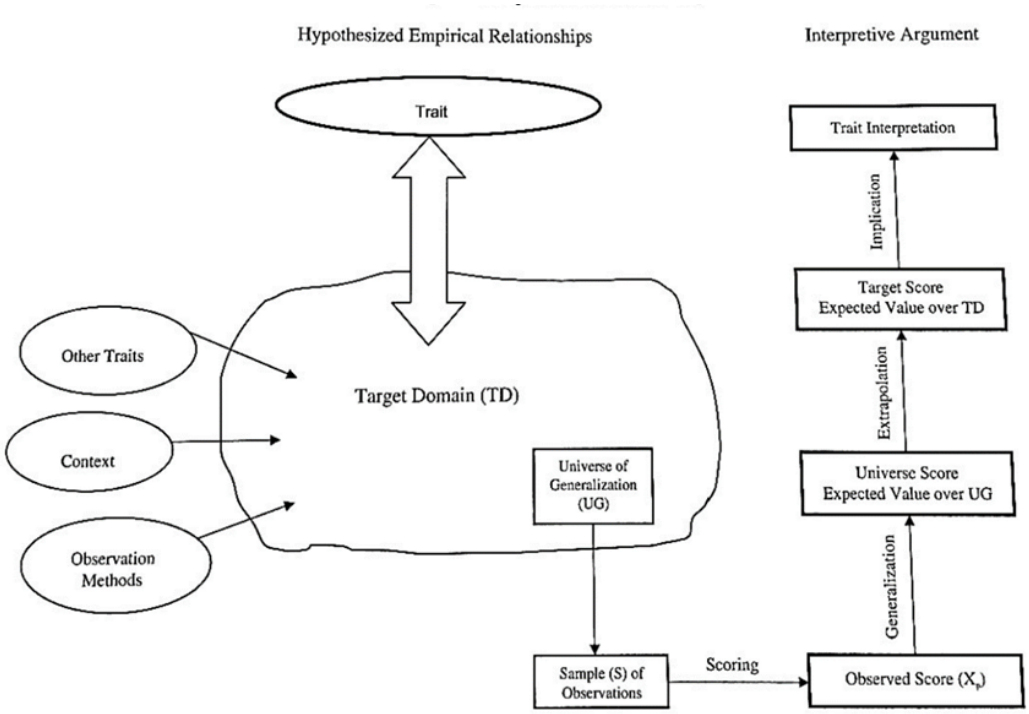
Interpretation and Use Argument (IA/IUA) (Kane 2006, 2013)

An argument-based approach to validity helps to operationalize test validation by providing "a place to start, guidance on how to proceed, [and] criteria for gauging progress and deciding when to stop" (Kane, 2012, p. 8). Kane's Interpretive Argument/Interpretive Use Argument (IA/IUA) shown in Figure 2, offers a roadmap for building a validity argument for trait-based interpretations. The left side of Kane's model, which he labels "hypothesized empirical relationships," represents the definition of the trait or construct. His argument-based approach to validity begins with an interpretive argument (IA). The IA, laid out on the right side of the model, specifies the claims or inferences with regards to score use and interpretation. Kane (2006, 2013) identifies four inferences in the IA: scoring, generalization, extrapolation, and implication. The validity argument (VA) is an overall evaluation of the claims or inferences being made. Research builds evidence to support the claims or inferences laid out in the IA. Kane concludes that the specified interpretations and uses for test scores are valid if the IA/IUA is complete, coherent, and plausible.

Kane (2006) describes a test development strategy involving three iterative steps: outline an interpretive argument, develop the test, and evaluate the inferences and assumptions in the interpretive argument. While test design considerations are inferred in Kane's "hypothesized

empirical relationships," his validity argument does not begin until the scoring inference. As such his argument-based validity model focuses on score inferences but largely overlooks the test development process in which the trait is defined and the contexts and methods for measurement are considered (Chapelle, 2012; Chalhoub-Deville, 2020).

Figure 2. From "Validation" by M. Kane, in R. Brennan (Ed.), *Educational Measurement* (4th ed., p. 33), 2006, Westport, CT: Greenwood Publishing. Copyright 2006 by Michael Kane. Reprinted with permission.



Kane (2006) acknowledged that social consequences of testing were of growing interest and that consequences (positive and negative) play a role in validation. Even though that role is "somewhat contentious" in the field, positive consequences should "outweigh" negative consequences in general (Kane 2006, p. 51). Furthermore, Kane (2006 p. 55) specifically noted that educational reform and accountability call for an evaluation of consequences. "The accountability program is an educational intervention, and a serious evaluation of an

accountability program would require an evaluation of both intended and unintended outcomes" (Kane, 2013, p. 54).

Kane's IA/IUA (2006, 2013, 2017, 2020) offers a useful roadmap for the operationalization of validation. However, despite his recognition of the importance of test design, consequences, and score interpretation, Kane's "…model, nevertheless, remains anchored in individual test scores, which does not accommodate accountability testing realities" Chalhoub-Deville (2020).

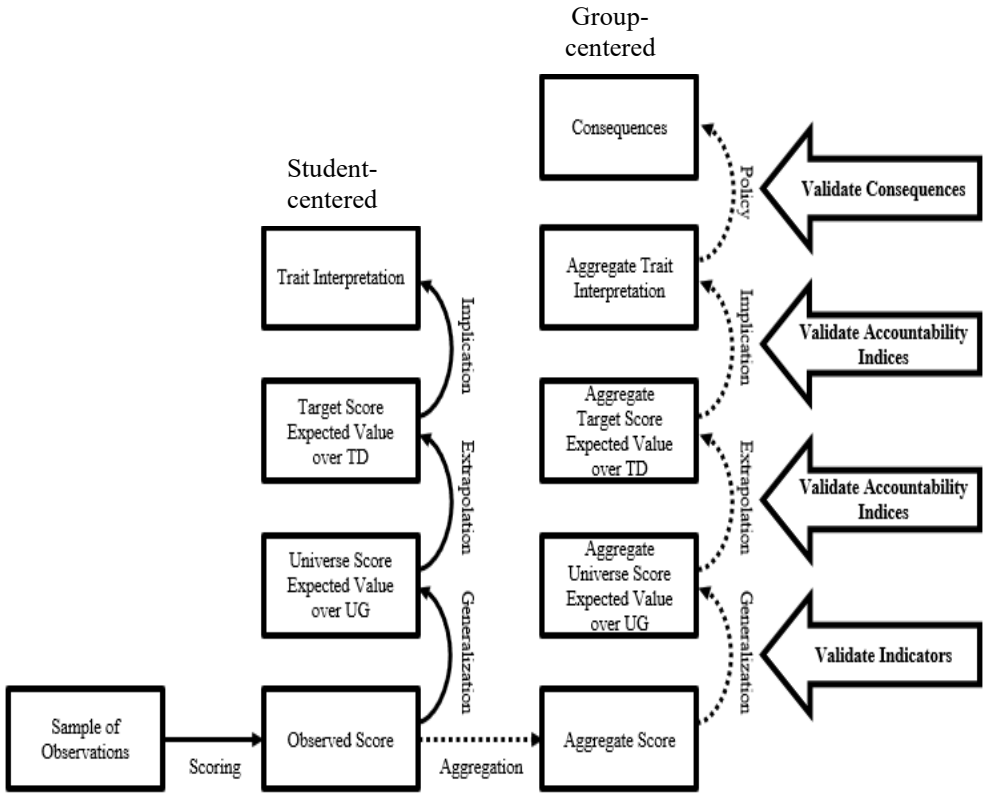
Parallel IUA (Acree, Hoeve, & Weir, 2016)

While Kane's technique addresses the test score itself, accountability systems present a different, though closely related problem in that the scores in question are aggregate in nature. Building on Kane's validation framework, Acree, Hoeve, and Weir (2016) proposed that Kane's (2006, 2013) IUA for validating uses of individual scores can be extrapolated and expanded for validating uses of aggregate and derivative scores in accountability systems. Individual scores and aggregate scores are similar enough that a common strategy may be used to evaluate the degree to which both are valid. In this framework (Figure 3), individual and aggregate scores are evaluated independently and in parallel. Both branches must be interrogated and interpreted systematically and separately. The validation of accountability systems concerns itself primarily with the group-centered branch.

The parallel, but independent nature of these evaluations maintains that even if strong evidence of validity is established along the individual or student-centered branch, this does not imply the same will hold for the aggregate or group-centered branch. Nor is validity evidence at the individual score level a necessary part of validation for use at an aggregate level. Similarly, this model holds that failure of an inference in the student-centered branch does not necessarily

undermine the validity of the group-centered branch. Validity evidence may be found for individual scores, aggregate scores, for both, or for neither.

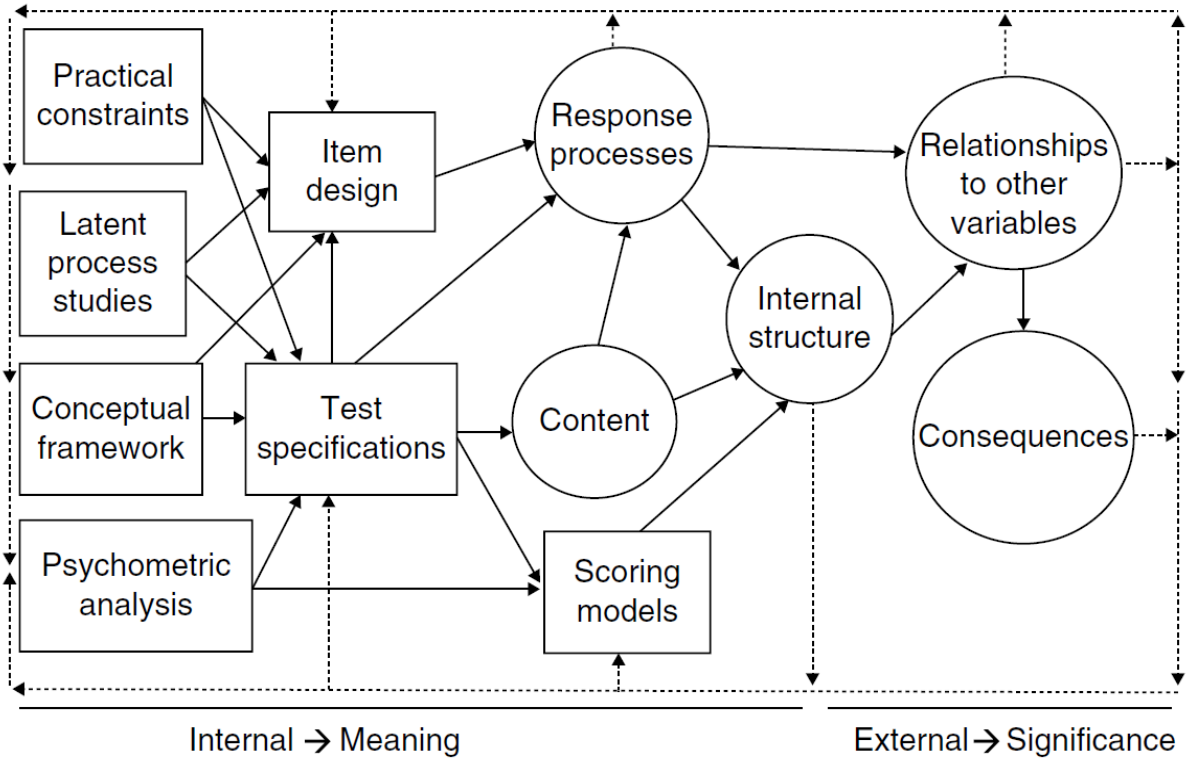
Figure 3. Parallel Validation for Accountability Testing by Acree, Hoeve, & Weir (2016)



Universal Design/Unified Framework (Embretson 2007, 2017)

Embretson (2007, 2008) proposed a validity framework that she described as universal and interactive. According to Embretson (2007), "the system is universal because all sources of evidence are included and may be appropriate for both educational and psychological tests" and "interactive because the adequacy of evidence in one category is influenced or informed by adequacy in the other categories" (p. 452). Embretson (2017) reconceptualized her universal system for validity as shown in Figure 4.

Figure 4. From "An Integrated Framework for Construct Validity," by S. Embretson in A. A. Rupp and J. P. Leighton (Ed.), *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*, (p. 104), 2017, Chichester, UK: John Wiley & Sons. Copyright 2017 by Susan Embretson. Reprinted with permission.



Embretson's unified framework is divided into internal and external aspects of construct validity. The seven rectangles on the left side represent test development processes. Embretson characterizes these test development processes as internal aspects of validity. The five circles on the right side are external and correspond to the five sources of evidence defined in the *Standards* (2014).

Embretson (2017) emphasizes test development processes as essential internal aspects of her validity framework. She identifies the need for a conceptual framework (like that offered by evidence-centered design). Embretson (2007, 2008, 2017) advocates for including categories of evidence that would be evaluated during the test development cycle as part of her validity system. She includes categories of evidence for practical constraints (e.g., test administration

methods and scoring mechanisms); item design principles (e.g., formats, context, complexity, and specific content); domain structure (specification of content areas and levels); and test specifications (e.g., blueprints).

Impact or consequences are also explicitly included as an aspect of Embretson's (2007, 2008, 2017) validity framework. Embretson describes concern for differential item functioning among groups and the potential impact on selection or placement at the individual level. She also recognizes a role for test developers in consequences saying, "there may be aspects of test specifications and item design that could be changed to reduce impact" (2017, p. 108).

Of the validity models reviewed, Embretson's (2017) framework is the most comprehensive. However, aggregate scores and the potential consequences of their use in accountability systems are not specifically addressed in Embretson's validity framework (2007, 2008, 2017).

Validation and Accountability Systems

Historically, validation approaches have been proposed for study at the individual score level where the test user wants to draw an inference about an individual test taker (e.g., placement testing, achievement testing, etc.). Accountability systems use group-level aggregate test scores and derivatives of test scores to draw conclusions about schools and teachers. Neither historical nor current theories of validity and validation explicitly address the use of group-level test scores, as used in accountability systems, nor have methodologies for building validity evidence for group-level scores been explored.

Proposed Validation Framework: Accountability IUA

To address the omission of validation during the test development phase of an accountability system, the Parallel IUA framework (Figure 5) proposed by Acree et al (2016) for

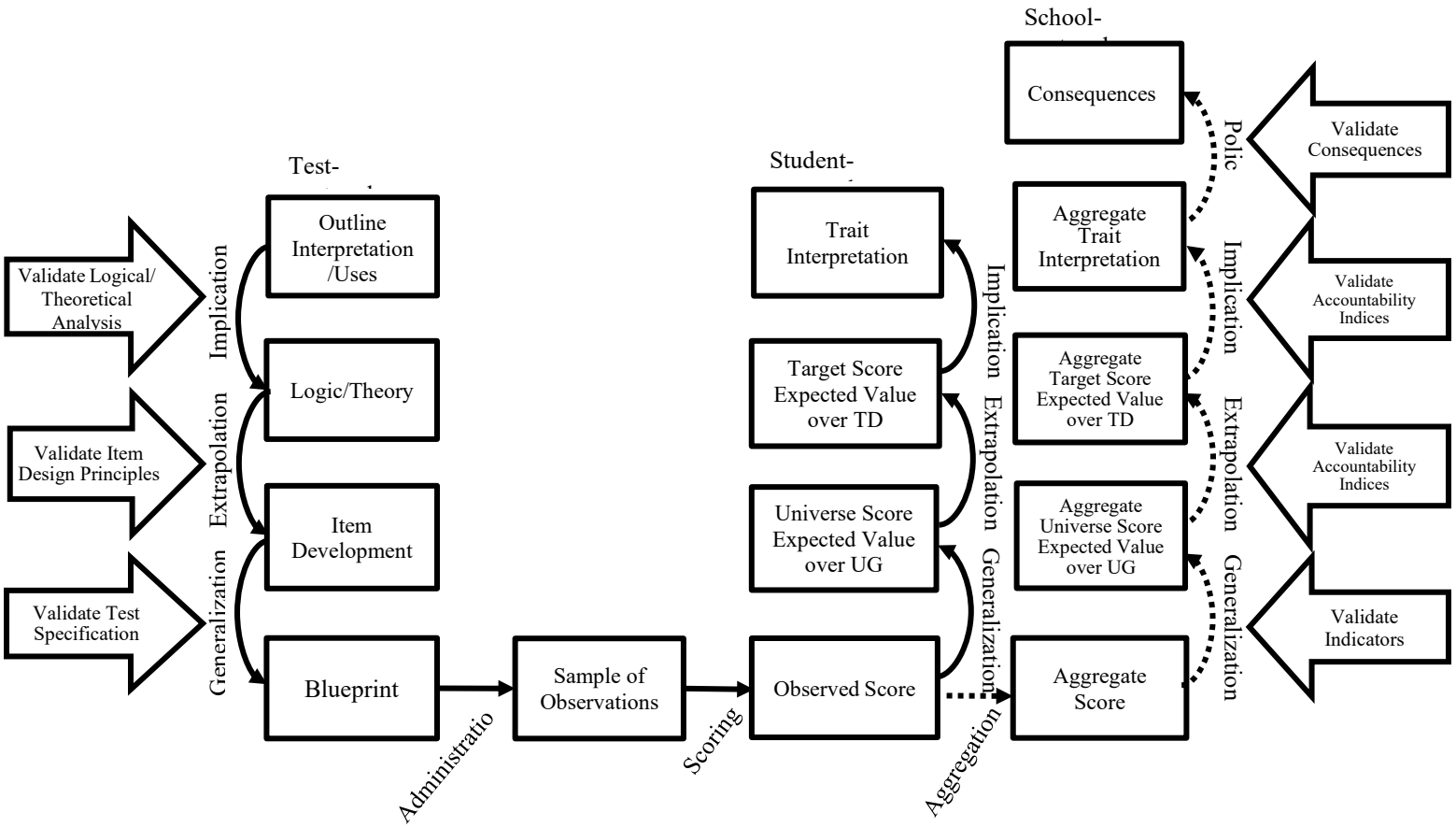
validating accountability systems has been expanded into an Accountability IUA framework for accountability systems that includes a test-centered branch.

