UNIVERISTY OF CALIFORNIA

Los Angeles

Examining the Validity of Classifications from an English Language Proficiency Assessment for English Language Learners and Native English Speakers in Fifth Grade

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Arts

in Education

by

Patricia Elaine Carroll

2012

ABSTRACT OF THE THESIS

Examining the Validity of Classifications from an English Language Proficiency Assessment for English Language Learners and Native English Speakers in Fifth Grade

by

Patricia Elaine Carroll

Master of Arts in Education

University of California, Los Angeles, 2012

Professor Alison Bailey, Chair

English language proficiency assessments are used to report English Language Learner (ELL) proficiency and progress, and classifications from these assessments help determine what educational services students receive. While classification validity is routinely reported by test vendors, empirical evidence is rarely used to justify the use of a certain classification model or scheme. In this study, fifth grade ELL (*n*=875) and native English speaker (non-ELL, *n*=92) performance data on a State's Standards-Based Achievement Assessment (SBAA) and the State's English Language Proficiency Assessment (ELPA) are used to evaluate "proficient" and "non-proficient" classifications on the ELPA. Findings indicate that the currently-used conjunctive model, compared to compensatory and mixed models, creates classifications that are the least congruent with other sources of proficiency evidence for ELL and non-ELL students. Findings from this study can directly inform how policy, practice and research communities define and verify ELPA classifications for use in high-stakes decision making.

The thesis of Patricia Elaine Carroll is approved.

José-Felipe Martínez-Fernandez

Jeffrey J. Wood

Alison Bailey, Committee Chair

University of California, Los Angeles

2012

To my parents, with gratitude.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I thank my graduate advisor and thesis chair, Alison Bailey, for her tremendous support, encouragement and guidance throughout this project. I also thank my graduate advisor Diane Haager and my committee members, Felipe Martínez and Jeffrey Wood, for their keen feedback and support. I also thank Robert Linquanti for his valuable insights on an earlier version of this work and Noreen Webb for her encouragement during the genesis of this project. Additionally, I thank the numerous graduate student colleagues whose enthusiasm for each evolving version kept me energized. In particular, I thank Anna Osipova, Cathy Coddington, Rachel Zwass, Sandy Chang, Yiching Huang, Stacey Beauregard, Sihan Xiao, Alejandra Priede, Cristal Byrne, Monica Esqueda, and Molly Faulkner-Bond for their support. Last, I wholeheartedly thank the Department of Education of "State A" for granting me the opportunity to analyze these unique data and for their constructive comments. All errors that remain in the paper are my own.

Introduction

Decisions about students are important, especially high-stakes decisions which affect access to a free, appropriate public education. Much of the research related to high-stakes decision making for students has focused on the validity of inferences based on assessments and, in particular, individual test scores (Brennan, 2006; see also Abedi, 2008a; Llosa, 2007, 2008; Sireci, Han & Wells, 2008). Decisions related to our educationally most vulnerable students, such as English Language Learner (ELL) students[1], are best made using multiple sources of data (Kim & Herman, 2010; Wolf, Herman, Bachman, Bailey & Griffin, 2008). English-language proficiency, for students whose home language influences are not "English Only", has been defined as a construct which can predict success in an English-only instructional setting (Solórzano, 2008). Yet, measurements of language proficiency often fall short in their ability to correctly classify as proficient the ELL students who are ready for such settings (Xiong & Zhou, 2006). How data from these measures are transformed and aggregated into Fluent English Proficient (FEP) and Limited English Proficient (LEP) classifications has a tremendous impact on access to learning opportunities for ELL students (Kim, 2011; MacSwan & Rolstad, 2006; Mahoney & MacSwan, 2005; Wang, Niemi & Wang, 2007; Xiong & Zhou, 2006). Especially for a proficient student preparing to enter secondary schooling, a misclassification of LEP could mean that instead of placement in general or gifted classes with pacing, rigor and content appropriate for his or her ability, that student would likely be assigned to classes designed for LEP students with curricula substantively different in the depth, breadth, pacing and complexity of content (Walqui et al., 2010).

In this way, classification errors can create a systemic barrier to learning, educational progress and opportunity for ELL students (Solórzano, 2008) and have been shown to have long-term effects on school persistence (Kim, 2011). Therefore, no study to validate inferences

[1] For this and all other acronyms, see the glossary provided in Appendix A.

about English Language Proficiency (ELP) would be complete without investigating the decision rules (henceforth, classification models and schemes) used to translate ELPA scores into "proficient" (FEP-eligible) and "non-proficient" classifications. Prior studies of reliability and validity of ELP assessments have examined the test development process, test formats, cut scores, and the performance of reclassified FEP students. However, no known studies to date have investigated the impact of classification models and schemes on the extent to which ELPA classifications predict FEP-eligible levels of English-language Proficiency.

Despite this gap in measurement accountability, numerous ELP assessments are already in use in all fifty States from kindergarten through twelfth grade to make decisions about the growing population of ELL students, predicted to be twenty-five percent of all students in the United States within the next decade (U.S. Department of Education, 2006). Classification consistency and accuracy are particularly high-stakes for fifth grade ELL students transitioning into secondary school settings. In sixth grade and beyond, classes designed for LEP students can supplant grade-level English Language Arts instructional minutes, leaving students with less opportunity to engage with at-grade-level content and vocabulary. Furthermore, the scheduling of specialized English Language Development instruction is often coordinated with remedial classes in other content areas in such a way that ELL students currently classified LEP are clustered into low-performing, within-subject tracks for the entire instructional day (Xiong & Zhou, 2006).

This study proposes an investigation of the validity of inferences made from the ELPA classification of "proficient" or "non-proficient" using data collected in 2010 by the State A Department of Education from native English speakers (henceforth "non-ELL")[2] and ELL students. The ELPA classification of "proficiency" (i.e., FEP-eligibility), which is created in State A by a conjunctive model, will be investigated in light of a second source of FEP-eligibility

[2] In study, non-ELL students are defined as those considered native English speakers with home language influences determined to be "English Only" and have never been considered to be or classified as ELL students. Students with designation of Special Education or testing accommodations have been excluded for this study.

provided by the SBAA scores. If classifications created from ELPA scores are an accurate determiner of FEP-eligible levels of English-language proficiency, it would be expected that non-ELL students would be FEP-eligible according to the ELPA, especially those scoring at FEP-eligible levels on the SBAA. It is also expected that FEP-eligibility on the SBAA will converge with FEP-eligibility on the ELPA, especially for those scoring at the highest levels on all four SBAA domains. Taken together, the convergence of SBAA and ELPA determinations of FEP-eligibility, especially for high-achieving students, would suggest that the conjunctive model is a good fit both logically and practically for the measurement of English-language proficiency. Lack of convergence would indicate a lack of model fit with high cost to States, districts and schools in terms of inflated measurement costs and poorly served long-term ELL students, with the highest cost to these students themselves and their families (Gándara, Rumberger, Maxwell-Jolly, & Callahan, 2003). For this reason, alternate classification models and schemes will also be explored and implications will be discussed.