In the proposed Accountability IUA framework, Kane's (2006, 2013) generalization, extrapolation, and implication inferences occur in the reverse order during the test-centered (test development) branch. In a sense, test development involves beginning with the end in mind, so the logical progression of building validity arguments is reversed. The intended interpretations and uses are defined and evidence gathered for the logical and theoretical analysis of the implications (the decision or action being taken from the scores). Item design principles, such as evidence centered design (ECD) provide evidence for extrapolation (using the scores as a reflection of real-world performance).

Validation of the test specifications supports generalization (using the scores as a reflection of performance in the test environment). Once the test administration is operationalized, building evidence for the student-centered and group-centered IUA branches begins. Evidence must be compiled to support the validity of inferences based on the accountability system indicators. Compiling evidence based on content, internal structure, and generalizability is especially pertinent to the validation process of accountability systems.

Evidence that is based on content links the features of a test to the construct of interest (*Standards*, 2014). Evidence based on internal structure (Wilson, 2008) is produced by comparing the results from statistical analyses of the relationships between items of a test (through factor analysis, structural equation modeling, etc.) to theoretical characterizations of the construct. Evidence supporting the generalizability of an indicator links the indicator to situations beyond the immediate interpretation of that indicator (*Standards*, 2014).

Figure 5. Accountability IUA Validity Framework



Extrapolation, Implication, and Accountability Indices

Indicators such as end-of-grade test scores, value-added model or growth scores and graduation rates are often combined in accountability systems to make judgments about teachers, schools, and school districts (*Standards*, 2014). This combining of scores serves a similar function to extrapolation and implication inferences, which extend interpretations of test scores to a target domain and trait interpretation in Kane's (2006, 2013) IUA framework.

Extrapolation and implication are distinct, but for the purpose of this discussion, both are defined as inferences intended to broaden test score interpretations to include "real-world" performances (Kane, 2013, p. 28). In accountability systems, decision rules and accountability indices are the mechanisms by which extrapolation and implication occur (Council of Chief State School Officers, referred to hereafter as CCSSO, 2004; Kane, 2013). Indices synthesize data based on decision rules to provide a single score that is used to make judgments about educational quality and student success (*Standards*, 2014). It is the interpretation and use of these indices that must be validated as an argument that is built for the system as a whole (Kane, 2013). The decision rules and indicators used to construct indices are to be evaluated in relation to the operational definitions of educational quality and school success, set by state and federal mandates such as the Every Student Succeeds Act (2015).

These definitions are, in effect, the target domain and trait for accountability systems. Analytic and empirical evidence are gathered to support extrapolation and implication inferences (Kane, 2006). For individual test interpretations, analytic evidence is found in development, when the content, tasks, and processes included in the test are compared with those encompassed by the target domain (Kane, 2013). The closer the test mirrors the target domain, the easier it is to support the extrapolation inference (Kane, 2013). Empirical evidence is derived from external

criteria. Test score interpretations are compared against other measures of the same target domain (Kane, 2006). This can include performance-based assessments that may only be practical for smaller sample sizes (Kane, 2006).

In aggregate score interpretations, analytic evidence should be gathered during the development of accountability indices (CCSSO, 2004). Because of the elusiveness of terms such as educational quality and school success that define the target domain, a pragmatic approach to validation is most appropriate (Moss, 2013; Kane, 2013). The weights given to indicators of quality and success in accountability indices should be scrutinized using stakeholder definitions and interpretations. Data triangulation should be used to inform judgments based on aggregate scores, giving voice to policy makers, school leaders, and teachers alongside strict, quantitative decision rules (Moss, 2013; Kane, 2013). The extent to which that triangulation occurs is evidence for the validity of aggregate score interpretations. Along the same line, extrapolation and implication arguments must include evidence of fairness and transparency (Kane, 2010, 2013). Empirical evidence stems from a critical perspective in both accountability and test-based validity (CCSSO, 2004; Kane, 2006). Once tests and accountability indices are operational, evidence is gathered and compared with other criteria linked to the same target domain or trait.

For accountability systems, longitudinal studies may be used to support score interpretations for groups of students by comparing them to long-term student outcomes (CCSSO, 2004). Other indices (e.g., Adequate Yearly Progress, EVAAS), stakeholder surveys, and document analysis could also be used as criteria for comparison (Lane & Stone, 2002). Empirical evidence must also refute threats of trait under-representation and irrelevant variance (Kane, 2006). To do so, it is important to consider the effects of external factors, such as

educational opportunity, English learner status, race, and socioeconomic status on aggregate score interpretations.

Building Empirical Evidence for the Use of Aggregate Scores

Using test scores as indicators in an accountability system often assumes that aggregations or derivations of test scores at the school level have the same meaning as the individual score has at the individual level. That assumption can be tested through hierarchical generalized linear modeling (HGLM).

Hierarchical Data

Student performance does not occur in isolation. Rather, it exists as part of a series of nested and hierarchical effects. Focusing on the individual student score alone does not account for the effects of the classroom or school in which the student is nested. Independence, an assumption of linear regression, is often violated in situations where data are nested. Independence is violated because students in schools have shared experiences and, as a result, they tend to be more like each other than they are to students from different schools. Simply aggregating data to the school level using mean test scores and performing linear regression to address these concerns ignores within group variation and interactions between the different levels thereby complicating the interpretations.

Hierarchical Linear Modeling (HLM)

A natural approach that can address the concerns of nested data is Hierarchical Linear Modeling (HLM, Raudenbush & Bryk, 2002). HLM allows researchers to study effects at the student and school level in addition to the interactions across. HLM is particularly useful for contextual analysis, which is also referred to as multilevel modeling. This analysis treats students (level-1 units) as nested within schools (level-2 units). In effect, HLM has the goal to explain

variation of individual-level behaviors in terms of individual effects within a school (within effects) and group-level effects between schools (between effects). HLM estimates the true or correctly specified model of the relationship between independent and dependent variables while also accounting for dependencies in the error terms

Hierarchical Generalized Linear Modeling (HGLM)

This analysis focuses on a two-level model where students (level-1) are nested within schools (level-2). Typically, an HLM model would assume a continuous dependent variable. However, in this study the dependent variables are nonnormal. Specifically, the dependent variables of interest are high school graduation, dropout, and senior intention survey responses. Because the dependent variables are nonnormal and the data is nested, a more general hierarchical linear model is considered that is referred to as the hierarchical generalized linear model (HGLM, Raudenbush & Bryk, 2002, p. 292). The HGLM is an extension of the generalized linear model (GLM, Lee & Nelder, 1996), which is typically used to model variables that are dichotomous, counts, ordinal, or nominal. For example, the GLM is also referred to as logistic regression when modeling a dichotomous dependent variable. Because HLM models are linear models, many of the typical features and assumptions of regression still apply. For example, multicollinearity within schools or between schools could become an issue in that, as variables are added that correlate with variables that are already in the model, there may be changes in both significance and in the coefficient.

Conceptually, the GLM uses a linear model to predict a specific function (called the link function) of a parameter that defines the distribution of the outcome. For example, the link function of the logistic model is the logit (i.e., the log-odds), which is a function of the probability of a success. Whereas the “identity link” indicates that a linear function is used to

predict the expected value (i.e., mean) of the dependent variable directly. For a more detailed description of the GLM see Carey (2013) and Nelder and Wedderburn (1972).

Because the outcomes of this study are either dichotomous or multinomial, two specific instances of the HGLM are discussed in the next section. These two instances are described as the hierarchical logistic model for dichotomous outcomes and the hierarchical multinomial model for multinomial outcomes

Hierarchical Logistic Model

When the dependent variable is dichotomous, the dependent variable can only take on one of two possible values, 0 or 1, and as a result, the residuals of these values cannot be normally distributed. In addition, the variance for the level-1 error, should one use a typical linear model, depends on the predicted value (i.e., the probability of a success) and is therefore not homogenous. Finally, the predicted value of the dichotomous outcome variable must be constrained to only those values between 0 and 1, inclusive, because it is a probability. Therefore, when the outcome variable is dichotomous (a logistic model), HGLM is a more robust multilevel analysis than HLM.

The level-1 HGLM model has three components: a sampling model, a link function, and a structural model. The sampling model describes the distribution of the dependent variable given the independent variable. Because the dependent variable is dichotomous it is assumed that the dependent variable follows a Bernoulli distribution with the probability of a success equal to ϕ_{ij}. The link function describes the relationship between what is being predicted and the relevant parameters for the sampling model. In this case the link function is the log-odds link (also known as the logit link). The structural model describes the linear combination of variables that predict the specific outcome. Specifically, the level-1 hierarchical logistic sampling model is specified:

$$Y_{ij}|\Phi_{ijj} \sim B(m_{ij}, \Phi_{ij}) \quad (1)$$

where Y_{ij} has a binomial distribution with m_{ij} trials, in this case m_{ij} = 1, and the probability of success on each trial is $\boldsymbol{\Phi_{ij}}$. The level-1 link function, the logit link, is specified as:

$$\eta_{ij} = \log\left(\frac{\Phi_{ij}}{1-\Phi_{ij}}\right) \quad (2)$$

where $\eta_{ij}$is the odds of success or the odds that $\eta_{ij} = 1$.

Level-1 structural model:

$$\eta_{ij} = \beta_{0j} + \sum_{q=1}^{Q} \beta_{qj} X_{qij} \quad (3)$$

where

β_{0j} is the expected change in the log-odds per unit increase in X_q

Thus, the final model is specified as:

$$\eta_{ij} = \beta_{ij} + \lambda_{ij} + \sum_{s=1}^{S_q} \gamma_{qs} W_{sj} + u_{qj} \quad (4)$$

where the random effects, u_{qj}, q = 0,…,Q, constitute a vector $\boldsymbol{u}_j$ having a multivariate normal distribution with component means of zero and a variance covariance matrix $\boldsymbol{\tau}$.

Multinomial Model

When the dependent variable has more than two categories, the multinomial HGLM is used. Multinomial HGLM extends the hierarchical logistic model to more than two possible outcomes. A referent group (*M*) is identified, and all other groups (*m*) are compared to the referent group such that η_{mij} is the log odds of the probability of group *m* versus the referent group *M*:

$$Y_{ij}|\boldsymbol{\Phi}_{\boldsymbol{ij}j} \sim Multi(m_{ij}, \boldsymbol{\Phi}_{\boldsymbol{ij}}) \tag{5}$$

where

$\boldsymbol{\phi}_{\boldsymbol{ij}} = \{\phi_{1ij}, \ldots, \phi_{Mij}\}$ such that ϕ_{mij} describes the probability of observing the m^{th} category

The following logit link function is used for the m^{th} category relative to the referent category:

$$\eta_{mij} = \log\left(\frac{\phi_{mij}}{\phi_{Mij}}\right) \tag{6}$$

where

$$\phi_{Mij} = 1 - \sum_{m=1}^{M-1} \phi_{mij} \tag{7}$$

where

ϕ_{mij} is the probability that person *i* in group *j* will be in category *m* for categories *m* = 1, ..., *M*, (a total of *M* categories). For *M* categories, there are (*M* – 1) sets of equations with membership in category *m* relative to category *M*.

Given the sampling model and the link functions, the structural model can be defined for level-1 and level-2.

Level-1 structural model:

$$\eta_{mij} = \beta_{0j(m)} + \sum_{q=1}^{Q} \beta_{qj(m)} X_{qij} \tag{8}$$

where

$\beta_{0j(m)}$ is the expected change in the log-odds of group (*m*) versus the referent group (*M*) per unit increase in X_q

Given the level-1 model, each one of the coefficients can be modeled using the level-2 model.

Level-2 multinomial model:

$$\beta_{qj(m)} = \gamma_{q0(m)} + \sum_{s=1}^{S_q} \gamma_{qs(m)} W_{sj} + u_{qj(m)} \quad (9)$$

where the level-1 coefficient is now predicted from a set of level-2 variables (W_{sj}), $\gamma_{q0(m)}$ is the expected level-1 coefficient while level-2 independent variables equal 0, and $\gamma_{qs(m)}$ is the expected change in the level-1 coefficient for unit increase in *W*.

Compositional Effects

One advantage of HGLM is that it can identify the effect of an independent variable on an outcome at the individual student level while at the same time identifying the effect of an aggregate or derivative of that same independent variable on the same outcome at the school level. For example, a student's proficiency status on a math test may predict the probability of that student graduating and simultaneously the school's percent of students proficient on that math test may also predict the school's graduation rate. Because HGLM allows for these both to be analyzed at the same time, direct comparisons can be made and tested for differences.

The comparison (i.e., difference) of the school level-2 effect versus the student level-1 effect is called a compositional effect. Thus, in the previous example, how predictive a school's percent proficient is for the school's graduation rate can be compared to how predictive math proficiency is with respect to an individual student's graduation. Note that the level-1 effect is a within effect and the level-2 effect is a between effect, so the compositional effect is defined as the between effect minus the within effect. If the difference is zero, there is no compositional

effect. In the context of this research, compositional effects are useful in examining validity evidence for indicators in accountability models because they allow the assessment of whether an indicator correlates with an outcome for a student (level-1) in the same way as it does for a school (level-2).

CHAPTER III: METHODS

Data and methodology were selected to address the following four research questions:

1. To what extent do key validity models consider accountability testing purposes where the focus is less on individual test scores, and more on aggregate scores?
2. Is there validity evidence to support the assumption that aggregate high school end-of-grade math and English language arts (ELA) scores at the school level are analogous to individual scores at the student level for predicting graduation?
3. Is there validity evidence to support the assumption that aggregate high school end-of-grade math and English language arts (ELA) scores at the school level are analogous to individual scores at the student level for predicting dropout status?
4. Is there validity evidence to support the assumption that aggregate high school end-of-grade math and English language arts (ELA) scores at the school level are analogous to individual scores at the student level for predicting graduating senior intention survey responses?

To answer research question one, the literature review in this dissertation included an analysis of three validity models, the Interpretive Argument/Interpretive Use Argument (IA/IUA) by Kane (2006, 2013), a Parallel IUA by Acree, Hoeve and Weir (2016), and the Universal or Unified Validity System by Embretson (2007, 2008) with regards to their applicability to test-based accountability systems. Three primary areas that have not been fully addressed in building a comprehensive validity argument for test-based accountability systems are identified: the test development process, evaluation of consequences, and the use of aggregate scores. A validation framework for accountability systems was proposed that

incorporates a systematic approach to beginning the development of a validity argument in the test design and development phase, a demarcated, parallel process for building a validity argument for aggregate scores in addition to individual scores, and consideration of consequences of aggregate score use in test-based accountability systems. Finally, to operationalize the validation of the use of aggregate scores, hierarchical generalized linear modeling (HGLM), is explored as an empirical method for comparing the predictive ability of individual and aggregate scores on the same outcome variables at the individual and aggregate levels. The use of HGLM is demonstrated as one approach to provide validity evidence for the use of standardized tests in an accountability system.

This chapter has two sections. First the study design is presented. Second, the hierarchical generalized linear modeling (HGLM) approach to answering research questions two through four is described along with the evaluation criteria.

Study Design

To answer research questions two through four, this research uses data from a southeastern state for high school students who were expected to graduate in 2018-19. Two end-of-grade standardized tests that are used as indicators in the state's accountability system are included in this analysis as independent variables for each student, i.e., high school math, and high school English language arts (ELA). In addition, the outcome variables of high school graduation and dropout status are collected. Finally, graduating senior intention survey responses for graduates in the 2018-19 school year are included as a proxy for a post-high school outcome since no post-high school outcome variables are available in the state's accountability data. The data is hierarchical in that the students are nested within schools.

Demographics

Available demographic data are race/ethnicity, economically disadvantaged status (EDS), and English learner status (EL). Race/ethnicity is a self-reported categorical variable and EDS and EL statuses are binary classifications. The descriptive statistics for these demographic variables are presented in Table 1 and Table 2.

Table 1. Race/Ethnicity of 2019 Cohort of Expected Graduates

Race/Ethnicity	N	Percent
American Indian or Alaska Native	1556	1.25
Asian	3536	2.84
Black	31765	25.53
Hispanic/Latino	18643	14.98
Two or more races	4960	3.99
White	63842	51.31
Native Hawaiian or Other Pacific Islander	133	0.11

n =124435

Table 2. Demographics of 2019 Cohort of Expected Graduates

	N	Percent
Economically disadvantaged	48021	38.59
English Learner	6060	4.87
Female	60951	48.98

n =124435

Dependent Variables

Ideally the outcome variables for validating the indicators in a high school accountability model would be post high school, evidence-based measures of college and career readiness or other metrics of educational quality and success as defined by stakeholders and policy makers who drive educational reform. Potential measures might include outcomes such as college GPA, college graduation, wages above poverty level, employer surveys indicating satisfaction with high school graduate's knowledge and skillset, etc. Evidence-based post high school outcome

measures were not available for this research. However, given that the goal of this research is to present a methodology that could be used to build validity evidence for test-based indicators in an accountability system, this research uses graduation status, dropout status, and graduating senior intention survey responses at the end of the 2018-19 school year as examples for the dependent variables.

Graduation is a binary variable indicating whether a student graduated on time, i.e., with their cohort. Dropout status is also a binary variable that indicates whether a student dropped out prior to their expected graduation date. If a student drops out, their graduation status is automatically set to not graduated. However, not all students who fail to graduate with their cohort are dropouts, so the two are not mutually exclusive. For example, a student may graduate late, in which case they count against the 4-year cohort graduation rate for the purposes of the accountability system. Or a student may be enrolled in a 5-year early college high school program in which case, they would not count against the high school's 4-year cohort graduation rate. Therefore, even though a student may eventually graduate, for this study they are considered not graduated. The descriptive statistics for the state's 2018-19 cohort graduation status and dropout status are presented in Table 3.

Table 3. Graduation and Dropout Status for Cohort Expected to Graduate in 2019

Outcome	N	Percent
Graduated with cohort	101539	81.6
Dropped out	5445	4.38

n =124435

The third dependent variable, graduating senior intention survey is administered to seniors just prior to graduation. The survey has twelve response options. One response option is a plan for employment; another is a plan to enter the military. Ten of the response options are

plans to attend various educational institutions for further education. These ten response options were collapsed into one category for further education. The final response option of "other plans or don't know" was included with the missing responses. This collapsing of the responses results in three mutually exclusive categories: employment, military, and education. The descriptive statistics for the graduating senior intention survey responses, prior to collapsing into three categories, are presented in Table 4. The collapsed graduating senior intention survey results that are used as the dependent variable in this study are presented in Table 5.

Table 4. 2018-19 Graduating Senior Intention Survey Results

Intention	N	Percent
Get a full-time job	13141	10.56
Go into military	4178	3.36
Education		
Attend in-state public or community or technical college	34007	27.33
Attend out-of-state public or community or technical college	671	0.54
Attend in-state private junior college	252	0.20
Attend out-of-state private junior college	73	0.06
Attend in-state public senior institution	31565	25.37
Attend out-of-state public senior institution	3311	2.66
Attend in-state private senior institution	6351	5.10
Attend out-of-state private senior institution	2446	1.97
Attend in-state trade, business, or nursing school	636	0.51
Attend out-of-state trade, business, or nursing school	129	0.10
Other plans or don't know	1538	1.24

n =124435 (frequency missing = 26137)

Table 5. 2018-19 Collapsed Graduating Senior Intention Survey Results

Intention	N	Percent
Employment (Get a full-time job)	13141	10.56
Military (Go into military)	4178	3.36
Education (Reference group)	79441	63.84

n =124435 (frequency missing and "other plans or don't know" = 27675)

Independent Variables

The independent variables in this study are binary indicators for whether a student's test score meets the state's proficiency level on high school accountability tests in math and ELA. The math and ELA tests are multiple-choice assessments aligned with the state's standard course of study. The tests are administered at the end of the school year, test forms are parallel in content coverage for the state's standard course of study, and total test scores are statistically equivalent across forms. While this study uses a binary indicator for meeting the state's minimum proficiency level on the tests, this same method would also work for other score representations such as scale scores, raw scores, percentiles, or an ordinal range of achievement levels. These tests are high stakes and are intended to hold students, staff, and schools accountable for academic performance. Educators may use the results in making promotion, remediation, acceleration, and graduation decisions at the individual student level.

So far this discussion of independent and dependent variables has focused on individual student proficiency on high school math and ELA tests and on the individual outcomes of graduation, dropout, and senior intention survey responses. However, for accountability purposes, the math and ELA tests are also aggregated at the school level. The aim of this research is to demonstrate a methodology to determine the usefulness of these same tests in evaluating both individual performance and school performance for accountability purposes. Therefore, in addition to student-level proficiency, proficiency status is aggregated as the proportion proficient on each test at the school level. The dependent variables are also aggregated at the high school level as proportions. That is, the proportion of graduates and dropouts, and the proportion of students responding to the graduating senior intention survey in each of the three categories: employment, military, or continuing education are calculated for

each school. Proportions are used in the model because this is what is used in the state's accountability system. The independent and dependent variables are then analyzed both at level-1, the student level, and at level-2, the school level.

Level-1 Variables

Student-level math proficiency and ELA proficiency status are included as separate independent variables in the HGLM model at the individual student level. The percent proficient in high school math and high school ELA for all students in the cohort, across all schools, are presented in Table 6.

Table 6. Level-1 (Student) Percent Proficient

Assessment Subject Area	Variable Name	N	Percent
High School Math	MathProf	63800	51.3
High School ELA	ELAProf	70379	56.6

n =124435

Level-2 Variables

The proportion of students who are proficient in math and ELA is used at Level 2, the aggregate or school level. The aggregate proportion proficient across all schools is presented in Table 7.

Table 7. Level-2 (Average School) Descriptive Statistics

Outcome Assessment	Variable Name	N	Average Proportion
High School Math	MathProp	637	49.7
High School ELA	ELAProp	637	55.2

n =124435

HGLM Approach

Data in this study are hierarchical such that students are nested within schools. Recall that Hierarchical Linear Modeling (HLM) partials total variance in the dependent variable (i.e.,

graduation, dropout, senior intention survey responses) into within and between-school variance, thereby providing the opportunity to disaggregate individual and group effects. Specifically, the relationship between proficiency and a dependent variable, e.g., graduation, can be explored with respect to both levels (student and school). Therefore, how proficiency predicts graduation at the student level can be compared to how proportion proficient predicts school graduation rates. By including both the level-1 and level-2 variables, a comparison can be made, and compositional effects can be explored to determine whether the effects are the same at both levels. Compositional effects compare the individual student level performance (level-1) to school level performance (level-2).

The dependent variables are dichotomous (graduation and dropout) and categorical (graduating senior intention survey responses) therefore, hierarchical generalized linear modeling (HGLM) is used to identify the compositional effects.

Addressing the Research Questions

Keeping in mind that end-of-grade tests are used as high stakes tests for students (e.g., in making decisions regarding mastery, placement, promotion, etc.) in addition to being used in accountability systems for the teacher and school, if the relationship between proficiency status and evidence based-outcome measures is different at the student and school level then it may not be reasonable to use these test scores as indicators of both student and school success in achieving college and career readiness. Recall that for this study, evidence-based measures were not available, therefore graduation, dropout, and senior intention survey responses are used as proxies for outcome measures for the purposes of demonstrating the methodology.

Three separate HGLM models are estimated using HLM 8 (Raudenbush, Bryk, Cheong, & Congdon, 2019) to compare the relationship between the tests at the individual level and at the

school level. For each of the two dichotomous dependent variables (graduation and dropout), a hierarchical logistic model is estimated. For the categorical dependent variable, graduating senior intention survey results, a hierarchical multinomial model is estimated.

Hierarchical Logistic Model

Research question two investigates whether the high school end-of-grade math and ELA tests are as good at predicting graduation at the student level as aggregate end-of-grade math and English tests are for predicting overall graduation rates at the high school level. Similarly, research question three investigates whether end-of-grade math and ELA tests are as good at predicting dropout at the student level as aggregate end-of-grade math and ELA scores are for predicting overall dropout rates at the high school level. Graduation and dropout are dichotomous variables such that 1 = yes, 0 = no, therefore a hierarchical logistic model [equations (1), (2) and (3)] are used.

Specifications for the level-1 hierarchical logistic model is:

$$log\frac{\Phi_{ij}}{1-\Phi_{ij}} = \beta_{oj} + \beta_{1j} * \left(MathProficient_{ij}\right) + \beta_{2j} * \left(ELAProficient_{ij}\right) \quad (10)$$

where

$MathProficient_{ij}$and $ELAProf_{icientij}$ are group mean centered, and

β_{0j} is the mean log odds of graduation (or dropout) in school j,

β_{qj} is the effect of X_{qij} on the log odds of graduation (or dropout) in school j,

X_{qij} is the level-1 predictor, or student variable, q.

The corresponding level-2 hierarchical logistic model is:

$$\beta 0j = \gamma 00 + \gamma 01*(MathProportion_j) + \gamma 02*(ELAProportion_j) + u_{0j} \quad (11)$$

$$\beta 1_j = \gamma 10$$

$$\beta 2_j = \gamma 20$$

where

$MathProportion_j$ and $ELAProportion_j$ are grand mean centered, and

γ_{00} is the grand mean of the log odds of graduation (or dropout) (across all schools),

γ_{01} is the between effect of the proportion of proficient math (or ELA) students with log-odds of graduation (or dropout) rates,

γ_{02} is the between effect of the proportion of proficient math (or ELA) students with log-odds of graduation (or dropout) rates,

γ_{10} is the within effect of whether a student is proficient in math (or ELA) with the log-odds of graduating (or dropping out)

γ_{20} is the within effect of whether a student is proficient in math (or ELA) with the log-odds of graduating (or dropping out).

Compositional Effects for the Hierarchical Logistic Model

The aim of this analysis is to explore whether high school math and ELA proficiency are related to graduation (or dropout) status for students in the same way that the aggregate proportion proficient is related to graduation (or dropout) rates at the school level. For math proficiency, these effects are represented by γ_{01} at the school level (the between effect) and by γ_{10} at the student level (the within effect). Therefore, the compositional effect for math proficiency can be computed by comparing the between effects to the within effects as follows: (γ_{01} - γ_{10}). Similarly, for ELA proficiency, γ_{02} represents the between effect at the school level and γ_{20} represents the within effect at the student level and the compositional effect for ELA

proficiency can be computed as (γ_{02} - γ_{20}). The compositional effects are tested to determine if they are significantly different from zero. If the compositional effect is not significantly different from zero, then this adds weight to a validity argument for using these test scores at both the individual and aggregate levels because they are behaving similarly at both levels. If the compositional effect is different from zero, then it provides evidence that the individual and aggregate scores are behaving differently.

Multinomial Model

Research question four investigates whether high school end-of-grade math and ELA tests are as good at predicting graduating senior intention survey responses at the student level as aggregate end-of-grade math and ELA proficiency are for predicting overall graduating senior intention survey responses at the high school level. The graduating senior intention survey response options are categorical (1 = employment, 2 = military, and 3 = education), therefore a hierarchical multinomial model is estimated.

In this study, further education is the referent group, *M* or category 3, and the following fully conditional model will be used to identify predictors of employment (category 1) relative to further education (category 3) and military (category 2) relative to further education (category 3) which is the reference group.

As was initially discussed with equations (4), (5), (6) and (7) the hierarchical multinomial models the log odds of being in each category relevant to a reference category therefore, in this study the multinomial model is specified accordingly.

Specifications for the level-1 multinomial model:

$$log\left[\frac{\phi_{1ij}}{\phi_{3ij}}\right] = \beta_{0j(1)} + \beta_{1j(1)} * \left(MathProficient_{ij}\right) + \beta_{2j(1)} * (ELAProficient_{ij}) \quad (12)$$

$$log\left[\frac{\phi_{2ij}}{\phi_{3ij}}\right] = \beta_{0j(2)} + \beta_{1j(2)} * \left(MathProficient_{ij}\right) + \beta_{2j(1)} * (ELAProficient_{ij})$$

where

$MathProficient_{ij}$and $ELAProficient_{ij}$ are group mean centered, and

ϕ_{1ij} is the probability that student i in school j will respond to the survey with the intention of employment after graduation.

ϕ_{2ij} is the probability that student i in school j will respond to the survey with the intention of military after graduation.

ϕ_{3ij} is the probability that student i in school j will respond to the survey with the intention of education after graduation.

$\beta_{0j(1)}$ is the mean log odds of the probability of employment versus education in school j,

$\beta_{0j(2)}$ is the mean log odds of the probability of military versus education in school j,

Specifications for the level-2 multinomial model:

$$\beta_{0(1)} = \gamma_{00(1)} + \gamma_{01(1)}*(MathProportion_j) + \gamma_{02(1)}*(ELAProportion_j) + u_{0j(1)} \quad (13)$$

$\beta_{1(1)} = \gamma_{10(1)}$

$\beta_{2(1)} = \gamma_{20(1)}$

$\beta_{0(2)} = \gamma_{00(2)} + \gamma_{01(2)}*(MathProportion_j) + \gamma_{02(2)}*(ELAProportion_j) +$

$u_{0j(2)}$

$\beta_{1(2)} = \gamma_{10(2)}$

$\beta_{2(2)} = \gamma_{20(2)}$

where

$MathProportion_j$ and $ELAProportion_j$ are grand mean centered, and

γ_{00} is the grand mean of the log odds of intention of education (across all schools),

γ_{01} is the between effect of the proportion of proficient math (or ELA) students with log-odds of intention of employment,

γ_{02} is the between effect of the proportion of proficient math (or ELA) students with log-odds of intention of military,

γ_{10} is the within effect of whether a student is proficient in math (or ELA) with the log-odds of intention of employment,

γ_{20} is the within effect of whether a student is proficient in math (or ELA) with the log-odds of intention of military.

Compositional Effects for the Multinomial Model

The aim of this research is to explore whether high school math and ELA proficiency are related to senior intention survey responses for students in the same way as aggregate senior intention survey responses at the school level. The multinomial HGLM model identifies predictors of the intention for employment or the intention to join the military relative to the referent category of further education by expanding the hierarchical logistic model.

For math proficiency, these effects are represented by γ_{01} at the school level (the between effect) and by γ_{10} at the student level (the within effect). Therefore, the compositional effect for

the intention of employment (category 1) relative to the referent group responding with an intention of further education (category 3) can be computed for math proficiency as follows: $\gamma_{01(1)} - \gamma_{10(1)}$. The compositional effect for the intention of military (category 2) relative to the referent group responding with an intention of further education (category 3) can be computed for math proficiency as follows: $\gamma_{01(2)} - \gamma_{10(2)}$. If either of these compositional effects are significantly different from zero, it indicates math proficiency is not predicting senior intention the same way at the student level as it is at the school level.

Similarly, for ELA proficiency, γ_{02} represents the between effect at the school level and γ_{20} represents the within effect at the student level. Therefore, the compositional effect for the intention of employment (category 1) relative to the referent group responding with an intention of further education (category 3) can be computed for math proficiency as follows: $\gamma_{02(1)} - \gamma_{20(1)}$. The compositional effect for the intention of military (category 2) relative to the referent group responding with an intention of further education (category 3) can be computed for math proficiency as follows: $\gamma_{02(2)} - \gamma_{20(2)}$. If either of these compositional effects are significantly different from zero, it indicates ELA proficiency is not predicting senior intention the same way at the student level as it is at the school level.

In summary, many of the tests used in accountability systems were developed to measure individual mastery of content specifications. Validity theory has evolved, and its application has become more practical and comprehensive with frameworks such as Kane's (2006, 2013) roadmap and Embretson's universal validity model. These frameworks are valuable for developing a validity argument for individual test scores, but they fall short in accommodating the reality of aggregate scoring in test-based accountability because they fail to address the need

for a validity argument for the use and interpretation of aggregate scores (Chalhoub-Deville, 2016, 2020). These validity frameworks remain primarily score-based and lack a focus on test development, the use and interpretation of aggregate scores, and the consequences of using aggregate scores in accountability systems. A validity framework that takes accountability testing purposes into account was proposed. However, empirical methodologies for building validity evidence for the use of aggregate scores have not been explored. For example, while work may have been done with regards to whether a student is proficient and how that predicts outcomes at the student level, less focus has been on aggregate scores such as percent proficient in a school and how that may be used as an accountability indicator at the school level. This research advances the theoretical work of researchers such as Chalhoub-Deville (2016, 2020), by exploring a methodology for building empirical evidence to investigate the legitimacy of a potential divide between score-based validation of individual scores and aggregate scores.

CHAPTER IV: RESULTS

This chapter describes the results from the methods and analyses selected to address these four research questions:

1. To what extent do key validity models consider accountability testing purposes where the focus is less on individual test scores, and more on aggregate scores?
2. Is there validity evidence to support the assumption that aggregate high school end-of-grade math and English language arts (ELA) scores at the school level are analogous to individual scores at the student level for predicting graduation?
3. Is there validity evidence to support the assumption that aggregate high school end-of-grade math and English language arts (ELA) scores at the school level are analogous to individual scores at the student level for predicting dropout status?
4. Is there validity evidence to support the assumption that aggregate high school end-of-grade math and English language arts (ELA) scores at the school level are analogous to individual scores at the student level for predicting graduating senior intention survey responses?

This chapter is organized in two sections. First, the results of a review and analysis of three validity models with respect to how well they address the needs of aggregate score use in accountability testing are summarized. Second, the results of the hierarchical generalized linear model (HGLM) and the compositional effects are presented.

Review and Analysis of Key Validity Models

For research question one, a literature review and analysis of key validity models was undertaken. The three validity models analyzed were Kane's IA/IUA (2006, 2013), Acree et al's

Parallel IUA (2016), and the Unified Framework by Embretson (2007, 2008, 2017). The goal of research question one was to assess the extent to which these models consider accountability testing purposes where the focus is less on individual test scores, and more on aggregate scores. The literature review established that a comprehensive validity argument for accountability systems needs to be initiated during the test development process, consequences need to be evaluated, and validity evidence for the use of aggregate scores needs to be gathered. None of the validity models fully addressed all three of these key elements. The findings are summarized in Table 8.

Table 8. Summary of Literature Review Findings for Research Question One

	Kane's IA/IUA[a]	Acree et al's Parallel IUA	Embretson's Unified Framework
Test development			✓
Consequences		✓	✓
Aggregate scores		✓	

[a]Kane (2006, 2013) acknowledges the importance of test development processes, consequences, and the use of test scores in accountability systems, however, he did not specifically address these issues in his IA/IUA.

Given that a comprehensive validation model that addresses all three of these key elements was not identified, an Accountability IUA framework was proposed (Chapter II, Figure 5). The proposed validation framework offers a roadmap for building a validity argument that promotes consideration of validity in the test development phase, outlines a parallel process for building validity evidence for aggregate scores in addition to individual scores, and considers the consequences of aggregate score use in test-based accountability systems.

Recognizing that methodologies for validation of aggregate scores have not been addressed in the literature, hierarchical generalized linear modeling (HGLM) is proposed as an

empirical method for comparing the predictive ability of individual and aggregate scores on outcome variables at the individual and aggregate level. Analyses for research questions two, three, and four demonstrate the use of HGLM as one approach for building validity evidence for the use of standardized tests in an accountability system.