Literature Review

Validity

Based on the work of Messick (1989), validity is viewed as a unitary concept which argues the suitability of assessment outcomes for interpretations and uses based on multiple sources of evidence and related analyses (Messick, 1989; see also Bachman & Palmer, 2010; Clark & Watson, 1995; Kane, 2006, 2010; Shepard, 1993, 1997). According to Forte, Perie & Paek (2012), the fundamental validity question regarding ELP assessments is “whether a student who is deemed proficient by an ELPA can successfully function without language supports in academic classes taught in English" (p. 8). To date, validity studies have typically investigated the test development process (Davidson, Kim, Lee, Li & Lopez, 2007; Garcia, Lawton & Diniz de Figueiredo, 2010; see also technical reports for individual ELP assessments, or Abedi, 2007 for a general review), the constructs tested (Bailey, 2007; Bailey

& Butler, 2003, 2004) the impact of test formats and measurement models (Zhang, 2010), the impact of cut scores (Florez, 2012; Wang et al., 2007), the performance of students who have been reclassified FEP (Abedi, 2008b; Kim & Herman, 2010; Ragan & Lesaux, 2006) and a comparison of standards-based achievement assessment mean scores of native English speaker to ELL students (Crane, Barrat & Huang, 2011). Guidance has also been provided for developing validity studies related to the assessment of ELL students (Sireci et al., 2008; Wolf, Farnsworth & Herman, 2008)[3] and overall ELL assessment systems (Wolf, Herman & Dietel, 2010). However, neither the guidelines nor the known studies to date have addressed validity in terms of the impact of a classification model or scheme on the inferences drawn regarding FEP-eligibility and thus the determination of readiness for English-only classroom placement.

Those responsible for ELPA systems are obligated by professional standards to investigate the validity of the inferences drawn from classifications. As stated in the *Standards for Educational and Psychological Testing*: "When raw score or derived score scales are designed for criterion-referenced interpretations, including the classification of examinees into separate categories…serious efforts should be made whenever possible to obtain independent evidence concerning the soundness of such score interpretations" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, p. 56-57). Furthermore, the goal of measurement professionals is to develop assessments, and classification systems, which minimize construct-irrelevant variance in order to approximate the inferences of the scores for the construct (AERA et al., 1999) so that "the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error" (Hambleton & Novick, 1973, in Zhang, 2010, p. 126). As the composite classification schemes applied to ELP assessments directly impact the inferences drawn from "proficient"

[3]See also the *Evaluating the Validity of English Language Proficiency Assessments (EVEA) Project,* a federally-funded initiative to develop an argument-based approach to validity evaluations (www.eveaproject.com).

(FEP-eligible) or "non-proficient" labels, the validity of those inferences depends on the validity of the schemes.

Attainment of the criteria for "proficiency" is measured on an ELPA first through the cut scores for each sub-domain which create several level determinations (e.g., ranging from "Beginner" to "Fluent"). The accuracy of those level classifications within each sub-domain is critically dependent on measurement accuracy at each cut score point (Ercikan & Julian, 2002; Mislevy, Wilson, Ercikan, & Chudowsky, 2002). Measurement error can lead to misclassifications, which exist in two forms: *false-positive* (Type I) resulting from being identified proficient when not; and *false-negative* (Type II) when identified as non-proficient when actually proficient. Measurement error, especially around the cut scores, is important to quantify or estimate in determining the accuracy and consistency of the classifications (Livingston & Lewis, 1995; see also Stone, Weissman, & Lane, 2005).

For this reason, it is crucial to also evaluate the validity of the classification scheme used to determine an overall cut-point as it is what ultimately quantifies the "master" or "non-master" level of a given attribute. The manner in which schemes are applied may be the most crucial factor in evaluating the reliability and validity of inferences from these classifications (Chester, 2003; Douglas, 2007). In the context of fourth grade school promotion decisions, Chester (2003) demonstrates an intricate logic used in the combination of multiple measures of academic achievement. In his framework, Chester (2003) gives a brief overview of three models (conjunctive, complementary, and compensatory) in terms of several types of indicators: measures of different constructs, different measures of the same constructs, multiple opportunities to take the same assessment, and the use of accommodations and alternate assessments (p. 34). Although in the context of a different academic decision-making process, Chester asserts that it is the logic by which measures are combined, not simply the

use of multiple measures, which determines the reliability and validity of inferences from a decision and directly impacts the efficacy of consequences.

To explore a model for determining the reliability of a composite created from sub-tests within one measure, Douglas (2007) applied the models laid out by Chester in terms of five schemes (conjunctive, complementary, compensatory, conjunctive-complementary, and conjunctive-compensatory) to simulated and authentic GED test data. The extent of misclassification for all schemes was determined by split-half reliability tests and was 12% or less for *false-negatives* and *false-positives* combined. Douglas called careful attention to the nuance of each finding, in one example noting that while a conjunctive scheme created more misclassification, primarily *false-negatives*, a compensatory scheme created less misclassification yet the percent of *false-positive* cases had increased (see Douglas, 2007; Douglas & Mislevy, 2010). What remains unexplored is criterion-related validity of the classification based on an external measure (specifically to determine the extent of *false-positives* in comparison or contrast to *false-negatives* under certain schemes), and the validity of the logic behind each scheme for the purposes of test score use (e.g., which scheme creates GED classifications which best predict the success of "masters" in their first job).

To begin that process in the present study, general discussions of schemes were found in literature related to education (e.g., "compensatory" vs. "non-compensatory" in Kane & Case, 2004; Chester, 2003 and Douglas, 2007 as described above) as well as licensure and certification in the medical field (e.g., "fully compensatory" vs. "partially compensatory" in Clauser, Clyman, Margolis, & Ross, 1996). However, none were complete. Thus, to facilitate a common understanding for the discussion to follow, Table 1 has been created to summarize the four main classification models with a definition of "mastery", example schemes, underlying logic, and likely use for each.