Hierarchical Generalized Linear Model (HGLM) Results

Because the data were hierarchical where students are nested within schools, HLM was used to examine the predictive ability of high-school end-of-grade test scores on three non-continuous outcome variables, high school graduation, dropout, and graduating senior intention survey responses at both the individual and aggregate levels. More specifically, HGLM was used because the outcome variables were dichotomous (graduation and dropout status) or multinomial (graduating senior intention survey results). Finally, compositional effects were analyzed to compare the predictive ability of the test scores on the outcome variables at the student and school level.

It is important to reiterate that true evidence-based outcome variables for the state's accountability system were not available. In the absence of true outcome measures, cohort graduation, dropout status, and senior intention survey responses were used solely for demonstration of the empirical methodology and not to draw any conclusions about the validity of the state's accountability system or the test-based indicators used therein. The HGLM results for each of the outcome variables are discussed below.

Graduation – Hierarchical Logistic Model

The final estimation of fixed effects for the unit specific model with robust standard errors for the logistic HGLM with graduation as the outcome variable are presented in Table 9. In addition to the coefficients, the odds ratios are also included. The complete HLM output for

graduation as the outcome variable is found in Appendix A. The data were group mean centered at level-1 (the individual or student level) and grand mean centered at level-2 (the aggregate or school level). Graduating with the cohort was coded as 1 (one) and failing to graduate with the cohort was coded as 0 (zero). HLM estimated a level-1 dispersion parameter based on a binomial distribution ($\sigma^2 = 0.97$). If the assumption of no dispersion holds, σ^2 is equal to 1. If $\sigma^2 < 1$, the data are under-dispersed. Overall, at the school and student level, both math and ELA proficiency were positively associated with cohort graduation.

Table 9. Results of Logistic Model with Graduation as Outcome

Fixed Effect	Coefficient	Odds Ratio	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For Intercept1, β_{0j}						
Intercept2, γ_{00}	1.864	6.447	0.026	73.054	629	<0.001
Math Proportion Proficient, γ_{01}	0.004	1.004	0.003	1.166	629	0.244
ELA Proportion Proficient, γ_{02}	0.027	1.028	0.004	7.577	629	<0.001
For Math Proficient slope, β_1						
Intercept2, γ_{10}	0.445	1.561	0.024	18.556	120255	<0.001
For ELA Proficient slope, β_2						
Intercept2, γ_{20}	1.358	3.888	0.027	50.270	120255	<0.001

On average, for all students across all schools, the odds of graduating with the cohort were 6.4 times higher than not graduating with the cohort ($\gamma_{00} = 1.864$, $p < 0.001$). At the school level, as the proportion of students in a school who were proficient in math increased, given ELA proficiency, the proportion of students graduating with their cohort also tended to increase, but this was not a statistically significant result ($\gamma_{01} = 0.004$, $p = 0.244$). Given math proficiency, if the proportion of students in a school who were ELA proficient increased by ten percent, the odds of students graduating with their cohort increased by 0.27 percent ($\gamma_{02} = 0.027$, $p < 0.001$).

A similar pattern of results was found at the student level. Students who were proficient in math, given ELA, were 1.6 times more likely to graduate with their cohort (γ_{10} = 0.445, p < 0.001). Students who were proficient in ELA, given math, were 3.9 times more likely to graduate with their cohort (γ_{20}= 1.358, p < 0.001).

Therefore, in the typical school in which 52 percent of students were proficient in math, and 57 percent were proficient in ELA, and assuming level-2 random effects were zero, the probability of cohort graduation for a student who was proficient in both math and ELA was 0.94. If a student was proficient in ELA, but not math, the probability of graduation was 0.90. The results for math, given ELA, were not significant and therefore, math proficiency did not provide any information above and beyond ELA proficiency. Finally, if a student was not proficient in either math or ELA, the probability of graduation was 0.70.

Compositional Effects for Graduation

This analysis explored whether high school math and ELA proficiency were related to graduation in the same way at the individual student level as the aggregate proportion proficient was related to the cohort graduation rate at the school level. The compositional effect for math proficiency was computed by comparing the between effects to the within effects ($\gamma_{01} - \gamma_{10}$). The compositional effect for ELA proficiency was computed similarly ($\gamma_{02} - \gamma_{20}$). The resulting compositional effects were tested to determine if they were significantly different from zero. The compositional effects for math and ELA proficiency with the outcome variable of graduation are presented in Table 10. Both compositional effects were statistically significant indicating that there is a difference in the predictive ability of math and ELA proficiency at the student and school levels for graduation.

Table 10. Compositional Effects for Graduation as Outcome

	Level-2 Model (School) Between Effects			Level-1 Model (Student) Within Effects			Difference (Between – Within)
Graduation	Coefficient	Model se	*p*-value	Coefficient	Model se	*p*-value	Compositional Effect
Math	0.004	0.003	0.244	0.445	0.024	<0.001	-0.44*
ELA	0.027	0.004	<0.001	1.358	0.027	<0.001	-1.33*

*Compositional effect is statistically significant

Dropout – Hierarchical Logistic Model

The results for the fixed effect logistic HGLM model with dropout as the outcome variable are presented in Table 11. The data were group mean centered at level-1 (the individual student level) and grand mean centered at level-2 (the aggregate school level). Dropout was coded as 1 (one) and not dropping out was coded as 0 (zero). The complete HLM output for dropout is found in Appendix B. The HLM estimate of level-1 dispersion indicated that the dropout data are under-dispersed ($\sigma^2 = 0.62$). Given that the dropout data exhibit considerably less dispersion than expected based on a binomial distribution, these results should be interpreted with caution. Overall, at the school and student level, math and ELA proficiency were negatively associated with dropping out (i.e., proficiency was associated with staying in school).

On average, for all students across all schools, the odds of dropout were significantly lower than the odds of staying in school (γ_{00} = -4.617, $p < 0.001$). At the school level, as the proportion of students in a school who were proficient in math increased, given ELA proficiency, the proportion of dropouts in the school tended to decrease, but this was not a statistically significant result (γ_{01} = -0.006, $p = 0.391$). Given math proficiency, as the proportion of students in a school who were ELA proficient increased by ten percent, the odds of dropping out decreased by 0.41 percent (γ_{02} = -0.041, $p < 0.001$).

The results at the student level for dropout followed a similar pattern to those at the school level. Students who were proficient in math, given ELA, were about 62 percent less likely to drop out (γ_{10} = -0.957, $p < 0.001$). Students who were proficient in ELA, given math, were about 73 percent less likely to drop out (γ_{20}= -1.312, $p < 0.001$).

Table 11. Results of Hierarchical Logistic Model with Dropout as Outcome

Fixed Effect	Coefficient	Odds Ratio	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For Intercept1, β_{0j}						
Intercept2, γ_{00}	-4.617	0.010	0.082	-56.063	629	<0.001
Math Proportion Proficient, γ_{01}	-0.006	0.994	0.007	-0.859	629	0.391
ELA Proportion Proficient, γ_{02}	-0.041	0.960	0.006	-6.459	629	<0.001
For Math Proficient slope, β_1						
Intercept2, γ_{10}	-0.957	0.384	0.047	-20.458	120255	<0.001
For ELA Proficient slope, β_2						
Intercept2, γ_{20}	-1.312	0.269	0.053	-24.597	120255	<0.001

Compositional Effects for Dropout

This analysis explored whether high school math and ELA proficiency were related to dropout in the same way at the individual student level as the aggregate proportion proficient was related to the school level dropout rate. The compositional effect for math proficiency was computed by comparing the between effects to the within effects (γ_{01} - γ_{10}). The compositional effect for ELA was computed similarly (γ_{02} - γ_{20}). The resulting compositional effects were tested to determine if they were significantly different from zero. These compositional effects are presented in Table 12. Both compositional effects for dropout were statistically significant indicating that there is a difference in the predictive ability of math and ELA proficiency at the student and school levels.

Table 12. Compositional Effects for Dropout as Outcome

	Level-2 Model (School) Between Effects			Level-1 Model (Student) Within Effects			Difference (Between – Within)
Dropout	Coefficient	Model se	*p*-value	Coefficient	Model se	*p*-value	Compositional Effect
Math	-0.006	0.007	0.391	-0.957	0.047	<0.001	0.95*
ELA	-0.041	0.006	<0.001	-1.312	0.053	<0.001	1.27*

*Compositional effect is statistically significant

Graduating Senior Intentions – Multinomial Model

The results for the fixed effect multinomial HGLM with graduating senior intention as the outcome variable are presented in Table 13. The categories for the post-high school intention outcome variable are 1 = Employment, 2 = Military, and the referent category, 3 = Education. The complete output from HLM for graduating senior intention as the outcome variable is found in Appendix C. The data were group mean centered at level-1 (the individual student level) and grand mean centered at level-2 (the aggregate school level).

Intention: Employment vs. Education

On average, for all students and all schools, the odds of having a post high school intention for full-time employment were 87% lower as compared to further education ($\gamma_{00(1)} = -2.013,\ p < 0.001$). At the school level, given the proportion proficient in ELA, there was no evidence of a relationship between the proportion proficient in math and post high school intention ($\gamma_{01(1)}$ = 0.007, p = 0.197). Those schools with higher proportions of ELA proficient students, given math proficiency, had lower proportions of students with intention for full time employment, relative to intention for further education such that as the proportion of students

who were proficient in ELA increased by ten percent, the odds of the intention for employment decreased by 0.35 percent ($\gamma_{02(1)}$ = -0.035, $p < 0.001$).

Table 13. Results of Multinomial Model with Senior Intentions as Outcome

Fixed Effect	Coefficient	Odds Ratio	Standard error	*t*-ratio	Approx *d.f.*	*p*-value
For Category 1 - *Employment*[a]						
For Intercept1, $\beta_{0(1)}$						
Intercept2, $\gamma_{00(1)}$	-2.013	0.134	0.047	-42.460	629	<0.001
Math Proportion Proficient, $\gamma_{01(1)}$	0.007	1.007	0.005	1.291	629	0.197
ELA Proportion Proficient, $\gamma_{02(1)}$	-0.035	0.965	0.005	-6.923	629	<0.001
For Math Proficient slope, $\beta_{1(1)}$						
Intercept2, $\gamma_{10(1)}$	-0.663	0.515	0.023	-29.159	119621	<0.001
For ELA Proficient slope, $\beta_{2(1)}$						
Intercept2, $\gamma_{20(1)}$	-0.846	0.429	0.025	-34.282	119621	<0.001
For Category 2 - *Military*[a]						
For Intercept1, $\beta_{0(2)}$						
Intercept2, $\gamma_{00(2)}$	-3.056	0.047	0.035	-87.357	629	<0.001
Math Proportion Proficient, $\gamma_{01(2)}$	-0.003	0.997	0.004	-0.792	629	0.429
ELA Proportion Proficient, $\gamma_{02(2)}$	-0.013	0.987	0.004	-2.998	629	0.003
For Math Proficient slope, $\beta_{1(2)}$						
Intercept2, $\gamma_{10(2)}$	-0.270	0.764	0.037	-7.371	119621	<0.001
For ELA Proficient slope, $\beta_{2(2)}$						
Intercept2, $\gamma_{20(2)}$	-0.407	0.666	0.036	-11.296	119621	<0.001

[a]Referent category is further education

At the student level, math proficiency, given ELA proficiency, was associated with a lower probability of post high school intention for full time employment relative to further education ($\gamma_{10(1)}$ = -0.663, $p < 0.001$). ELA proficiency, given math proficiency, was also

associated with a lower probability of post high school intention for employment relative to further education ($\gamma_{20(1)}$ = -0.846, $p < 0.001$).

Therefore, in the typical school in which 52 percent of students are proficient in math and 57 percent are proficient in ELA, and assuming level-2 random effects are zero, students who were proficient in both math and ELA had a 0.06 probability of an intention for full time employment compared to further education. The probability of employment vs. education for students who were proficient in ELA, but not proficient in math was 0.12. At the school level, the results for math, given ELA, were not significant. The probability of employment vs. education was 0.23 for students who were not proficient in either ELA or math.

Intention: Military vs. Education

The probability of post high school intention to go into the military vs. education followed the same pattern as the probability of employment vs. education. On average, for all students and all schools, the probability of intention for military was smaller than the probability of intention for education ($\gamma_{00(2)}$ = -3.056, $p < 0.001$). There was no evidence of a relationship between a school's proportion proficient in math, given the proportion proficient in ELA, and intention for military vs. education ($\gamma_{01(2)}$ = -0.003, $p = 0.429$). Those schools with higher proportions of ELA proficient students, given math proficiency, had smaller proportions of students with a post high school intention of military, relative to further education such that as the proportion proficient in ELA increased by ten percent, the odds of graduating decreased by 0.13 percent ($\gamma_{02(2)}$ = -0.013, $p = 0.003$).

At the student level, math proficiency, given ELA proficiency, was associated with a lower probability of post high school intention for military relative to further education ($\gamma_{10(2)}$ =

-0.270, $p < 0.001$). ELA proficiency, given math proficiency, was also associated with a lower probability of post high school intention for going into the military relative to further education ($\gamma_{20(2)}$ = -0.407, $p < 0.001$).

In the typical school, a student who is proficient in both math and ELA, has a 0.03 probability of an intention for going into the military vs. further education. The probability of military vs. education for a student who is proficient in math, but not proficient in ELA, is 0.05 in the typical school. Again, the results for math, given ELA, were not significant so math proficiency does not provide any information above and beyond ELA proficiency at the school level. The probability of military vs. education for student who was proficient in ELA, but not proficient in math was 0.04 in the typical school. Finally, in the typical school, the probability of an intention for going into the military vs. an intention for further education was 0.06 for students who were not proficient in either ELA or math.

In summary, the multinomial HGLM model results offer evidence that proficiency increases the probability of a post high school intention for education relative to intention for employment or military. Proficiency increased the odds of an intention for further education.

Compositional Effects for Graduating Senior Intention

Results for the multinomial HGLM compositional effects are presented in Table 14. The multinomial HGLM model identified predictors of intention for employment or the intention to join the military relative to the referent category of further education by expanding the hierarchical logistic model and comparing category 1 (employment) with referent category 3 (education) and then also comparing category 2 (military) with referent category 3 (education).

For math proficiency, these effects were represented by γ_{01} at the school level (the between effect) and by γ_{10} at the student level (the within effect). Therefore, the compositional

effect for the intention of employment (category 1) relative to the referent group of those responding with an intention of further education (category 3) was computed for math proficiency as follows: $\gamma_{01(1)} - \gamma_{10(1)}$. The compositional effect for the intention of military (category 2) relative to the referent group responding with an intention of further education (category 3) was computed for math proficiency as follows: $\gamma_{01(2)} - \gamma_{10(2)}$. The compositional effects for ELA proficiency and employment vs education were calculated as $\gamma_{02(1)} - \gamma_{20(1)}$ and $\gamma_{02(2)} - \gamma_{20(2)}$. If either compositional effect is significantly different from zero, it indicates math proficiency is not predicting graduating senior intention the same way at the student level as it is at the school level.

Table 14. Compositional Effects for Senior Intention as Outcome

	Level-2 Model (School) Between Effects			Level-1 Model (Student) Within Effects			Difference (Between – Within)
Employment[†]	Coefficient	Model se	*p*-value	Coefficient	Model se	*p*-value	Compositional Effect
Math	0.007	0.006	0.197	-0.664	0.023	<0.001	0.67[*]
ELA	-0.036	0.006	<0.001	-0.846	0.025	<0.001	0.81[*]
	Level-2 Model (School) Between Effects			Level-1 Model (Student) Within Effects			Difference (Between – Within)
Military[†]	Coefficient	Model se	*p*-value	Coefficient	Model se	*p*-value	Compositional Effect
Math	-0.003	0.004	0.492	-0.270	0.038	<0.001	0.27[*]
ELA	-0.013	0.004	<0.001	-0.407	0.036	<0.001	0.39[*]

[†]Referent group is intention for further education after high school
[*]Compositional effect is statistically significant

Similarly, for ELA proficiency, γ_{02} represents the between effect at the school level and γ_{20} represents the within effect at the student level. Therefore, the compositional effect for the intention of employment (category 1) relative to the referent group responding with an intention

of further education (category 3) was computed for math proficiency as follows: $\gamma_{02(1)} - \gamma_{20(1)}$. The compositional effect for the intention of military (category 2) relative to the referent group responding with an intention of further education (category 3) was computed for math proficiency as follows: $\gamma_{02(2)} - \gamma_{20(2)}$. Again, if either compositional effect is significantly different from zero, it indicates ELA proficiency did not predicting graduating senior intention the same way at the student level as it did at the school level.

In summary, the compositional effect was significant for both math and ELA proficiency for all three outcome variables (graduation, dropout, and post-high school intention). A compositional effect offers evidence that the probability of predicting graduation, dropout, and post high school intention based on math and ELA proficiency at the individual student level is different from predicting the school-level proportions for graduation, dropout, and post high school intention. In other words, math and ELA proficiency did not predict the outcome variables of graduation, dropout, or intention in the same way at the individual student level as they did at the aggregate or school level.

CHAPTER V: DISCUSSION

This dissertation aims to move the needle in addressing the demands of validating test-based accountability systems. Validity frameworks need to be reconceptualized in consideration of test-based accountability systems where scores are aggregated to measure the performance of teachers, administrators, and schools (Chalhoub-Deville, 2020). Reform-driven policies such as NCLB and ESSA mandate and attach consequences to schools and teachers based on aggregates and derivatives of student test scores. Therefore, consequences are inextricably tied to the validation of test use and interpretation in accountability. Lastly, empirical methodologies for building validity evidence are required to support the use of aggregate or derivative scores at the school level.

This chapter is divided into five sections: a summary of key findings, a discussion of implications, recognition of limitations, directions for future research, and conclusions.

Summary of Key Findings

As a review, the research questions are:

1. To what extent do key validity models consider accountability testing purposes where the focus is less on individual test scores, and more on aggregate scores?
2. Is there validity evidence to support the assumption that aggregate high school end-of-grade math and English language arts (ELA) scores at the school level are analogous to individual scores at the student level for predicting graduation?
3. Is there validity evidence to support the assumption that aggregate high school end-of-grade math and English language arts (ELA) scores at the school level are analogous to individual scores at the student level for predicting dropout status?

4. Is there validity evidence to support the assumption that aggregate high school end-of-grade math and English language arts (ELA) scores at the school level are analogous to individual scores at the student level for predicting graduating senior intention survey responses?

The first research question explored the extent to which key validity models consider accountability testing purposes. Researchers and validity theorists have argued that a comprehensive validity argument needs to include both test development and measurement evidence (Chalhoub-Deville, 2020; Chalhoub-Deville & O'Sullivan, 2020), documentation of consequences (e.g., Chalhoub-Deville, 2016, 2020; Embretson, 2007, 2008, 2017; Kane, 2006, 2013), and consideration for the validation of aggregate-level data in addition to individual student level data (Chalhoub-Deville, 2020, p.254). Notably, the *Standards* (2014) limit the role of test developers in these areas (Chalhoub-Deville, 2020) and none of the three validity models reviewed, Kane's IA/IUA (2006, 2013), Acree et al's Parallel IUA (2016), nor the Unified Framework by Embretson (2007, 2008) address all these demands of accountability testing.

To address the gaps in these validity models, the proposed Accountability IUA offers a systematic approach for building a validity argument beginning in the test design and development phase, includes a parallel process for building validity evidence for aggregate scores, and considers the consequences of accountability systems. While there are calls for the evaluation of aggregate scores used in test-based accountability systems, the validity literature is lacking a discussion of empirical methodologies for doing so.

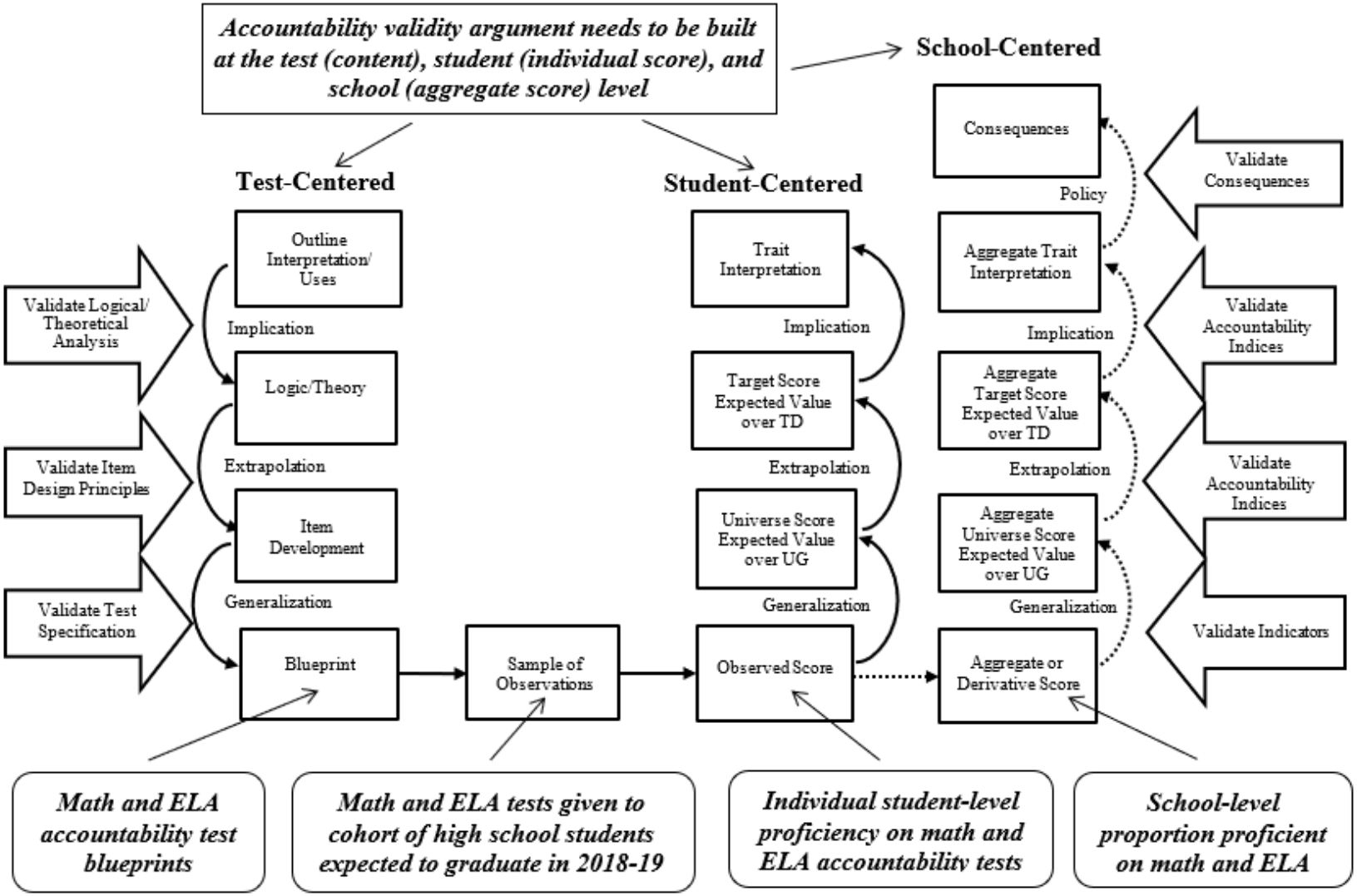
In choosing an empirical methodology for accountability systems, it is helpful to remember that student performance does not occur in isolation. Rather, it exists as part of a series of nested and hierarchical effects. Focusing on individual student scores alone does not account

for the effects of the classroom or school in which the student is found. It is possible to simply aggregate the data at the school level, using mean test scores for example, and then perform linear regression, but this approach ignores within group variation and interactions between the students. Independence, an assumption of linear regression, is often violated in accountability data because students in schools have shared experiences and, as a result, they tend to be more like each other than they are like students from different schools.

An empirical method that lends itself to the comparison of individual and group-level outcomes is hierarchical generalized linear modeling (HGLM). HGLM allows an analysis of effects at both the student and school level in addition to the interactions across levels. HGLM is particularly useful for contextual analysis or multilevel modeling. Contextual analysis allows the researcher to explain variation of individual-level performance or behaviors in terms of individual effects within a school and group-level effects between schools. The ability to simultaneously evaluate the relationship between test scores and outcome variables at both the student and school level makes HGLM particularly appealing as an empirical methodology for accountability data.

Application of the Accountability IUA framework for research questions two through four is depicted in Figure 6. The validity argument is addressed in the test development phase (in the test-centered branch), at the student level (in the student-centered branch), and at the school level (in the school-centered branch). Objectives as well as positive and negative consequences should be anticipated when scores are aggregated for accountability purposes.

Figure 6. Application of the Accountability IUA



Research questions two through four demonstrate the use of HGLM in building validity evidence to support the assumption that aggregate high school end-of-grade math and English language arts (ELA) scores predict cohort graduation, dropout status, and graduating senior intention at the school level in the same way that they do at the student level. The *observed scores* or independent variables in this study were dichotomous – proficiency on accountability math and ELA tests. The tests were administered to the cohort of high school students in a southeastern state who were expected to graduate in the 2018-19 school year. The *aggregate or derivative scores* are the proportion of students who are proficient on the math and ELA tests in each high school. An assumption was made that the math and ELA tests being used in the accountability system are valid and reliable predictors of high school graduation, high school dropout, and graduating senior intention at the individual student level. It bears repeating that the outcome variables used in this demonstration of the methodology are not true evidence-based outcome variables for the accountability system. Such outcome variables were not available; therefore, the outcome variables in this analysis were chosen to demonstrate the methodology and not to draw any conclusions about the validity of the state's accountability system or the test-based indicators used therein.

Overall, the results indicate that at both the student and school level, proficiency on the math and ELA tests used in the accountability system is associated with a greater likelihood of graduation, staying in school (not dropping out), and the intention for further education after high school graduation (as compared to going into the military or seeking full-time employment). However, the association for math proficiency (given ELA proficiency) was not statistically significant at the school level. The compositional effects comparing the strength of the association of proficiency at the individual student level with that at the school level for the

outcome variables are the analysis of interest in building validity evidence for aggregate scores used in accountability systems. The significant compositional effect for all three outcome variables indicates that if these were true evidence-based outcome measures for the accountability system and were valid predictors of individual student success and college and career readiness, then these results would not support the use of the aggregate proportion proficient at the school level as an accountability system indicator of college and career readiness or teacher and school success.

Implications

Historically, validity has been framed in terms of empirical methodologies (Chalhoub-Deville & O'Sullivan, 2020). The development of the correlation coefficient by Pearson in the late 1800s shaped the discussion around criterion-related validity. Current thinking on validity evolved from this discussion of validity concomitantly with an empirical methodology. As per this tradition in the literature, this dissertation proposes a new way to conceptualize validity in accountability systems and links the validation of aggregate scores to the methodology of hierarchical generalized linear modeling (HGLM) and hierarchical linear modeling (HLM), a special case of HGLM. While other methodological approaches could be utilized, HGLM and HLM are uniquely suited to deal with nested data such as students within schools. Focusing on individual student scores does not account for the effects of the school in which the student is nested and simply aggregating data to the school level using mean test scores and then performing linear regression ignores within group variation and interactions between the student and school. The advantage of HGLM and HLM is that it allows evaluation of the relationship between test scores and outcome variables at both the student and school level simultaneously. As a result, a direct comparison, called a compositional effect, can be tested for differences

between the predictive ability of the individual and aggregate measures on the outcomes. For these reasons HGLM and HLM are an ideal empirical methodology for building validity evidence for aggregate scores.

This dissertation helps bridge the gap between theory and operationalization of the use of aggregate scores in accountability systems. The Accountability IUA offers a reconceptualization of validity frameworks to account for the demands of accountability systems where aggregate scores are used as measures of the success of teachers and schools in educating students.

Test Development

Operationalizing the proposed Accountability IUA requires education test developers to consider the potential for aggregation of test scores and lay a foundation for an accountability validity argument in the test development phase. Potential interpretations and uses at the student or aggregate level should be anticipated. Principled assessment design approaches like evidence-centered design (ECD) guide test developers in articulating the chain of reasoning that links evidence to claims about target constructs. Logic and theory guides test specification and item development. However, even when using tests and indicators that were not originally designed to be aggregated for accountability systems, the HGLM or HLM methodology still offers an opportunity to retroactively build validity evidence for the aggregated indicators in an accountability system or to build a case for modifying the accountability model if the validity argument is weak.

Consequences

Bachman and Palmer's (2010) Assessment Use Argument (AUA), the preeminent validity model in language testing, merits highlighting for its unique focus on consequences as the basis for the validation argument in the test development and design phase. However,

Chalhoub-Deville (2020) advocates for investigating consequences at the policy-making stage before test development begins. She distinguishes this approach "from what is proposed by Bachman and Palmer (2010) where consequences are considered at the beginning of test development, after a policy has been finalized and rolled out" (p. 257). In the Accountability IUA, implications and consequences are evaluated as part of the use of aggregate scores when educational reform policies are being defined. Consideration should be given to maximizing the opportunity to meet the objectives of the accountability system while anticipating and minimizing unintended negative consequences. Theory of Action (TOA) offers a framework within which to consider the impact of accountability systems on students, teachers, and schools. According to Chalhoub-Deville (2020, p. 259),

> The use of frameworks such as TOA invites systematic and anticipatory research that can help us move beyond traditional, individual-focused test scores and related technical quality documentation. Such frameworks can help us attend to actual desired socio-educational goals embedded in a policy (or a client's request) and address research into unintended outcomes.

Future editions of the *Standards* need to hold test developers accountable for anticipating consequences of the interpretation and use of their tests at the individual and aggregate level in accordance with the zone of negotiated responsibility (ZNR) described by Chalhoub-Deville (2016) (see Figure 1).

Aggregate Scores

Often the use of aggregate scores and the design of accountability systems is a policy decision. This means test developers must engage with policymakers regarding their testing needs and their goals for interpretation and use of tests. Test developers cannot monitor, or

control all uses of their tests, but as Chalhoub-Deville (2020) describes, "[t]Est providers create tests with *some* understanding of the consequences entailed by the testing program, but they are reluctant to engage in validation to uphold those consequences" (pp. 255-256, emphasis in original). Historically test developers have hidden behind test specifications and the interpretations and uses laid out therein to absolve themselves of responsibility for unintended consequences of the use of their tests beyond the originally defined scope. Knowing that test scores may be aggregated and used in accountability systems and that accountability systems impose consequences such as sanctions and rewards, obligates education test developers to consider the consequences of such use.

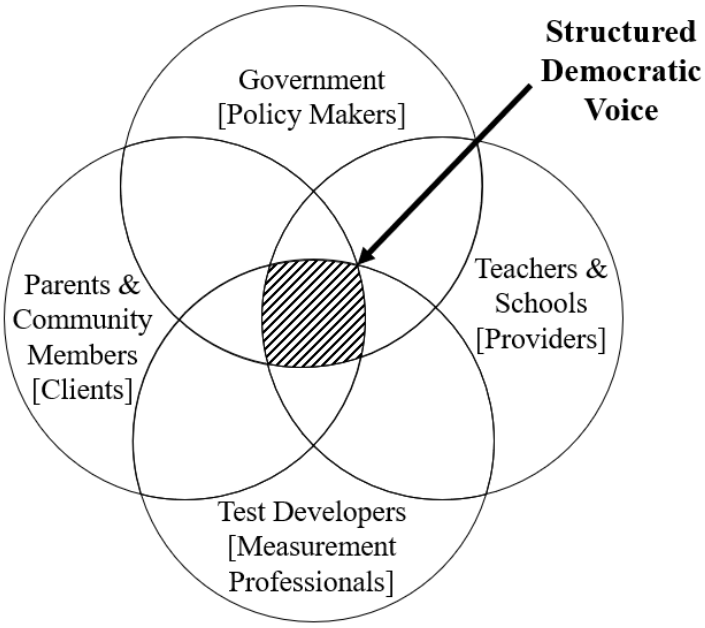
Roles and Responsibilities of Test Developers, Test Users, and Policy Makers

Sireci and Forte (2012) point out that tests are at the center of accountability and that "the use of tests is initiated and mandated by policy makers" (p. 27). Elected officials such as general assemblies, governors, and politically motivated and directed chiefs of state education have become the decision makers in the design of many state accountability systems. Decisions regarding what indicators are included in an accountability system and how they are calculated is often a political conversation as evidenced by the fact that state statute often regulates the definition of accountability systems (Education Commission of the States, 2018). Sireci and Forte (2012) argue that "it is an ethical imperative for the measurement community to do all we can to inform policy makers of the strengths, benefits, and limitations of educational tests" (p. 27). Testing programs and accountability systems necessitate "[c]ommunication and engagement with policy makers, education professionals, and other key stakeholder groups beyond the measurement community" (Chalhoub-Deville, 2020, p. 245). A one-page document on designing accountability systems is found in Appendix D. This document can be shared with state

education agency staff or others who work with education policy makers to help guide them in developing a Theory of Action before determining indicators to be used in the accountability system.

Smith and Benavot (2019) have argued that discussions of accountability exclude the "voices of stakeholders who work, learn, and teach in schools and other educational institutions" (p. 193). They advocate for the inclusion of these stakeholders, particularly in discussions of planning and evaluation through what they have labeled "structured democratic voice." Their "collaboration for structured democratic voice" diagram (Smith & Benavot, 2019, p. 202) offers a useful vision for the engagement of stakeholders, however, they have not included the measurement community in that collaboration. A modification to their depiction of a "collaboration for structured democratic voice" to include the measurement community is shown in Figure 7.

Figure 7. Adapted from "Improving accountability in education: The importance of structured democratic voice," W. C. Smith and A. Benavot, 2019, *Asia Pacific Education Review, 20*, p. 202. CC BY 4.0.



Writing to the measurement community, Sireci (2019) said, "[e]ducation policy makers, state and local department of education staff, superintendents, principals, and teachers are all involved in educational testing. It is time for us to get involved with them." Researchers and experts in measurement have an obligation to step outside of theory, and perhaps outside of their comfort zone, to engage with and inform test users and policy makers. Future editions of the *Standards* need to hold test developers accountable for engaging in this collaboration. Chalhoub-Deville's zone of negotiated responsibility (ZNR) offers guidance for test developers, test users, and policy makers to engage in meaningful discussion of the shared responsibility for outcomes and consequences (2016, 2020). Further guidance is given by Sireci and Forte (2012) who "discuss the types of information that are important to communicate to policy makers, how to best convey this information in a manner in which it can be understood, and how to be seen as a valuable source of information to education policy makers" (p. 27).

Design of Accountability Systems

An effective accountability system requires resources and thoughtful planning. States are often limited by the data they have available and the tests they already administer. Evidence-based measures of the intended outcomes of accountability systems must be defined and researched. Without evidence-based outcome variables of college and career readiness and school quality and success, it is not possible to determine if education reform is working. Testing is integral to education as a measure of learning and achievement, but in accountability it is the aggregation of student performance that is intended to reflect on or represent the success of a teacher or school in developing college and career ready high school graduates. There are three main reasons why tests are heavily relied upon by education policy makers for accountability. First, they are perceived to be objective and quantifiable measures of student achievement;

second, relatively speaking, they are inexpensive; and third, often there are few other options readily available (Sireci & Forte, 2012, p. 27).