Table 1: *Classification models for combining multiple indicators*

	Conjunctive	Compensatory	Complementary	Mixed
Defining Mastery	Performances on all tests, subtests or indicators must be at or above a predetermined cut score or mastery level	Performance(s) below "mastery" on one (or more) test, subtest or indicator can be compensated for by a performance at or above "mastery" on another	Performances at or above "mastery" on any measure can fulfill the requirement	Definition of "mastery" can be achieved with any predetermined mix of conjunctive, compensatory, and/or complementary schemes
Example Scheme(s)	Graduate admissions criteria requiring a minimum score of 160 on both the verbal and quantitative domains of the GRE; driver's licensure criteria requiring passing scores on both the written and road tests.	Graduate admissions criteria requiring a minimum GRE score of 320 (overall); a high school exit exam passing score set at 80% overall	Undergraduate admissions criteria requiring a minimum score on the ACT or the SAT.	Graduate admissions criteria requiring a minimum GRE score of 320 (overall) and verbal and quantitative scores at or above 155
Logic	All indicators are equally important for predicting a latent variable, e.g., "readiness for college", "driver competency".	All indicators make necessary contributions, but not all indicators are expected to be at equally high levels in order to predict the latent variable. Rather, a predictive composite could include indicators to coexist in an uneven profile, e.g., for some considered gifted in math, very high GRE quantitative scores could co-occur with moderately low verbal scores in students possessing "readiness for college"	Both indicators provide inferences that are interchangeable for the purpose of the decision, therefore only one is needed.	All indicators are equally important for determining and/or predicting the presence of the latent variable by being at least at a set minimum level, with mastery levels expected for most indicators.
Likely Use	When uneven profiles among indicators (e.g., high scores on a written driver's licensure test, but low scores on the road test) are unacceptable based on known risk factors correlated with these failures (e.g., likelihood of accident proneness)	When known misclassification risk from measurement error (e.g., augmented from combining multiple indicators of differing reliabilities) is greater than any known risk from uneven profiles of indicators	When multiple indicators are considered equivalent and yet not equally accessible for all examinees	When the risk of misclassification due to measurement error is matched or exceeded by a known risk from uneven profiles of indicators (e.g., very low verbal scores are unlikely to correlate with "readiness for college", even if quantitative scores are very high)

Classification Models and Schemes

The choice of a classification model, defined here as a decision rule or an approach for combining multiple indicators, is based on a theoretical stance related to the nature of the scores or indicators being combined and the purposes for which the classifications will be used. Within a chosen model, a composite classification scheme (algorithm, formula or combination rule) is chosen to convert multiple scores into a dichotomous "master" or "non-master" determination. In ELP assessments, classification models and schemes are used to combine scores from sub-domains (e.g., *Listening, Speaking, Reading* and *Writing*) and/or levels (e.g., Beginner, Advanced Beginner, Intermediate, Early Fluent, Fluent) into "proficient" and "non-proficient" classifications.

To determine the reliability of composites, Mosier (1943) provides a generalized formula which can be estimated from the variance, weights, reliabilities and intercorrelations of the components. Mosier emphasized that "for mutually uncorrelated components, the reliability of the weighted composite is a weighted mean of the reliabilities… (and) for equal weights, this reduces to the mean reliability of the components, a value generally less than that of the most reliable of the sub-tests" (p. 165; see also Wang & Stanley, 1970). Ideally, when devising a reliable composite for an ELPA the intercorrelation of sub-domains (Abedi, 2007), dependence of items within sub-domains (e.g., testlet format), as well as the differing reliabilities of sub-domains (due to different item types, modes of construct measurement, modes of delivery, length, rater reliability, etc.) should be considered during the test development process (see also Abedi, 2002, 2004; Bachman & Palmer, 2010). As individual item data is not available for the study reported here, such investigations will need to be explored in future research. Alternatively, a substantive review of models and schemes was conducted with available data.

The 2011 National Research Council panel report "*Allocating Federal Funds for State Programs for English Language Learners"* provides a broad overview of the ELP assessments

in use in the 2009-2010 school year. While all the tests reviewed reported scores with an overall composite classification summarizing performance in the four sub-domains (plus a comprehension score composite of listening and reading tests), it is noted that "these composites are not consistently based on either equally or unequally weighted subscale scores" (p. 69). Some States use equally weighted[4] sub-domains but employ a composite classification scheme that is either conjunctive (e.g., Colorado, with an overall score of "5 - Advanced" and a mandatory pass of all four sub-domains at "4.5 or higher" on the CELA), compensatory (e.g., Washington, with an overall score composite of "4"), or mixed (e.g., California, with a mandatory pass at least three of four sub-domains at or above "Level 4 - Early Fluent", with the fourth sub-domain at or above "Level 3 - Intermediate" on the CELDT). In addition to these models, many States also implement weighted formulas. For the approximately 23 States using ACCESS for ELLs® (WIDA, 2009) in the 2009-10 school year, each State partnered with WIDA to customize of their own mixed, weighted classification schemes. For example, in North Carolina a weighted formula is employed (e.g., Listening= 0.15, Speaking=0.15, Reading=0.35, Writing=0.35) and applied to a classification scheme proprietary to ACCESS for ELLs® including what can be interpreted as a compensatory-conjunctive rule requiring an overall level of 4.8 with a level of 4.0 or higher on both *Reading* and *Writing* (National Research Council, 2011, p. 89). Although an investigation of the WIDA model used in each participating State is beyond the scope of this paper, it is included here to demonstrate the complexity of models applied to ELP assessments nationwide (see also Porter & Vega, 2007 for a general review).

In the case of the State A ELPA, a conjunctive model was chosen where "proficient" is defined as reaching or surpassing the cut score for Level 4 (Early Fluent) in all four ELPA sub-

[4] Although "equally weighted" can mean no additional weights are assigned, it is understood that only tests whose sub-domains have the exact same number of items and the exact same cut score in each are "equally weighted".

domains. To set cut scores for each sub-domain test, the State A ELP standard-setting procedure was conducted in 2006 and revisited in 2009 using the enrolled State population of ELL students as a standard-setting population, although the proportions of "not-yet-proficient" and "ready-to-be-classified-proficient" students contributing test score data was unknown. Levels were set using the Bookmark standard-setting procedure (Mitzel, Lewis, Patz & Green, 2001; for a review, see Karatonis & Sireci, 2006). For each grade-level within each cluster test, cut scores were nominated by a panel of experts for each sub-domain over three successive rounds, then smoothed to produce five levels (Beginner, Advanced Beginner, Intermediate, Early Fluent, and Fluent). As there was not enough data to support five levels within each of the sub-domain tests, those are described by three reportable levels (Beginner, Advanced Beginner/Intermediate, and Early Fluent/Fluent). While cut scores have been adjusted regularly to maintain the expected proportion of students in these levels, no analysis of the percent of proficient and non-proficient students has been conducted using external sources of proficiency data. Since the first pilot exam in 2006, the scores and proficiency classifications from the annual ELP assessments have been used as a first step in determining FEP eligibility.

English Language Proficiency Assessments and Accountability

There are multiple uses for the ELPA levels and classification at the State and district level, some more high-stakes at the individual student level than others. For compliance related to Title I and Title III funding of The No Child Left Behind Act (NCLB) of 2001, States report ELPA and SBAA scores to the federal government for three Annual Measurable Achievement Objectives (AMAO) which are as follows: (1) Progress or Growth based on the proportion of ELL students receiving Title III services who gained an ELP level in a given year, (2) English language proficiency based on the proportion of ELL students receiving Title III services who attained an overall "proficient" classification, and (3) Core Content Proficiency based on the Title I Annual Yearly Progress (AYP) outcome for the LEP student group or the

subset that represents Title III-served ELL students (NCLB, 2001a, 2001b, 2001c). All States must assess ELL students annually on the ELPA and SBAA. Goals for AMAO 1 and 2 can be calculated with ELPA scores using one of these options: (a) the sub-domain scores for reading, writing, speaking and listening; (b) a composite score that reflects performance in all four of these sub-domains; or (c) a combination of one or more sub-domain scores and the composite score that reflects performance in all four of these sub-domains (NCLB, 2001c; see also Forte & Faulkner-Bond, 2010; Linquanti & George, 2007). States are allowed to set different progress targets for *Listening, Speaking, Reading* and *Writing* for AMAO 1, and are allowed to choose any type of ELPA composite score for AMAO 2 as long as it is justified with evidence. Thus, the classification scheme used to create that composite has been up to each State to decide.

In 2008, the United States Department of Education clarified that the scheme used to create the composite used to classify students does not have to abide by a conjunctive classification model, rather "a State can use a composite score so long as the State can demonstrate that the composite score meaningfully measures student progress and proficiency in each of the language domains and, overall, is a valid and reliable measure of student progress and proficiency in English, consistent with the purpose for which the assessment is used" (USDE, Federal Register, Nov 18, 2008 in Forte & Faulkner-Bond, 2010, Appendix D, p.6). Again, the demonstration of the "meaningful measurement" of proficiency through a chosen model or scheme is expected, but no known studies or guidelines to date have provided States with guidance on collecting and evaluating evidence towards this end.

Besides the use of the ELPA for federal accountability purposes, individual districts also use these classifications in the larger aggregate decision-making process of FEP reclassification in terms of how students are placed in academic programming. Even though States can stipulate the use of multiple criteria to determine how and when students can be

classified FEP (Mahoney & MacSwan, 2005), the way sources of evidence are prioritized, weighted and combined may be determined at the district or school level depending on the State (Abedi, 2008b; Ragan & Leseaux, 2006) but are typically conjunctive. Of all available evidence (such as ELPA scores and classifications, SBAA scores and levels, teacher evaluations, parental opinions, and a comparison of skills relative to native English speaker students) the ELPA classifications are considered prima facie evidence for redesignation. As such, "non-proficient" on the ELPA can be the sole determiner of ineligibility for reclassification when conjunctive decision rules are applied (Ragan & Leseaux, 2006; Robinson, 2011). This is especially undesirable in States where ELPA data is the first evidence collected in a given school year, yet not available for use until many months after the fact (e.g., California CELDT data collected in October, delivered back to districts in March-June; see Walqui et al., 2010). The order in which sources of FEP eligibility evidence are collected, and how these sources are ultimately combined for high-stakes decision-making for individual student FEP decisions at the school level is beyond the scope of this paper, but is in need of careful investigation.

External Criteria for Convergent Validity

The relationship between ELPA and SBAA data in this context, however, informs the design of this study. Specifically, the inference of "FEP-eligible" as determined from scores on the ELPA will be considered here in light of the inference of "FEP-eligible" as determined from scores on the SBAA as both assessments play a role in this context-specific inference of English language proficiency. In a perfect measurement world, an ELPA classification of "proficient" (i.e., FEP-eligible) implies the following: (1) the latent variable of "English-language Proficiency" is present at FEP-eligible levels; (2) the influence of "Home Language Other Than English" has been overcome to the extent that it is no longer a source of error on the ELPA; and (3) English-language proficiency is not likely to be a source of error on academic content tests administered in English. Furthermore, this FEP-eligible level of English-language

proficiency is the *sine qua non* of high academic achievement on content-based tests given in English, and thus is expected to co-occur with the highest performance levels of SBAA in all domains (e.g., "Proficient" or "Advanced" in *Reading, Language Usage, Math, and Science*). However, FEP-eligible levels of English-language proficiency can also co-occur with even the lowest levels of achievement (e.g., "Below Basic") insomuch as the SBAA domains assess content, not just communicative abilities in English. In point of fact, the lack of achievement in SBAA content can be observed in native English speakers.

By this same logic, an ELPA classification of "non-proficient" (i.e., not FEP-eligible) implies the following: (1) the latent variable of English-language proficiency is not present at FEP-eligible levels; (2) the influence of "Home Language Other Than English" has not been overcome and is likely a source of error on the ELPA (or, that another influence known to impact language acquisition – e.g., specific learning disability, lack of opportunity to learn, poor attendance, poverty – has not been overcome); and (3) lack of English-language proficiency is likely to be a source of error on tests administered in English, thus predicting inability to show content knowledge and acquisition of standards through SBAA given in English even in cases where testing accommodations have been provided.

For the purposes of determining the likely number of English-language proficient students in the sample used in this study, FEP-eligible determinations from SBAA scores will be used as an external criterion related to convergent validity. It is clear that high performance levels on the SBAA can likely confirm the presence of English-language proficiency at FEP-eligible levels, but no SBAA performance level(s) can confirm the lack of English-language proficiency. Therefore, the only prediction that can be made using the SBAA is of FEP-eligible levels of English-language proficiency for students with high achievement on the SBAA. The FEP-eligible levels of English-language proficiency will be estimated using two criteria: First, State A FEP-eligible criterion of "Basic", "Proficient" or "Advanced" in *Reading* (also commonly

used nationwide) will be used to determine the percentage of students likely to be FEP-eligible on the ELPA. Second, to identify a level of performance where FEP-eligible English-language proficiency is most defensibly present, "Proficient" or "Advanced" in all domains (*Reading, Language Usage, Math* and *Science)* will be used to identify a sub-set of "High-Performing SBAA" cases. The ELPA classification of these cases should be "proficient", thus can be used to test the classification sensitivity (ability to predict *true-positives*) of each classification scheme.

Using the substantive argument that English-language proficiency is defined as a construct which can predict success in an English-only instructional setting (Solórzano, 2008), it would reasonably follow that non-ELL students who are currently receiving instruction in English-only settings and performing at or above the SBAA FEP-eligibility level could provide a "known-to-be-proficient" comparison group from which to draw inferences related to the validity of assessment classifications. Thus, it is expected that all non-ELL cases, but most certainly those at SBAA FEP-eligible levels, will be classified ELPA "proficient". If any of these FEP-eligible non-ELL cases are considered by the ELPA to be "non-proficient" and thus "not likely to be successful in an English-only classroom", it seems logical to explore other classification schemes that would be more predictive of actual proficiency.

Studies with ELL and non-ELL Students

Verification of classification models is rarely attempted using external measures (see Kane & Case, 2004). For ELPA validation research using non-ELL students, the known studies to date have stopped short of providing SBAA FEP-eligibility levels, or additional sources of data, to support their assumptions about the expected proficiency levels of their non-ELL (or ELL) students.