The direct consequences to students for poor performance on accountability tests are often minimal and, in many cases, students are held harmless for their test performance while teachers and schools are sanctioned or rewarded based on aggregate scores. ESSA offers some flexibility and opportunity to use indicators other than tests in accountability systems. Examples include student growth measures, early warning indicators such as chronic absenteeism; being on track to graduate (Martin, Sargrad, & Batel, 2016); and measures of student engagement such as eye tracking (Kaakinen, 2021). Regardless of the indicators used, if they are student level measures that are being aggregated at the teacher or school level, validation studies are needed for that aggregation. HGLM or HLM are empirical methodologies that can be applied to any aggregate or derivative measure as long as there is an evidence-based outcome measure.

Limitations

The goal of this research was to offer a validation framework and methodology for building validity evidence for the use of aggregate scores in accountability systems. One limitation in this study is the absence of post high school, evidence-based outcome measures of college and career readiness or evidence-based outcome measures of school quality and success. A second limitation is the oversimplicity of the HGLM models presented for the purposes of demonstrating the methodology.

The HGLM or HLM analysis should use the same aggregate or derivative form of the variables as are used for the indicators in the accountability model. Therefore, this study used proficiency on math and ELA tests from the state's accountability model as the independent

variables. It is possible that using actual test scores, rather than proficiency, may give different results because test scores are continuous, and proficiency is dichotomous.

States need to clearly define the intended outcomes and objectives as they build their accountability systems. Once equipped with clear and measurable definitions of college and career readiness and school quality and success, states need to collect evidence-based outcome measures for their accountability systems to establish the validity of the indicators and the calculation of indices used in their accountability systems. It is not enough to establish that an indicator is valid at predicting an outcome at the individual student level because accountability systems change the unit of measurement from the individual student level to a group level by aggregating scores at the school or teacher level. Without evidence-based outcome measures it is not possible to build a sound validity argument for accountability systems.

Furthermore, building validity arguments for aggregate scores requires addressing other factors that may influence test scores at both the individual and school level such as socioeconomic status, English learner (EL) status, and race or ethnicity. HGLM and HLM allow for the inclusion of these factors to build evidence that the accountability system is a valid and equitable measure of school success. HGLM and HLM also allow for the parsing of results by subgroup to determine if aggregate scores are equally valid indicators across groups, however because this is not part of the calculation of school performance in this state's accountability system, it was not included. The state in which this accountability test data was collected is one of twelve states that do not include subgroups in school performance calculations (Hunt Institute, 2019). According to Alliance for Education (2018, emphasis added) this is not in compliance with ESSA law:

> The Every Student Succeeds Act (ESSA) is a civil rights law that works to ensure states provide all children with equal access to a high-quality education...Despite this legal mandate, many states fail to include student subgroups *meaningfully* across two of the law's most important accountability provisions: (1) school ratings and (2) the definitions used to identify schools for TSI [Targeted Support and Improvement].

The backdrop of society at the time of the writing of this dissertation was one in which "[s]ocial justice has reached more urgent and heightened importance this past year, with overdue attention and national discussions" (Tong, 2021). As president of NCME, Tong wrote, "I also believe that as measurement professionals, we need to continue to support unbiased, equitable, and fair assessments…We owe it to society to put our talents together to create culturally responsive, equitable, and fair assessments for all populations." Measurement professionals must engage and help guide states in designing and implementing unbiased, equitable, and fair accountability systems that identify opportunities to address injustice in our educational systems.

Future Research

Predictive criterion-related validity, defined as "evidence that a test score or other measurement correlates with a variable that can only be assessed at some point after the test has been administered or the measurement made (American Psychological Association (n.d.)," has been applied extensively in college admissions and employment testing (e.g., Burrus, Way, Bobek, Stoeffler, & O'Connor, 2020; Nichols-Barrer, Place, Dillon, & Gill, 2016; Rogelberg, 2008; Robbins, Lauver, Le, & Davis, 2004; Roth, BeVier, Switzer, & Schippman, 1996; Guion, 2011), however, it has not been applied in accountability. Future research needs to address the limitations of this study by identifying and using evidence-based criterion or outcome measures for accountability systems that represent the success of schools and teachers in educating

students as defined by the stakeholders. Stakeholders include school administrators, teachers, and the community at large such as businesses who may hire high school graduates, and higher educational institutions who may enroll them.

Additionally, HGLM/HLM models should replicate the accountability model as closely as possible including all aggregated indicators in the accountability system. While HGLM/HLM can analyze the variables in any form (e.g., continuous, binary, count, ordinal, or multinomial values), they need to be input in HGLM/HLM in the same form that they are used in the accountability model. If the HGLM/HLM results do not offer validity evidence for the indicators in the form they are currently used in the accountability system, then exploration of the use of other aggregates or derivatives may be warranted to see if outcomes change and if compositional effects are eliminated. For example, if the accountability system assigns school performance grades, are schools assigned the same grade when scaled scores are used as they are when proficiency status or achievement levels are used? Are compositional effects still found?

Further research is needed to determine the most meaningful and informative indicators for accountability systems. ESSA (2015) requires state accountability systems to include these five indicators: achievement on annual reading/language arts and mathematics assessments which may include growth in high school; growth in grades below high school or another academic indicator, high school graduation rates, progress of English language learners toward proficiency, and a non-academic indicator of school quality or student success. Differentiation is needed between indicators or independent variables and outcomes or dependent variables such a graduation.

ESSA requires both math and ELA scores, however, the high correlation between them means that including both may not be providing any additional information. The relationship

between English language proficiency and reading comprehension on math performance is established (e.g., Adelson, Dickinson, & Cunningham, 2015; Beal, Adams, & Cohen, 2010; Fuchs, Fuchs, Compton, Hamlett, & Wang, 2015; Kieffer, Lesaux, Rivera, & Francis, 2009). Language skills feature prominently in solving math word problems (Abedi & Lord, 2001; Vilenius-Tuohimaa, Aunola, & Nurmi, 2008). The results of the HLM analyses in this dissertation showed that math, given ELA, was not a significant predictor of the outcome variables. If performance on a math test is overly dependent on reading comprehension, then reading comprehension becomes a construct-irrelevant factor, because the definition of math proficiency does not include reading comprehension (Haladyna & Downing, 2004). Further research is needed to address construct irrelevance and to develop valid indicators for accountability systems.

Conclusion

The proposed IUA Framework for Accountability (Figure 6) maintains not only that a test may be valid at the individual level and not at the aggregate level, but also that a test may be valid at the aggregate level and not at the individual level, depending on the use and interpretation. If tests are going to be used at both the individual and the aggregate level with the same interpretation, then the individual and aggregate scores must both be validated for that interpretation and use. Methods other than compositional effects are needed to build validity evidence for tests designed specifically for accountability systems that are not also used as a measure of college and career readiness at the individual student level. A strength of the proposed IUA Framework for Accountability systems is that it allows for separate and parallel validation of individual and aggregate scores. Similarly, a strength of HGLM/HLM is the methodology's ability to parse out how well individual or aggregate scores predict outcome

variables while accounting for variations among students nested in schools and controlling for other factors such as race, EL status, school locale (e.g., urban vs rural), etc. This dissertation provides a validity framework for the validation of accountability systems and demonstrates an empirical methodology for building validity evidence for the use of aggregate scores.

REFERENCES

Abedi, J., & Lord, C. (2001). The Language Factor in Mathematics Tests. *Applied Measurement in Education*, 14(3), 219-234.

Acree, J., Hoeve, K.B., Weir, J.B. (2016). Approaching the validation of accountability systems. Unpublished paper and presentation. ERM 600: Validity and Validation, University of North Carolina at Greensboro.

Adelson, J. L., Dickinson, E. R., & Cunningham, B. C. (2015). Differences in the reading–mathematics relationship: A multi-grade, multi-year statewide examination. *Learning and Individual Differences*, 43, 118–123.

Alliance for Education. (2018). *Too Many States Minimize Student Subgroup Performance in ESSA Accountability Systems*. Retrieved from https://all4ed.org/wp-content/uploads/2018/09/ESSA-Subgroup-Performance-State-Accountability-Systems.pdf.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1966). *Standards for educational and psychological testing*. Washington, DC: Author.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1974). *Standards for educational and psychological testing*. Washington, DC: Author.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: Author.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

American Psychological Association. (n.d.). *APA Dictionary of Psychology*. American Psychological Association. Retrieved October 3, 2021, from https://dictionary.apa.org/predictive-validity.

American Psychological Association, & American Educational Research Association. National Council on Measurements Used in Education (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Psychological Bulletin, 51(2), 1-38.

Appropriations Act. (2016). General Assembly of North Carolina. Session 2015. HB1030. https://www.ncleg.net/sessions/2015/bills/house/html/h1030v8.html

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

Bandalos, D. L., Ferster, A. E., Davis, S. L., & Samuelsen, K. M. (2011). Validity arguments for high-stakes testing and accountability systems. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K–12 settings* (p. 155–175). American Psychological Association.

Beal, C. R., Adams, N. M., & Cohen, P. R. (2010). Reading proficiency and mathematics problem solving by high school English language learners. *Urban Education*, *45*(1), 58–74.

Bennett, R., (2015). Validity Considerations for Next-Generation Assessment: A "Theory of Action" Perspective. Paper presented at National Conference on Student Assessment, San Diego, CA.

Burrus, J., Way, J., Bobek, B., Stoeffler, K., & O'Connor, R. (2020). The ACT Holistic Framework® of Education and Workplace Success. In M. Oliveri & C. Wendler (Eds.), *Higher Education Admissions Practices: An International Perspective* (Educational and Psychological Testing in a Global Context, pp. 307-332). Cambridge: Cambridge University Press. doi:10.1017/9781108559607.017.

Carey, G. (2013). The General Linear Model (GLM): A gentle introduction. *Quantitative Methods in Neuroscience.* Retrieved from: http://psych.colorado.edu/~carey/qmin/qminChapters/QMIN09-GLMIntro.pdf.

Chalhoub-Deville, M. (2009). The Intersection of Test Impact, Validation, and Educational Reform Policy. *Annual Review of Applied Linguistics, 29*, 118-131. doi:http://dx.doi.org/10.1017/S0267190509090102.

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

Chalhoub-Deville, M. B. (2020). Toward a model of validity in accountability testing. *Assessing English language proficiency in US K–12 schools. New York, NY: Routledge*.

Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity: Theoretical Development and Integrated Arguments*. Equinox Publishing Limited.

Cizek, G.J. (2016). Validating test score meaning and defending test score use: different aims, different methods. *Assessment in Education: Principles, Policy & Practice*, *23*, 212-225.

Council of Chief State School Officers (CCSSO). (2004). *A Framework for Examining Validity in State Accountability Systems.* Washington, DC: Council of Chief State School Officers.

Council of Chief State School Officers (CCSSO). (2016). *Exploring College and Career Readiness Indicators: A webinar series about state accountability systems under ESSA.* Washington, DC: Council of Chief State School Officers.

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

Education Commission of the States. (2018). *50-State Comparison: States' School Accountability Systems.* Retrieved from https://www.ecs.org/50-state-comparison-states-school-accountability-systems/.

Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher*, *36*(8), 449-455.

Embretson, S. (2008). Construct Validity: *A Universal Validity System [PowerPoint Slides]*. Retrieved from https://marces.org/conference/validity/8Susan%20Embretson.ppt

Embretson, S. (2017). An integrative framework for construct validity. *The Handbook of Cognition and Assessment, Frameworks, Methodologies and Applications*, 102-123.

Every Student Succeeds Act (ESSA), 20 U.S.C. § 6301 (2015). Retrieved from https://www.congress.gov/bill/114th-congress/senate-bill/1177.

Ferrara, S., Lai, E., Reilly, A., & Nichols, P. D. (2017). Principled approaches to assessment design, development, and implementation. *The Handbook of Cognition and Assessment, Frameworks, Methodologies and Applications*, 41-74.

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

Fuchs, L. S., Fuchs, D., Compton, D. L., Hamlett, C. L., & Wang, A. Y. (2015). Is Word-Problem Solving a Form of Text Comprehension? *Scientific Studies of Reading: The official journal of the Society for the Scientific Study of Reading*, 19(3), 204–223.

Guion, R. M. (2011). *Assessment, Measurement, and Prediction for Personnel Decisions* (2nd ed.). Routledge.

Haladyna, T. M., & Downing, S. M. (2005). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Hunt Institute. (2019) *School Accountability and Performance Grades Issue Brief*. Retrieved from http://www.hunt-institute.org/wp-content/uploads/2019/04/HI-NC-SAA-ISSUEBRIEF-0419-FULL.pdf.

Im, G.H., Shin, D. & Cheng, L. (2019). Critical review of validation models and practices in language testing: their limitations and future directions for validation research. *Language Testing in Asia, 9*(14), Retrieved from https://doi.org/10.1186/s40468-019-0089-4.

Kaakinen, J. K. (2021). What can eye movements tell us about visual perception processes in classroom contexts? Commentary on a special issue. *Educational Psychology Review*, *33*(1), 169-179.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Westport, CT: Greenwood Publishing.

Kane, M. (2010). Validity and fairness. *Language Testing, 27*(2), 177-182.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3-17.

Kane, M. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1-73.

Kane, M. (2015). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In *Handbook of test development* (pp. 80-96). Routledge.

Kane, M. (2020) Validity Studies Commentary, Educational Assessment, 25:1, 83-89, DOI: 10.1080/10627197.2019.1702465.

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English Language Learners Taking Large-Scale Assessments: A Meta-Analysis on Effectiveness and Validity. *Review of Educational Research, 79*(3), 1168-1201.

Lee, Y., & Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(4), 619–678.

Lissitz, R.W. & Samuelsen. K. (2007). A Suggested Change in Terminology and Emphasis Regarding Validity and Education. *Educational Researcher, 36*, 437-448.

Martin, C., Sargrad, S., & Batel, S. (2016). *Making the Grade: A 50-State Analysis of School Accountability Systems*. Retrieved from https://files.eric.ed.gov/fulltext/ED567858.pdf.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 13-103). New York: Macmillan.

Moss, P. (2007). Reconstructing validity. Educational Researcher, 36, 470-476.

Moss, P. (2013). Validity in action: lessons from studies of data use. *Journal of Educational Measurement*, *501*(1), 91-98.

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. Sage.

Nichols-Barrer, I., Place, K., Dillon, E., & Gill, B. (2016). Testing college readiness: Massachusetts compares the validity of two standardized tests. *Education Next*, *16*(3), 70-77. Retrieved October 3, 2021 from https://www.educationnext.org/testing-college-readiness-massachusetts-parcc-mcas-standardized-tests/.

No Child Left Behind Act of 2001 (NCLB), Pub. L. No. 107–110, 115 Stat. 1425 (2002). Retrieved from https://www2.ed.gov/policy/elsec/leg/esea02/index.html.

Lane, S. (1999). Validity evidence for assessments. *Reidy interactive lecture series. Pittsburgh, PA: University Pittsburgh*.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–7.

Plake, B. S., Huff, K., Reshetar, R. R., Kaliski, P., & Chajewski, M. (2015). Validity in the making: From evidenced-centered design to the validations of the interpretations of test performance. In Faulkner-Bond, M. & Wells, C. (Eds). *Educational measurement: Foundations to future* (p. 62-73).

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S.W., Bryk, A.S, Cheong, Y.F. & Congdon, R. (2019). HLM 8 for Windows [Computer software]. Skokie, IL: Scientific Software International, Inc.

Riconscente, M. M., Mislevy, R. J., & Corrigan, S. (2016). *Evidence-centered design.* In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (p. 40–63). Routledge/Taylor & Francis Group.

Robbins, S. B., Lauver, K., Le, H., & Davis, D. (2004). Do psychosocial and study skill factors predict college outcomes? a meta-analysis. *Psychological Bulletin*, *130*(2), 261–288. https://doi.org/10.1037/0033-2909.130.2.261.

Rogelberg, S. G. (Ed.). (2008). *Handbook of research methods in industrial and organizational psychology* (Vol. 5). John Wiley & Sons.

Roth, P. L., BeVier, C. A., Switzer, F. S. I. I. I., & Schippman, J. S. (1996). Meta-analyzing the relationship between grades and job performance. *Journal of Applied Psychology*, *81*(5), 548–548.

Sahlberg, P. (2012). *Global educational reform movement is here!* Retrieved from https://pasisahlberg.com/global-educational-reform-movement-is-here/.

Sahlberg, P. (2014). *Finnish lessons 2.0: What can the world learn from educational change in Finland?* Teachers College Press.

Shepard, L. A. (2016). Evaluating test validity: reprise and progress. *Assessment in Education, 23*, 2, 268-280. Amherst, MA: Center for Educational Assessment, University of Massachusetts.

Shaw, S., & Crisp, V. (2011). Tracing the evolution of validity in educational measurement: Past issues and contemporary challenges. *Research Matters: A Cambridge Assessment Publication*, *11*, 14-17.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In *The Concept of Validity: Revisions, New Directions and Applications, Oct, 2008*. IAP Information Age Publishing.

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*(1), 100–7. https://doi.org/10.7334/psicothema2013.256.

Sireci, S. G., & Soto, A. (2016, January). Test Validation for 21st-Century Educational Assessments. In *Meeting the challenges to measurement in an era of accountability*. Routledge.

Sireci, S. G. (2019). *From the President: You, Me, and NCME!* Retrieved from https://www.ncme.org/blogs/megan-welsh1/2019/06/30/you-me-and-ncme.

Sireci, S. G. (2020). De-"constructing" test validation. *Chinese/English Journal of Educational Measurement and Evaluation*, *1*(1), 3.

Smith, W. C., & Benavot, A. (2019). Improving accountability in education: the importance of structured democratic voice. *Asia Pacific Education Review*, *20*(2), 193-205.

Tong, Y. (2021). *NCME President's Letter*. Retrieved from https://www.ncme.org/news/president-message.

U.S. Department of Education, (2009). Race to the Top fund. Retrieved from https://www2.ed.gov/programs/racetothetop/index.html.

Vilenius-Tuohimaa, P. M., Aunola, K., & Nurmi, J. E. (2008). The association between mathematical word problems and reading comprehension. *Educational Psychology*, 28(4), 409–426.