An internal, unpublished study for Pearson Assessment conducted by Stephenson, Jiao, and Wall (2004) compared the performance of ELL and non-ELL students[5] on outcomes of the Stanford English Language Proficiency Assessment (Stanford ELP, 2003). The sample of students was drawn from 70 school districts in 20 States and students were administered the sub-domain tests of the SELP that were multiple-choice, untimed and group-administered only (*Listening, Writing Conventions,* and *Reading*). Sub-domain tests were administered in grade clusters: Primary (second grade only), Elementary (third, fourth and fifth grade), Middle Grades (sixth, seventh and eighth grade) and High School (ninth through twelfth grade). A sample of 400 students was randomly selected for each cluster (ELL, *n=200;* non-ELL, *n=200*) and means scores were used to identify group differences. Researchers conducted Analyses of Variance (ANOVA) and the results showed statistically significant differences between groups, with non-ELL students scoring higher than ELL students at each grade cluster and on every assessment. Secondly, they used a discriminant analysis to determine if group membership (ELL or non-ELL) could be reliably predicted from the total of the three sub-domain scores, presumably with the expectation that all non-ELL students would be classified "proficient". The percentage of non-ELL cases classified proficient was 64% in Primary, 78% in Elementary, 87% in Middle Grades, and 83% in High School. Researchers concluded that the Primary and Elementary classifications are not very reliable, but considered the Middle Grades and High School classifications to be sufficiently reliable. The percentage of ELL cases classified proficient were reported as 52% in Primary, 44% in Elementary, 27% in Middle Grades, and 23% in High School but these results were not discussed. What is left unexplored in this study is a discussion of whether the misclassification rates of non-ELL students can inform the reliability and validity of classifications for ELL students, what percentage of non-ELL

[5] Stephenson et al. use "Non-Native Speakers" and "Native Speakers" to refer to ELL and non-ELL students, respectively. We have continued to use our terminology of "ELL" and "non-ELL" for the ease of the reader as we compare studies in this section while also acknowledging these terms may not be fully comparable in all studies.

students are likely to be considered “non-proficient” on an ELP test, and whether this is a meaningful comparison without other sources of proficiency data.

In a related study, Stephenson, Johnson, Jorgensen, and Young (2003) administered the SAT9 Reading Comprehension subtest to a group of ELL students who had taken the Stanford ELP test (no sample numbers provided). The results for second, third and fourth grade students were divided into “proficient” and “non-proficient” groups according to ELP results, and mean scores on the SAT9 were compared. The authors provide mean raw score, mean scale score and mean percentile rank of the SAT9 scores to demonstrate that students in the “non-proficient” group scored lower on average than students in the “proficient” group. Although the authors report that “the data show how language proficiency as measured by one instrument is confirmed by results obtained from a separate, equally valid, instrument” (p. 16), it is not clear that the validity of inferences of the measures were considered. What is left unexplored is the extent to which ELL students classified “proficient” on the ELPA are those who are subsequently high-performing on the SAT9 – evidence which could confirm the predictive validity of the ELPA for inferences related to successful application of basic English-language skills on content-based tests.

In the case of the present study, ELPA and SBAA scores and classifications for both non-ELL and ELL students provide a unique opportunity to investigate not just mean differences and classification accuracy, but also the predictive ability of the classification for educational program placement as well as federal reporting. However, there are some limitations of these data to consider. While the non-ELL students are a representative random sample of the entire population of native English speakers, there is potential for selection bias due to randomization at the school level and thus opportunity sampling at the student level. The ELL students are a non-random sample as the entire population that was tested; however, there is still potential for selection bias due to a reduced range of proficiency distributions within

the subset of cases selected for this study. The choice to not include reclassified Fully English Proficient or ELL-newcomer students was based on the non-availability of FEP data and the fact that ELL-newcomers take a different test form (Cluster 3-5, Form 1), and the author acknowledges the limited generalizability that will result. In addition, unmeasured factors (such as absenteeism, student motivation, or access to high-quality instruction) may have had an effect on scores and thus the inferences made about those scores.

Children in Transition

While all ELL students are in need of careful consideration on the part of decision-makers, perhaps no child is more in need of an appropriate class placement than an ELL fifth grader going into sixth grade. From a socio-cultural perspective, children entering secondary schooling (which generally starts in sixth and continues through twelfth grade) face unprecedented changes, both formatively and academically. For ELL students especially, the demands of school take a steep climb starting in sixth grade where the use of academic language by multiple teachers in multiple classroom settings creates new challenges to expressing knowledge and being understood (Bailey, 2007). How ELL students are labeled and where they spend their instructional minutes starting in sixth grade has been shown to be predictive of future academic outcomes (Kim & Herman, 2009, Parker, Louis, & O'Dwyer, 2009). This raises the stakes for the accuracy and consistency of the composite scheme and points to the need for its careful calibration in light of, and in conjunction with, other sources of proficiency data. In high-stakes decision making – such as placement of FEP English learners into general or gifted classes – the cost of *false-positives* or *false-negatives* should be carefully weighed in terms of access to instruction at each grade level (Bailey & Kelly, 2010; Xiong & Zhou, 2006). Getting the FEP placement decision right for students entering sixth grade can mean equity and opportunity for future high school placement and likely positive academic outcomes including a greater chance of high school graduation (Kim, 2011).

Research Questions

To examine available evidence to substantiate the claim that these ELPA classifications are valid for predicting FEP-eligible levels of English-language proficiency, the following research questions were explored:

(1) Does the performance of English Language Learner (ELL) cases differ significantly from that of native English speaking (non-ELL) cases on all subsections of the State A English Language Proficiency Assessment (ELPA) and Standards-Based Achievement Assessment (SBAA)?

As was found in prior studies, it was hypothesized that non-ELL cases may score higher on average than ELL cases due to the higher percentage of fluent English proficiency in the non-ELL group. Effect sizes were expected to be largest on the SBAA as its domains measure content standards which depend on not only basic language skills, but also academic language skills and, importantly, meaningful access to grade-level content instruction. Effect sizes on the ELPA were expected to be small if there are a large number of ELL students ready to be reclassified as FEP, but no pre-existing data was available to make a prediction about the prevalence of likely FEP-eligible students. For non-ELL cases, however, it was hypothesized that using the current conjunctive classification scheme, 100% of non-ELL cases would be classified "proficient" and thus confirm what was already known — that these non-ELL cases have a pre-existing status as FEP-eligible, and already enjoy successful placement in English-only classrooms albeit practically by default.

(2) To what extent does the State A ELP using a conjunctive scheme classify as "proficient" all ELL and non-ELL cases who are FEP-eligible on the SBAA by State A criteria ("Basic", "Proficient", or "Advanced" in *Reading*) within the full sample, and

within a group of High-Performing SBAA cases ("Proficient" or "Advanced" in all domains, *Reading, Language Usage, Math* and *Science)*?

It was expected that there would be a convergence of FEP-eligible rates between the ELPA and SBAA. For cases scoring "Proficient" or "Advanced" on the SBAA, a high level of convergence (100%) was expected. For all cases in the High-Performing SBAA group, a high level of convergence (100%) was expected. For all non-ELL cases that were FEP-eligible on the SBAA, an ELPA "proficient" rate of 100% was expected.

(3) Which classification scheme (Conjunctive, Compensatory-1, Compensatory-2, or Mixed) provides the best convergence of FEP-eligible ELPA classifications with FEP-eligible SBAA performances using State A criteria ("Basic", "Proficient" or "Advanced" in *Reading*) for the full sample, and for those in a group of High-Performing SBAA cases ("Proficient" or "Advanced" in all domains (*Reading*, *Language Usage, Math*, and *Science*)?

Differences between FEP-eligible percentages on SBAA and those generated by ELPA schemes provide an approximation of the number of students whose FEP-eligible performances would not be convergent and thus would likely require a special, individualized classification review for an overall FEP decision. Based on the substantive argument that FEP eligible levels of English-language proficiency could be present even with somewhat uneven profiles among the four modalities tested on the ELPA, it was expected that a mixed model scheme would provide the best convergence with FEP-eligible levels on the SBAA for both ELL and non-ELL students. Furthermore, as English-language proficiency is *sine qua non* for reaching "Proficient" and "Advanced" levels on the SBAA, the SBAA scores at these levels were expected to be highly concordant with ELPA "proficient". It was hypothesized that all

schemes would be 100% predictive of proficiency for the non-ELL students scoring "Proficient" or "Advanced" in all SBAA domains, and for all cases in the "High-Performing SBAA" groups.

Method

Data Source

Data were collected by the State A Department of Education in the spring of 2010 from two groups of K-12 students: English Language Learners (ELL, *n*= 14,513), and native English speakers (non-ELL, *n*=1049). The ELL students in the original data set are all those enrolled in the State, and the non-ELL students in this sample were chosen at random from districts who agreed to participate in the study. All subjects completed the State A ELPA and SBAA.

Sample

For this study, one grade level (fifth) was extracted for analysis comprised of ELL (*n* = 1119) and non-ELL (*n* = 103) students. Cases designated as FEP (formerly classified as ELL) were not part of these data. Cases designated as newcomers (in the United States less than 12 months) are given the "Cluster 3-5, Form 1" test and thus were excluded (ELL, *n*=60*)* as well as any other cases administered the "Form 1" test (ELL, *n*=10; non-ELL, *n*=4). Therefore all cases retained had been administered the "Cluster 3-5, Form 2" test. All cases that received testing accommodations or had a special education designation were purposively excluded and await separate analysis (ELL, *n*=153). Furthermore, cases, were duplicates (ELL, *n*=1), were missing ELPA and SBAA-*Reading* or SBAA-*Language Usage* data (ELL, *n*=12, non-ELL, *n*=3) or had conflicting information about native language status (ELL students listed as native English speakers *n=8*, non-ELL students listed as native Spanish speakers *n*=2) were excluded for the purposes of this study. The final sample used for this study (*N*=967) was comprised of fifth grade ELL (*n* = 875) and non-ELL (*n* = 92) students who all took the same ELPA ("Cluster 3-5, Form 2") and SBAA (fifth grade).

Measures[6]

State A English Language Proficiency Assessment (ELPA)

Structure: The State A ELPA is comprised of tests in four sub-domains – *Listening, Speaking, Reading,* and *Writing* - and is administered to each student by grade cluster and test form difficulty. Fifth graders in this study took the "Cluster 3-5" (third, fourth and fifth grade), at form level two which is for non-newcomer ELL students (i.e., students who have been in the United States for more than twelve months).

Test Administration Procedures: The ELPA is an untimed test which is group administered for some sub-domains (*Writing*, part of *Reading*) and individually administered for others (*Listening, Speaking* and part of *Reading*). *Listening* is administered with a compact disc recording which the examiner pauses so students may respond to the recorded questions.

Scoring: Answers to multiple-choice items are recorded by students on test booklets and sent to the test vendor for machine scoring. The constructed-response items in *Speaking* are scored by the examiner in real time; however, the student responses are not recorded. *Writing* and *Reading* constructed-response items are sent to the test vendor for scoring. Most raters selected for hand-scoring hold four-year degrees plus prior experience as raters, and were selected based on teaching credentials, past scoring experience, and performance data. Scoring guides are used to train raters and include test items, rubrics, sample student responses, and annotations. Each student response is read and scored by one rater, with 20% read by a second rater.

Reported Scores: Student performances in each sub-domain are reported in terms of raw score, scale score, and proficiency level. An additional sub-domain of *Comprehension* is also reported and is calculated by combining *Listening* and *Speaking* raw scores but is not used in the calculation of "proficiency". The raw score is the total number of correct answers

[6] The information in this section was adapted from publically available technical documents with the permission of the State A Department of Education.

on multiple-choice items plus the number of points earned on open-ended items. Raw scores on the ELPA can only be compared for the same sub-domain, grade cluster and test form. Scale scores are derived from raw scores and can be compared for the same sub-domain and grade cluster. For the total score, five proficiency levels are reported: Beginning (1), Advanced Beginning (2), Intermediate (3), Early Fluent (4), and Fluent (5). Within each sub-domain, three proficiency levels are reported: Beginning (B), Advanced Beginning to Intermediate (AB+), and Early Fluent to Fluent (EF+).

State A Standards-Based Achievement Assessment (SBAA)

Structure: The State A SBAA is a computer-administered (but not computer-adaptive), multiple-choice test administered to students in grades 3-10 which measures academic content standards in four domains: *Language Arts – Reading, Language Arts – Language Usage, Mathematics,* and *Science.*

Language Arts – Reading: The reading standards, goals and objectives for each grade are covered in two sections: (1) Reading Process (e.g., vocabulary including context clues, affixes, synonyms, antonyms, use of headings and graphics), and (2) Comprehension & Interpretation (e.g., main idea, relevant details, inference, comprehension of literary and expository texts, literary devices and figurative language, plot structure and text organization).

Language Arts - Language Usage: The language usage standards, goals and objectives for each grade are covered in two categories: (1) Writing process (e.g., writing for a specific purpose and audience; selecting a main idea, support, and relevant details; organizing ideas into paragraphs/essays; revising to clarify meaning; using a variety of sentences) and (2) Writing components (e.g., sentence structure, spelling, capitalization, punctuation, and grammar).

Mathematics: The math standards, goals and objectives for each grade are distributed among five categories: (1) Numbers and Operations; (2) Concepts and Principles of Measurement; (3) Concepts of Algebra and Functions; (4) Concepts and Principles of Geometry; (5) Data Analysis.

Science: The science standards, goals and objectives for grades 5, 7 and 10 are distributed among five categories: (1) Nature of Science; (2) Physical Science; (3) Biology; (4) Earth and Space Systems; and (5) Personal and Social Perspectives: Technology.

Scoring: Students complete the computer-based test online which the test vendor stores, scores, and processes to produce reports. Automatic and management-initiated audits are in place ensure the accuracy and reliability of the reports. Performances are reported in raw scores (not available for this study), scale scores, and four proficiency levels ("Below Basic", "Basic", "Proficient" and "Advanced").

Procedures

The ELPA was administered to all ELL students in the State as well as the sample of non-ELL students randomly selected by the State for research purposes during the regularly scheduled testing window [February 22 – April 2, 2010]. The SBAA was administered to all ELL and non-ELL students during the regular testing window [April 12 through May 7, 2010]. Administration of all assessments was coordinated through the State A Department of Education.

Data Analysis Plan

To answer the first research question, the first set of analyses replicated prior studies in conducting an independent samples t-test[7] to compare scores on the ELPA and SBAA. To extend prior studies, effect sizes were calculated and percent ELPA proficient and non-

[7] Statistical analyses were performed using the statistical software package PASW (SPSS) for Windows (version 18.0, IBM Inc., Chicago, IL).

proficient for each group was calculated according to the conjunctive classification model currently in use.