Wilson, M. (2008). *Constructing measures: An item response modeling approach*. Mahwah, N.J: Lawrence Erlbaum Associates.

Program: HLM 8 Hierarchical Linear and Nonlinear Modeling
Authors: Stephen Raudenbush, Tony Bryk, & Richard Congdon
Publisher: Scientific Software International, Inc. (c) 2019
hlm@ssicentral.com
www.ssicentral.com

Module: HLM2.EXE (8.0.2010.18)
Date: 14 February 2021, Sunday
Time: 19:12:37
License:

Specifications for this Overdispersed Bernoulli HLM2 run

The maximum number of level-1 units = 120889
The maximum number of level-2 units = 632
The maximum number of micro iterations = 14
Method of estimation: full PQL

Maximum number of macro iterations = 100

Distribution at Level-1: Bernoulli

The outcome variable is GRAD

Summary of the model specified

Step 2 model

Level-1 Model

$\text{Prob}(GRAD_{ij}=1|\beta_j) = \phi_{ij}$
$\log[\phi_{ij}/(1 - \phi_{ij})] = \eta_{ij}$
$\eta_{ij} = \beta_{0j} + \beta_{1j}*(MATHPROF_{ij}) + \beta_{2j}*(ELAPROF_{ij})$

Level-2 Model

$\beta_{0j} = \gamma_{00} + \gamma_{01}*(MATHPROP_j) + \gamma_{02}*(ELAPROP_j) + u_{0j}$
$\beta_{1j} = \gamma_{10}$
$\beta_{2j} = \gamma_{20}$

MATHPROF ELAPROF have been centered around the group mean.
MATHPROP ELAPROP have been centered around the grand mean.

Level-1 variance = $\sigma^2/[\phi_{ij}(1-\phi_{ij})]$

Mixed Model

$\eta_{ij} = \gamma_{00} + \gamma_{01}*MATHPROP_j + \gamma_{02}*ELAPROP_j$
$+ \gamma_{10}*MATHPROF_{ij}$
$+ \gamma_{20}*ELAPROF_{ij}$
$+ u_{0j}$

The value of the log-likelihood function at iteration 11 = -4.447076E+04

Results for Non-linear Model with the Logit Link Function

Unit-Specific Model, PQL Estimation - (macro iteration 7)

$\sigma^2 = 0.96650$

Standard error of $\sigma^2 = 0.00394$

τ

INTRCPT1,β_0	0.29810

Standard error of τ

INTRCPT1,β_0	0.02184

Approximate confidence intervals of tau variances
INTRCPT1 : (0.258,0.344)

Random level-1 coefficient	Reliability estimate
INTRCPT1,β_0	0.729

The value of the log-likelihood function at iteration 2 = -1.700150E+05

Final estimation of fixed effects: (Unit-specific model)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	1.863650	0.025790	72.263	629	<0.001
MATHPROP, γ_{01}	0.004062	0.002591	1.568	629	0.117
ELAPROP, γ_{02}	0.027394	0.002534	10.809	629	<0.001
For MATHPROF slope, β_1					
INTRCPT2, γ_{10}	0.445129	0.019755	22.533	120255	<0.001
For ELAPROF slope, β_2					
INTRCPT2, γ_{20}	1.357849	0.020038	67.763	120255	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	1.863650	6.447229	(6.129,6.782)
MATHPROP, γ_{01}	0.004062	1.004070	(0.999,1.009)
ELAPROP, γ_{02}	0.027394	1.027772	(1.023,1.033)
For MATHPROF slope, β_1			
INTRCPT2, γ_{10}	0.445129	1.560691	(1.501,1.622)
For ELAPROF slope, β_2			
INTRCPT2, γ_{20}	1.357849	3.887822	(3.738,4.044)

Results of General Linear Hypothesis Testing - Test 1

	Coefficients	Contrast	
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	1.863650	0.0000	0.0000
MATHPROP, γ_{01}	0.004062	1.0000	0.0000
ELAPROP, γ_{02}	0.027394	0.0000	1.0000
For MATHPROF slope, β_1			
INTRCPT2, γ_{10}	0.445129	-1.0000	0.0000
For ELAPROF slope, β_2			
INTRCPT2, γ_{20}	1.357849	0.0000	-1.0000
Estimate		-0.4411	-1.3305
Standard error of estimate		0.0199	0.0202

χ^2 statistic = 7244.694443
Degrees of freedom = 2
p-value = <0.001

Final estimation of fixed effects
(Unit-specific model with robust standard errors)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	1.863650	0.025511	73.054	629	<0.001
MATHPROP, γ_{01}	0.004062	0.003484	1.166	629	0.244
ELAPROP, γ_{02}	0.027394	0.003616	7.577	629	<0.001
For MATHPROF slope, β_1					
INTRCPT2, γ_{10}	0.445129	0.023988	18.556	120255	<0.001
For ELAPROF slope, β_2					
INTRCPT2, γ_{20}	1.357849	0.027011	50.270	120255	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	1.863650	6.447229	(6.132,6.778)
MATHPROP, γ_{01}	0.004062	1.004070	(0.997,1.011)
ELAPROP, γ_{02}	0.027394	1.027772	(1.021,1.035)
For MATHPROF slope, β_1			
INTRCPT2, γ_{10}	0.445129	1.560691	(1.489,1.636)
For ELAPROF slope, β_2			
INTRCPT2, γ_{20}	1.357849	3.887822	(3.687,4.099)

Results of General Linear Hypothesis Testing - Test 1

	Coefficients	Contrast	
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	1.863650	0.0000	0.0000
MATHPROP, γ_{01}	0.004062	1.0000	0.0000
ELAPROP, γ_{02}	0.027394	0.0000	1.0000
For MATHPROF slope, β_1			
INTRCPT2, γ_{10}	0.445129	-1.0000	0.0000
For ELAPROF slope, β_2			
INTRCPT2, γ_{20}	1.357849	0.0000	-1.0000
Estimate		-0.4411	-1.3305
Standard error of estimate		0.0243	0.0274

χ^2 statistic = 4378.290985
Degrees of freedom = 2
p-value = <0.001

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	0.54599	0.29810	629	4257.33960	<0.001
level-1, r	0.98311	0.96650			

A residual file, called Grad Proficiency resfil2.sas, has been created. Note, some statistics could not be computed and a value of -99 has been entered. These should be recoded to 'missing values' before any analyses are performed.

Results for Population-Average Model

The value of the log-likelihood function at iteration 3 = -1.676875E+05

Final estimation of fixed effects: (Population-average model)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	1.804085	0.025488	70.781	629	<0.001
MATHPROP, γ_{01}	0.004144	0.002576	1.609	629	0.108
ELAPROP, γ_{02}	0.027071	0.002524	10.727	629	<0.001
For MATHPROF slope, β_1					
INTRCPT2, γ_{10}	0.432746	0.019122	22.631	120255	<0.001
For ELAPROF slope, β_2					
INTRCPT2, γ_{20}	1.318584	0.019317	68.261	120255	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	1.804085	6.074408	(5.778,6.386)
MATHPROP, γ_{01}	0.004144	1.004153	(0.999,1.009)
ELAPROP, γ_{02}	0.027071	1.027441	(1.022,1.033)
For MATHPROF slope, β_1			
INTRCPT2, γ_{10}	0.432746	1.541484	(1.485,1.600)
For ELAPROF slope, β_2			
INTRCPT2, γ_{20}	1.318584	3.738126	(3.599,3.882)

Results of General Linear Hypothesis Testing - Test 1

	Coefficients	Contrast	
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	1.804085	0.0000	0.0000
MATHPROP, γ_{01}	0.004144	1.0000	0.0000
ELAPROP, γ_{02}	0.027071	0.0000	1.0000
For MATHPROF slope, β_1			
INTRCPT2, γ_{10}	0.432746	-1.0000	0.0000
For ELAPROF slope, β_2			
INTRCPT2, γ_{20}	1.318584	0.0000	-1.0000
Estimate		-0.4286	-1.2915
Standard error of estimate		0.0193	0.0195

χ^2 statistic = 7428.998937
Degrees of freedom = 2
p-value = <0.001

Final estimation of fixed effects
(Population-average model with robust standard errors)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	1.804085	0.024860	72.570	629	<0.001
MATHPROP, γ_{01}	0.004144	0.003425	1.210	629	0.227
ELAPROP, γ_{02}	0.027071	0.003553	7.619	629	<0.001
For MATHPROF slope, β_1					
INTRCPT2, γ_{10}	0.432746	0.022607	19.142	120255	<0.001
For ELAPROF slope, β_2					
INTRCPT2, γ_{20}	1.318584	0.026274	50.186	120255	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	1.804085	6.074408	(5.785,6.378)
MATHPROP, γ_{01}	0.004144	1.004153	(0.997,1.011)
ELAPROP, γ_{02}	0.027071	1.027441	(1.020,1.035)
For MATHPROF slope, β_1			
INTRCPT2, γ_{10}	0.432746	1.541484	(1.475,1.611)
For ELAPROF slope, β_2			
INTRCPT2, γ_{20}	1.318584	3.738126	(3.550,3.936)

Results of General Linear Hypothesis Testing - Test 1

	Coefficients	Contrast	
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	1.804085	0.0000	0.0000
MATHPROP, γ_{01}	0.004144	1.0000	0.0000
ELAPROP, γ_{02}	0.027071	0.0000	1.0000
For MATHPROF slope, β_1			
INTRCPT2, γ_{10}	0.432746	-1.0000	0.0000
For ELAPROF slope, β_2			
INTRCPT2, γ_{20}	1.318584	0.0000	-1.0000
Estimate		-0.4286	-1.2915
Standard error of estimate		0.0229	0.0266

χ^2 statistic = 4213.179721
Degrees of freedom = 2
p-value = <0.001

APPENDIX B: HLM RESULTS FOR DROPOUT

Program:	HLM 8 Hierarchical Linear and Nonlinear Modeling
Authors:	Stephen Raudenbush, Tony Bryk, & Richard Congdon
Publisher:	Scientific Software International, Inc. (c) 2019
	hlm@ssicentral.com
	www.ssicentral.com

Module:	HLM2.EXE (8.0.2010.18)
Date:	14 February 2021, Sunday
Time:	18:27:42

Specifications for this Overdispersed Bernoulli HLM2 run

The maximum number of level-1 units = 120889
The maximum number of level-2 units = 632
The maximum number of micro iterations = 14
Method of estimation: full PQL

Maximum number of macro iterations = 100

Distribution at Level-1: Bernoulli

The outcome variable is DROPOUT

Summary of the model specified

Step 2 model

Level-1 Model

$\text{Prob}(DROPOUT_{ij}=1|\beta_j) = \phi_{ij}$
$\log[\phi_{ij}/(1 - \phi_{ij})] = \eta_{ij}$
$\eta_{ij} = \beta_{0j} + \beta_{1j}*(MATHPROF_{ij}) + \beta_{2j}*(ELAPROF_{ij})$

Level-2 Model

$\beta_{0j} = \gamma_{00} + \gamma_{01}*(MATHPROP_j) + \gamma_{02}*(ELAPROP_j) + u_{0j}$
$\beta_{1j} = \gamma_{10}$
$\beta_{2j} = \gamma_{20}$

MATHPROF ELAPROF have been centered around the group mean.

MATHPROP ELAPROP have been centered around the grand mean.

Level-1 variance = $\sigma^2/[\phi_{ij}(1-\phi_{ij})]$

Mixed Model

$\eta_{ij} = \gamma_{00} + \gamma_{01}*MATHPROP_j + \gamma_{02}*ELAPROP_j$
$+ \gamma_{10}*MATHPROF_{ij}$
$+ \gamma_{20}*ELAPROF_{ij}$
$+ u_{0j}$

The value of the log-likelihood function at iteration 7 = 3.460688E+04

Results for Non-linear Model with the Logit Link Function

Unit-Specific Model, PQL Estimation - (macro iteration 11)

$\sigma^2 = 0.61531$

Standard error of $\sigma^2 = 0.00251$

τ
INTRCPT1,β_0 2.98191

Standard error of τ
INTRCPT1,β_0 0.21504

Approximate confidence intervals of tau variances
INTRCPT1 : (2.588,3.436)

Random level-1 coefficient	Reliability estimate
INTRCPT1,β_0	0.732

The value of the log-likelihood function at iteration 2 = -1.428919E+05

Final estimation of fixed effects: (Unit-specific model)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	-4.616509	0.082024	-56.283	629	<0.001
MATHPROP, γ_{01}	-0.005647	0.007732	-0.730	629	0.465
ELAPROP, γ_{02}	-0.040971	0.007351	-5.573	629	<0.001
For MATHPROF slope, β_1					
INTRCPT2, γ_{10}	-0.956832	0.035539	-26.923	120255	<0.001
For ELAPROF slope, β_2					
INTRCPT2, γ_{20}	-1.311500	0.035278	-37.176	120255	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-4.616509	0.009887	(0.008,0.012)
MATHPROP, γ_{01}	-0.005647	0.994369	(0.979,1.010)
ELAPROP, γ_{02}	-0.040971	0.959857	(0.946,0.974)
For MATHPROF slope, β_1			
INTRCPT2, γ_{10}	-0.956832	0.384108	(0.358,0.412)
For ELAPROF slope, β_2			
INTRCPT2, γ_{20}	-1.311500	0.269416	(0.251,0.289)

Results of General Linear Hypothesis Testing - Test 1

	Coefficients	Contrast	
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-4.616509	0.0000	0.0000
MATHPROP, γ_{01}	-0.005647	1.0000	0.0000
ELAPROP, γ_{02}	-0.040971	0.0000	1.0000
For MATHPROF slope, β_1			
INTRCPT2, γ_{10}	-0.956832	-1.0000	0.0000
For ELAPROF slope, β_2			
INTRCPT2, γ_{20}	-1.311500	0.0000	-1.0000
Estimate		0.9512	1.2705
Standard error of estimate		0.0363	0.0360

χ^2 statistic = 3147.887971
Degrees of freedom = 2
p-value = <0.001

Final estimation of fixed effects
(Unit-specific model with robust standard errors)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	-4.616509	0.082344	-56.063	629	<0.001
MATHPROP, γ_{01}	-0.005647	0.006576	-0.859	629	0.391
ELAPROP, γ_{02}	-0.040971	0.006343	-6.459	629	<0.001
For MATHPROF slope, β_1					
INTRCPT2, γ_{10}	-0.956832	0.046770	-20.458	120255	<0.001
For ELAPROF slope, β_2					
INTRCPT2, γ_{20}	-1.311500	0.053318	-24.597	120255	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-4.616509	0.009887	(0.008,0.012)
MATHPROP, γ_{01}	-0.005647	0.994369	(0.982,1.007)
ELAPROP, γ_{02}	-0.040971	0.959857	(0.948,0.972)
For MATHPROF slope, β_1			
INTRCPT2, γ_{10}	-0.956832	0.384108	(0.350,0.421)
For ELAPROF slope, β_2			
INTRCPT2, γ_{20}	-1.311500	0.269416	(0.243,0.299)

Results of General Linear Hypothesis Testing - Test 1

	Coefficients	Contrast	
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-4.616509	0.0000	0.0000
MATHPROP, γ_{01}	-0.005647	1.0000	0.0000
ELAPROP, γ_{02}	-0.040971	0.0000	1.0000
For MATHPROF slope, β_1			
INTRCPT2, γ_{10}	-0.956832	-1.0000	0.0000
For ELAPROF slope, β_2			
INTRCPT2, γ_{20}	-1.311500	0.0000	-1.0000
Estimate		0.9512	1.2705

Standard error of estimate	0.0475	0.0539

χ^2 statistic = 1125.658806
Degrees of freedom = 2
p-value = <0.001

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	1.72682	2.98191	629	15093.15648	<0.001
level-1, *r*	0.78441	0.61531			

A residual file, called Dropout Proficiency resfil2.sas, has been created. Note, some statistics could not be computed and a value of -99 has been entered. These should be recoded to 'missing values' before any analyses are performed.