To answer the second research question, descriptive statistics for SBAA levels, and ELPA overall levels were compiled by group in order to identify likely numbers of ELPA proficient cases. Additionally, a "High-Performing SBAA" group was created (scoring "Proficient" or "Advanced" in *all* domains: *Reading, Language Usage, Math and Science)* to represent a population with nearly indisputable English-language proficiency. Background variables as well as percent ELPA proficient and non-proficient according to the conjunctive classification model were reported.

To answer the third research question, "proficient" determinations were re-evaluated using three classification schemes from two additional models. The first scheme was Compensatory-1, which set "proficient" (FEP-eligible) as achieving passing levels on at least three of four sub-domains. The second scheme, Compensatory-2, set "proficient" as an overall ELPA level of four (Early Fluent) or five (Fluent), calculated by the test vendor as a total raw score cutoff of 76 out of 100. The third scheme, Mixed, set "proficient" as an overall raw score of 80 out of 100 with an additional criteria of passing all four sub-domains at or above the lower-bound of the 95% confidence interval of each cut score (*Listening* cut score = 21/25, lower-bound = 17; *Speaking* cut score = 21/25, lower-bound = 17; *Reading* cut score = 19/25, lower-bound = 15; and *Writing* cut score = 16/25, lower-bound =12)[8]. Descriptive statistics were compiled to compare the findings from all four classification schemes in light of predicted levels of proficiency based on State A SBAA FEP-eligible criteria. The number of cases with incongruent FEP-eligible data (ELPA "non-proficient" at SBAA FEP-eligible levels of "Basic" or above; and ELPA "proficient" at SBAA FEP-ineligible level of "Below Basic") were calculated for each scheme. All schemes were then applied to High-Performing SBAA group and descriptive

[8] See Appendix B, Table B2 for the calculation of 95% confidence intervals.

statistics were similarly compiled to compare each classification rate to a predicted level of 100% proficient.

Results

This sample contained ELL (*n*=875) and non-ELL (*n*=92) students who were similar in gender distribution (*percent male*: ELL= 52.2%; non-ELL = 46.7%). On standard background variables, differences were seen in *percent free or reduced-price lunch* (ELL= 80.3%, non-ELL=22.8%), *percent identified as gifted and talented* (ELL= 0.1%, non-ELL=8.7%), *percent homeless* (ELL=1.6%, non-ELL=0%) and *percent Title I* (ELL = 75.1%, non-ELL = 40.2%). Due to exclusion criteria applied for this study, there remained no cases identified as receiving special education services, nor testing accommodations.

None of the non-ELL students retained in this sample reported a native language other than "English", while the ELL students had thirty-two specified and four unspecified languages. The largest native language group was Spanish (85.9%) followed by North American Indian (2.2%), Russian (1.5%), Arabic (1.1%) and Bosnian (0.9%). The ethnicity reported for the non-ELL students was 93.5% White, 2.2% American Indian/Alaskan Native, 1.1% Multiracial, and 3.3% unreported. For the ELL students, the ethnicity reported was 85.1% Hispanic of any race, 5.7% White, 3.4% Black/African-American, 2.4% Asian, 2.3% American Indian/Alaskan Native, 0.5% Native Hawaiian/Pacific Islander, 0.3% Multiracial, and 0.2% unreported.

Descriptive statistics of all performance variables (ELPA, SBAA) were compiled to determine the mean, standard error of the mean, median, range, standard deviation, variance and distributional qualities (see Appendix B, Table B1). Cases of missing data were excluded pairwise to take advantage of all possible analyses. Data were analyzed to check that the assumptions of independence, normality, absence of outliers, and homogeneity of variance were not violated.

Mean Differences, Effect Sizes and Classifications

To answer the first research question, "*Does the performance of English Language Learners (ELL) students differ significantly from that of native English speaking (non-ELL) students on all subsections of the State A English Language Proficiency Assessment (ELPA) and Standards-Based Achievement Assessment (SBAA)?"* means and standard deviations were complied for both groups on both assessments (see Table 2).

Table 2: *Means (and standard deviations) of scores on English Language Proficiency Assessment [ELPA] and Standards-Based Achievement Assessment [SBAA] for ELL and non-ELL students*

	ELPA sub-domains and overall				
	Listening	*Speaking*	*Reading*	*Writing*	*Total*
ELL	21.37 (2.97)	21.20 (3.48)	21.04 (3.63)	18.13 (3.50)	81.75 (10.77)
Non-ELL	22.36 (2.60)	23.21 (1.90)	23.14 (2.55)	21.54 (2.51)	90.25 (6.97)
	SBAA domains				
	Reading	*Language Usage*	*Math*	*Science*	
ELL	205.57 (8.06)	206.81 (8.52)	210.10 (8.13)	200.00 (7.26)	
Non-ELL	218.99 (11.77)	218.77 (13.01)	221.77 (13.38)	210.42 (10.91)	

Note: N = 967; ELL (*n* = 875), non-ELL (*n* = 92). For the ELPA, raw scores are reported out of 100 total (25 for each sub-domain); for the SBAA, scale scores are reported (max scores for Reading=257, Language Usage=258, Math =263, and Science=253). For *Math* (ELL, *n=873, non-ELL n=92) and Science (ELL, n=873; non-ELL, n=89)* missing data were excluded pairwise.

The means and standard deviations reported in Table 2 illustrate that on average, non-ELL students scored higher on all assessments although a notably high mean level of performance was achieved by both groups on the ELPA (total score mean for ELL = 81.75, non-ELL = 90.25). An independent sample t-test was conducted and effect sizes were calculated (see Table 3).

Table 3: *Independent sample t-test between ELL and non-ELL students on measures of ELPA and SBAA*

	df	t	Effect size (η^2)[a]
ELPA – Raw			
Listening	965	-3.08***	0.01
Speaking	965	-5.43***	0.03
Reading	965	-5.40***	0.03
Writing	965	-9.11***	0.08
Total Raw Score	965	-7.41***	0.05
SBAA – Scale			
Reading	965	-14.43***	0.18
Language Usage	965	-12.07***	0.13
Math	963	-12.15***	0.13
Science	960	-12.22***	0.13

Note: N = 967; ELL (*n* = 875), non-ELL (*n* = 92), with missing data excluded pairwise. [a]Eta squared calculated by dividing t^2 by the product of t^2 and df. ****p* ≤ 0.002 *(2-tailed significance test)*

For all assessments, differences between mean scores for ELL and non-ELL students were statistically significant ($p \leq 0.002$). Effect sizes varied indicating that the differences, while statistically significant, may not be practically significant in all cases. For the ELPA sub-domains, effect sizes considered small in Cohen's terms were seen in *Listening* ($\eta^2 = 0.01$), *Speaking* and *Reading* (both $\eta^2 = 0.03$), while Writing ($\eta^2 = 0.08$) had the largest effect, considered moderate in Cohen's terms. The overall effect size for the total ELPA score ($\eta^2 = 0.05$) indicates that the practical difference between ELL and non-ELL scores on the ELPA is small. Whereas, the effect sizes of the SBAA subtests ranged from $\eta^2 = 0.13$ to $\eta^2 = 0.18$ which in Cohen's terms are large effects indicating that the practical difference between groups is large.