Results for Population-Average Model

The value of the log-likelihood function at iteration 3 = -1.393653E+05

Final estimation of fixed effects: (Population-average model)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	-3.636492	0.072497	-50.161	629	<0.001
MATHPROP, γ_{01}	-0.005627	0.006173	-0.912	629	0.362
ELAPROP, γ_{02}	-0.035481	0.005865	-6.050	629	<0.001
For MATHPROF slope, β_1					
INTRCPT2, γ_{10}	-0.833930	0.027975	-29.809	120255	<0.001
For ELAPROF slope, β_2					
INTRCPT2, γ_{20}	-1.144094	0.027381	-41.784	120255	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-3.636492	0.026345	(0.023,0.030)
MATHPROP, γ_{01}	-0.005627	0.994389	(0.982,1.007)
ELAPROP, γ_{02}	-0.035481	0.965141	(0.954,0.976)
For MATHPROF slope, β_1			
INTRCPT2, γ_{10}	-0.833930	0.434339	(0.411,0.459)
For ELAPROF slope, β_2			
INTRCPT2, γ_{20}	-1.144094	0.318513	(0.302,0.336)

Results of General Linear Hypothesis Testing - Test 1

	Coefficients	Contrast	
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-3.636492	0.0000	0.0000
MATHPROP, γ_{01}	-0.005627	1.0000	0.0000
ELAPROP, γ_{02}	-0.035481	0.0000	1.0000
For MATHPROF slope, β_1			
INTRCPT2, γ_{10}	-0.833930	-1.0000	0.0000
For ELAPROF slope, β_2			
INTRCPT2, γ_{20}	-1.144094	0.0000	-1.0000
Estimate		0.8283	1.1086
Standard error of estimate		0.0286	0.0280

χ^2 statistic = 4168.834021
Degrees of freedom = 2
p-value = <0.001

Final estimation of fixed effects
(Population-average model with robust standard errors)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	-3.636492	0.037086	-98.056	629	<0.001
MATHPROP, γ_{01}	-0.005627	0.002974	-1.892	629	0.059
ELAPROP, γ_{02}	-0.035481	0.003283	-10.806	629	<0.001
For MATHPROF slope, β_1					
INTRCPT2, γ_{10}	-0.833930	0.031421	-26.541	120255	<0.001
For ELAPROF slope, β_2					
INTRCPT2, γ_{20}	-1.144094	0.034535	-33.129	120255	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-3.636492	0.026345	(0.024,0.028)
MATHPROP, γ_{01}	-0.005627	0.994389	(0.989,1.000)
ELAPROP, γ_{02}	-0.035481	0.965141	(0.959,0.971)
For MATHPROF slope, β_1			
INTRCPT2, γ_{10}	-0.833930	0.434339	(0.408,0.462)
For ELAPROF slope, β_2			
INTRCPT2, γ_{20}	-1.144094	0.318513	(0.298,0.341)

Results of General Linear Hypothesis Testing - Test 1

	Coefficients	Contrast	
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-3.636492	0.0000	0.0000
MATHPROP, γ_{01}	-0.005627	1.0000	0.0000
ELAPROP, γ_{02}	-0.035481	0.0000	1.0000
For MATHPROF slope, β_1			
INTRCPT2, γ_{10}	-0.833930	-1.0000	0.0000
For ELAPROF slope, β_2			
INTRCPT2, γ_{20}	-1.144094	0.0000	-1.0000
Estimate		0.8283	1.1086
Standard error of estimate		0.0317	0.0349

χ^2 statistic = 1792.643623
Degrees of freedom = 2
p-value = <0.001

Program: HLM 8 Hierarchical Linear and Nonlinear Modeling
Authors: Stephen Raudenbush, Tony Bryk, & Richard Congdon
Publisher: Scientific Software International, Inc. (c) 2019
hlm@ssicentral.com
www.ssicentral.com

Module: HLM2.EXE (8.0.2010.18)
Date: 14 February 2021, Sunday
Time: 19:26: 0
License: HLM Standard

Specifications for this Multinomial HLM2 run

Problem Title: Intent Proficiency

The maximum number of level-1 units = 120889
The maximum number of level-2 units = 632
The maximum number of micro iterations = 14
Number of categories = 3
Method of estimation: full PQL

Maximum number of macro iterations = 100

Distribution at Level-1: Multinomial

The outcome variable is INTENT

Summary of the model specified

Step 2 model

Level-1 Model

$\text{Prob}[INTENT(1) = 1|\beta_j] = \phi_{1ij}$
$\text{Prob}[INTENT(2) = 1|\beta_j] = \phi_{2ij}$
$\text{Prob}[INTENT(3) = 1|\beta_j] = \phi_{3ij} = 1 - \phi_{1ij} - \phi_{2ij}$
$\log[\phi_{1ij}/\phi_{3ij}] = \beta_{0j(1)} + \beta_{1j(1)}*(MATHPROF_{ij}) + \beta_{2j(1)}*(ELAPROF_{ij})$
$\log[\phi_{2ij}/\phi_{3ij}] = \beta_{0j(2)} + \beta_{1j(2)}*(MATHPROF_{ij}) + \beta_{2j(2)}*(ELAPROF_{ij})$

Level-2 Model

$\beta_{0(1)} = \gamma_{00(1)} + \gamma_{01(1)}*(MATHPROP_j) + \gamma_{02(1)}*(ELAPROP_j) + u_{0j(1)}$
$\beta_{1(1)} = \gamma_{10(1)}$
$\beta_{2(1)} = \gamma_{20(1)}$

$\beta_{0(2)} = \gamma_{00(2)} + \gamma_{01(2)}*(MATHPROP_j) + \gamma_{02(2)}*(ELAPROP_j) + u_{0j(2)}$
$\beta_{1(2)} = \gamma_{10(2)}$
$\beta_{2(2)} = \gamma_{20(2)}$

MATHPROF ELAPROF have been centered around the group mean.

MATHPROP ELAPROP have been centered around the grand mean.

Final Results for Multinomial Iteration 11

$\sigma^2 = 1.00000$

τ

INTRCPT1(1)	1.17165	0.36938
INTRCPT1(2)	0.36938	0.48550

Standard errors of τ

INTRCPT1(1)	0.07589	0.04281
INTRCPT1(2)	0.04281	0.04026

Approximate confidence intervals of tau variances
INTRCPT1 : (1.032,1.331)
INTRCPT1 : (0.413,0.571)

τ (as correlations)

INTRCPT1(1),β_0	1.000	0.490
INTRCPT1(2),β_0	0.490	1.000

Confidence intervals of τ correlations

INTRCPT1(1),β_0	(1.000, 1.000)	(0.307, 0.638)
INTRCPT1(2),β_0	(0.307, 0.638)	(1.000, 1.000)

Random level-1 coefficient	Reliability estimate
INTRCPT1(1), $\beta_{0(1)}$	0.845
INTRCPT1(2), $\beta_{0(2)}$	0.616

The value of the log-likelihood function at iteration 2 = -2.240060E+05

Final estimation of fixed effects:

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For Category 1					
For INTRCPT1, $\beta_{0(1)}$					
INTRCPT2, $\gamma_{00(1)}$	-2.013169	0.046924	-42.903	629	<0.001
MATHPROP, $\gamma_{01(1)}$	0.006661	0.004395	1.516	629	0.130
ELAPROP, $\gamma_{02(1)}$	-0.035198	0.004247	-8.288	629	<0.001
For MATHPROF slope, $\beta_{1(1)}$					
INTRCPT2, $\gamma_{10(1)}$	-0.663314	0.021567	-30.756	119621	<0.001
For ELAPROF slope, $\beta_{2(1)}$					
INTRCPT2, $\gamma_{20(1)}$	-0.845618	0.021061	-40.150	119621	<0.001
For Category 2					
For INTRCPT1, $\beta_{0(2)}$					
INTRCPT2, $\gamma_{00(2)}$	-3.056187	0.035086	-87.106	629	<0.001
MATHPROP, $\gamma_{01(2)}$	-0.003294	0.003856	-0.854	629	0.393
ELAPROP, $\gamma_{02(2)}$	-0.013223	0.003847	-3.438	629	<0.001
For MATHPROF slope, $\beta_{1(2)}$					
INTRCPT2, $\gamma_{10(2)}$	-0.269606	0.033487	-8.051	119621	<0.001
For ELAPROF slope, $\beta_{2(2)}$					
INTRCPT2, $\gamma_{20(2)}$	-0.406966	0.033156	-12.274	119621	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For Category 1			
For INTRCPT1, $\beta_{0(1)}$			
INTRCPT2, $\gamma_{00(1)}$	-2.013169	0.133565	(0.122,0.146)
MATHPROP, $\gamma_{01(1)}$	0.006661	1.006683	(0.998,1.015)
ELAPROP, $\gamma_{02(1)}$	-0.035198	0.965414	(0.957,0.974)
For MATHPROF slope, $\beta_{1(1)}$			

INTRCPT2, $\gamma_{10(1)}$	-0.663314	0.515141	(0.494,0.537)
For ELAPROF slope, $\beta_{2(1)}$			
INTRCPT2, $\gamma_{20(1)}$	-0.845618	0.429292	(0.412,0.447)
For Category 2			
For INTRCPT1, $\beta_{0(2)}$			
INTRCPT2, $\gamma_{00(2)}$	-3.056187	0.047067	(0.044,0.050)
MATHPROP, $\gamma_{01(2)}$	-0.003294	0.996711	(0.989,1.004)
ELAPROP, $\gamma_{02(2)}$	-0.013223	0.986864	(0.979,0.994)
For MATHPROF slope, $\beta_{1(2)}$			
INTRCPT2, $\gamma_{10(2)}$	-0.269606	0.763680	(0.715,0.815)
For ELAPROF slope, $\beta_{2(2)}$			
INTRCPT2, $\gamma_{20(2)}$	-0.406966	0.665667	(0.624,0.710)

Results of General Linear Hypothesis Testing - Test 1

	Coefficients	Contrast			
For INTRCPT1, $\beta_{0(1)}$					
INTRCPT2, $\gamma_{00(1)}$	-2.013169	0.0000	0.0000	0.0000	0.0000
MATHPROP, $\gamma_{01(1)}$	0.006661	1.0000	0.0000	0.0000	0.0000
ELAPROP, $\gamma_{02(1)}$	-0.035198	0.0000	1.0000	0.0000	0.0000
For MATHPROF slope, $\beta_{1(1)}$					
INTRCPT2, $\gamma_{10(1)}$	-0.663314	-1.0000	0.0000	0.0000	0.0000
For ELAPROF slope, $\beta_{2(1)}$					
INTRCPT2, $\gamma_{20(1)}$	-0.845618	0.0000	-1.0000	0.0000	0.0000
For INTRCPT1, $\beta_{0(2)}$					
INTRCPT2, $\gamma_{00(2)}$	-3.056187	0.0000	0.0000	0.0000	0.0000
MATHPROP, $\gamma_{01(2)}$	-0.003294	0.0000	0.0000	1.0000	0.0000
ELAPROP, $\gamma_{02(2)}$	-0.013223	0.0000	0.0000	0.0000	1.0000
For MATHPROF slope, $\beta_{1(2)}$					
INTRCPT2, $\gamma_{10(2)}$	-0.269606	0.0000	0.0000	-1.0000	0.0000
For ELAPROF slope, $\beta_{2(2)}$					
INTRCPT2, $\gamma_{20(2)}$	-0.406966	0.0000	0.0000	0.0000	-1.0000
Estimate		0.6700	0.8104	0.2663	0.3937
Standard error of estimate		0.0220	0.0215	0.0337	0.0334

χ^2 statistic = 4281.667634
Degrees of freedom = 4
p-value = <0.001

Final estimation of fixed effects
(with robust standard errors)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For Category 1					
For INTRCPT1, $\beta_{0(1)}$					
INTRCPT2, $\gamma_{00(1)}$	-2.013169	0.047414	-42.460	629	<0.001
MATHPROP, $\gamma_{01(1)}$	0.006661	0.005159	1.291	629	0.197
ELAPROP, $\gamma_{02(1)}$	-0.035198	0.005084	-6.923	629	<0.001
For MATHPROF slope, $\beta_{1(1)}$					
INTRCPT2, $\gamma_{10(1)}$	-0.663314	0.022749	-29.159	119621	<0.001
For ELAPROF slope, $\beta_{2(1)}$					
INTRCPT2, $\gamma_{20(1)}$	-0.845618	0.024666	-34.282	119621	<0.001
For Category 2					
For INTRCPT1, $\beta_{0(2)}$					
INTRCPT2, $\gamma_{00(2)}$	-3.056187	0.034985	-87.357	629	<0.001
MATHPROP, $\gamma_{01(2)}$	-0.003294	0.004162	-0.792	629	0.429
ELAPROP, $\gamma_{02(2)}$	-0.013223	0.004410	-2.998	629	0.003
For MATHPROF slope, $\beta_{1(2)}$					
INTRCPT2, $\gamma_{10(2)}$	-0.269606	0.036576	-7.371	119621	<0.001
For ELAPROF slope, $\beta_{2(2)}$					
INTRCPT2, $\gamma_{20(2)}$	-0.406966	0.036028	-11.296	119621	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For Category 1			
For INTRCPT1, $\beta_{0(1)}$			
INTRCPT2, $\gamma_{00(1)}$	-2.013169	0.133565	(0.122,0.147)
MATHPROP, $\gamma_{01(1)}$	0.006661	1.006683	(0.997,1.017)
ELAPROP, $\gamma_{02(1)}$	-0.035198	0.965414	(0.956,0.975)
For MATHPROF slope, $\beta_{1(1)}$			
INTRCPT2, $\gamma_{10(1)}$	-0.663314	0.515141	(0.493,0.539)
For ELAPROF slope, $\beta_{2(1)}$			
INTRCPT2, $\gamma_{20(1)}$	-0.845618	0.429292	(0.409,0.451)
For Category 2			
For INTRCPT1, $\beta_{0(2)}$			
INTRCPT2, $\gamma_{00(2)}$	-3.056187	0.047067	(0.044,0.050)
MATHPROP, $\gamma_{01(2)}$	-0.003294	0.996711	(0.989,1.005)
ELAPROP, $\gamma_{02(2)}$	-0.013223	0.986864	(0.978,0.995)
For MATHPROF slope, $\beta_{1(2)}$			

INTRCPT2, $\gamma_{10(2)}$	-0.269606	0.763680	(0.711,0.820)
For ELAPROF slope, $\beta_{2(2)}$			
INTRCPT2, $\gamma_{20(2)}$	-0.406966	0.665667	(0.620,0.714)

Results of General Linear Hypothesis Testing - Test 1

	Coefficients	Contrast			
For INTRCPT1, $\beta_{0(1)}$					
INTRCPT2, $\gamma_{00(1)}$	-2.013169	0.0000	0.0000	0.0000	0.0000
MATHPROP, $\gamma_{01(1)}$	0.006661	1.0000	0.0000	0.0000	0.0000
ELAPROP, $\gamma_{02(1)}$	-0.035198	0.0000	1.0000	0.0000	0.0000
For MATHPROF slope, $\beta_{1(1)}$					
INTRCPT2, $\gamma_{10(1)}$	-0.663314	-1.0000	0.0000	0.0000	0.0000
For ELAPROF slope, $\beta_{2(1)}$					
INTRCPT2, $\gamma_{20(1)}$	-0.845618	0.0000	-1.0000	0.0000	0.0000
For INTRCPT1, $\beta_{0(2)}$					
INTRCPT2, $\gamma_{00(2)}$	-3.056187	0.0000	0.0000	0.0000	0.0000
MATHPROP, $\gamma_{01(2)}$	-0.003294	0.0000	0.0000	1.0000	0.0000
ELAPROP, $\gamma_{02(2)}$	-0.013223	0.0000	0.0000	0.0000	1.0000
For MATHPROF slope, $\beta_{1(2)}$					
INTRCPT2, $\gamma_{10(2)}$	-0.269606	0.0000	0.0000	-1.0000	0.0000
For ELAPROF slope, $\beta_{2(2)}$					
INTRCPT2, $\gamma_{20(2)}$	-0.406966	0.0000	0.0000	0.0000	-1.0000
Estimate		0.6700	0.8104	0.2663	0.3937
Standard error of estimate		0.0234	0.0251	0.0364	0.0360

χ^2 statistic = 2208.199909
Degrees of freedom = 4
p-value = <0.001

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1(1), $u_{0(1)}$	1.08243	1.17165	629	8168.25373	<0.001
INTRCPT1(2), $u_{0(2)}$	0.69678	0.48550	629	3066.36724	<0.001

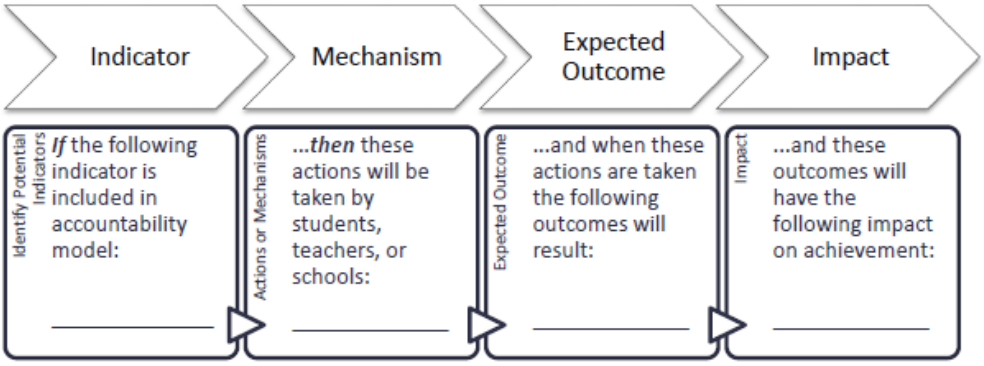
A residual file, called Intent Proficiency resfil2.sas, has been created. Note, some statistics could not be computed and a value of -99 has been entered. These should be recoded to 'missing values' before any analyses are performed.

DESIGNING AN ACCOUNTABILITY SYSTEM

The goal of accountability is to increase student achievement so students will be college and career ready after graduation.

THEORY OF ACTION

Theory of Action (TOA) is an if then statement that explicitly states the intended outcomes as well as the mechanisms through which they will occur conceptually and operationally. Drafting a TOA focuses on achieving desired goals and outcomes and avoiding unintended consequences.



Once the Theory of Action is affirmed, the following steps are recommended for identifying valid indicators for the system.

Identify Potential Indicators of Student, Teacher, and School Success

Accountability models include a variety of indicators to provide information on school performance. Accountability systems shift the unit of measurement from individual students to aggregate scores at the teacher or school level. Traditionally test developers have focused on the validity of the student level score. However, attention needs to be given to the validity of the teacher or school level indicators included in the accountability model. The validity of the model, and by association the indicators, is critical to equity and opportunity in targeting resources.

Seek Stakeholder Input

Stakeholders include school administrators, teachers, and the community at large such as businesses who may hire high school graduates, and higher educational institutions who may enroll them. Providing validity evidence for proposed accountability indicators centers the discussion, and thus any outcomes, on data rather than perceptions.

Analyze Indicators

When evaluating whether to include an indicator in an accountability model, validity evidence for the measure affirms the appropriateness and the contribution of the indicator to the model. Hierarchical Linear Modeling (HLM) is particularly useful in providing empirical validity evidence for the use of aggregate scores that have already been validated at the individual student level. HLM is also useful in determining whether scores or derivatives of scores, such as proficiency or percent proficient, lead to more reliability and validity in the outcomes.

Evaluate Outcomes

Accountability models' outcomes may be used to compare schools, identify schools for technical assistance that may be accompanied by funding, and in cases, award bonuses for teachers and principals. The positive consequences of an accountability system should outweigh the negative.