Descriptive statistics were then compiled to determine the number of cases classified proficient by the ELPA which uses the conjunctive model which defines "proficient" (FEP-eligible) as an "all four sub-domain pass" (see Table 4).

Table 4: *ELPA classification: Conjunctive (all four sub-domain pass)*

	Proficient	Non-Proficient
ELL (*n=875)*	377 (43.1)[a]	498 (56.9)
Non-ELL (*n*=92)	67 (72.8)	25 (27.2)

[a]*All values in parentheses are percentages according to row.*

For this first research question, most of the findings were as expected. As hypothesized, non-ELL students scored higher on average than ELL students. Also as expected, the effect size differences on the SBAA provided greater distinction between groups than the ELPA. However, an unexpected finding was the smaller-than-predicted percentage of non-ELL students who were classified "proficient" by the ELPA: 72.8% versus the predicted 100%. For ELL students, 43.1% were classified "proficient" despite high mean scores on all ELPA sub-domains.

Descriptive Analyses of SBAA and ELPA Levels

For the second research question, "*To what extent does the State A ELP using a conjunctive scheme classify as "proficient" all ELL and non-ELL cases who are FEP-eligible on the SBAA by State A criteria ("Basic", "Proficient" or "Advanced" in Reading) within the full sample, and within a group of High-Performing SBAA cases ("Proficient" or "Advanced" in all domains, Reading, Language Usage, Math and Science)?"* descriptive statistics were compiled and analyzed.

In Table 5, SBAA levels are reported for ELL and non-ELL cases. Of ELL cases, 13% were "Below Basic" on *Reading*, 23.8% on *Language Usage*, 13.2% on *Math*, and 16.6% on *Science*. Of non-ELL cases, 1.1% were "Below Basic" on *Reading*, 6.5% on *Language Usage*, 4.3% on *Math*, and 3.4% on *Science*. For ELL cases, 114 (13%) failed to meet the FEP-eligible criterion according to the SBAA by scoring "Below Basic" in *Reading*. For non-ELL cases, 1 (1.1%) failed to meet this criterion. The majority of cases met the SBAA FEP-eligible criterion by scoring "Basic" or above in *Reading:* 761 ELL (87%) and 91 non-ELL (98.9%).

Table 5: *Counts (and percentages) of ELL and non-ELL cases by SBAA levels by domain*

	Below Basic	Basic	Proficient	Advanced
SBAA Reading				
ELL (*n*=875)	114 (13.0)[a]	249 (28.5)	453 (51.8)	59 (6.7)
Non-ELL(*n*=92)	1 (1.1)	8 (8.7)	36 (39.4)	47 (51.1)
Total (*N*=967)	115 (11.9)	257 (26.6)	489 (50.6)	106 (11.0)
SBAA Language Usage				
ELL (*n*= 875)	208 (23.8)	314(35.9)	321 (36.7)	32 (3.7)
Non-ELL (*n*=92)	6 (6.5)	11 (12.0)	38 (41.3)	37 (40.2)
Total (*N*=967)	214 (22.1)	325 (33.6)	359 (37.1)	69 (7.1)
SBAA Math				
ELL (*n*= 873)	115 (13.2)	336 (38.5)	372 (42.6)	50 (5.7)
Non-ELL (*n*=92)	4 (4.3)	13 (14.1)	31 (33.7)	44 (47.8)
Total (*N*=965)	119(12.3)	349 (36.2)	403 (41.8)	94 (9.7)
SBAA Science				
ELL (*n*= 873)	145 (16.6)	558 (63.9)	149 (17.1)	21 (2.4)
Non-ELL (*n*=89)	3 (3.4)	26(29.2)	37 (41.6)	23 (25.8)
Total (*N*=962)	148 (15.4)	584 (60.7)	186 (19.3)	44 (4.6)

[a]*All values in parentheses are percentages according to row*

These findings provide one estimate of how many ELL (87%) and non-ELL (98.9%) cases may have also reached FEP-eligible levels on the ELPA. Another indicator of proficiency is the "Overall Level of ELPA" (as seen in Table 6), where data indicate that a majority of cases achieved Level 4 (Early Fluent) or 5 (Fluent): 79.3%[9] of ELL (*n*=694), and 94.5% of non-ELL (*n*=87). However, as reported in the previous research question, only 43.1%[10] of ELL (*n*=377) and 72.8% of non-ELL (*n*=67) cases and were classified ELPA "proficient" by the conjunctive model (see Table 4).

Table 6*: Distribution of overall level on English Language Proficiency Assessment for ELL and non-ELL cases*

	Level 1 Beginner	Level 2 Advanced Beginner	Level 3 Intermediate	Level 4 Early Fluent	Level 5 Fluent
ELL (*n*=875)	3 (0.3)[a]	15 (1.7)	163 (18.6)	372 (42.5)	322 (36.8)
Non-ELL (*n*=92)	-	-	5 (5.5)	17 (18.5)	70 (76.1)

[a]*All values in parentheses are percentages according to row.*

One additional set of analyses was conducted to evaluate the ability of the conjunctive classification to predict proficiency of "High-Performing SBAA" cases ("Proficient" or "Advanced" in all SBAA domains: *Reading, Language Usage, Math,* and *Science*). Analysis of background variables (see Table 7) illustrate that these High-Performing SBAA students are similar to the full sample.

Table 7*: Background variables on "High-Performing SBAA"[a] cases and full sample*

	Gender (% Male)	Free or Reduced-Price Lunch	Gifted and Talented	Homeless	Title I
ELL					
High-Performing (*n*=115)	62 (54)[b]	92 (80)	1 (1)	1 (1)	90 (78)
Full Sample (*n*=875)	457 (52)	703 (80)	1(<1)	14 (2)	657 (75)
Non-ELL					
High-Performing (*n*=59)	30 (51)	11(19)	8(14)	0	25 (42)
Full Sample (*n*=92)	43 (47)	21(23)	8 (9)	0	27 (40)

[a]*"High-Performing SBAA" cases scored "Proficient" or "Advanced" in all SBAA domains (Reading, Language Usage, Math, and Science).* [b]*All values in parentheses are percentages according to row, rounded to the nearest integer.*

[9] The proportion of "Early Fluent"/"Fluent" cases seen here for fifth grade is not unusual for State A. For comparison, the proportion for fifth grade ELL students in 2009-10, when not employing this study's exclusion criteria, was 71.0%. For the three years prior, it was 71.9% (2008-09), 67.2% (2007-08), and 70.1% (2006-07).

[10] The percent proficient for fifth grade ELL students in 2009-10, when not employing this study's exclusion criteria, was 38.9%. In the prior year (2008-09) it was 39%.

When classified by the conjunctive model, only 70.4% of ELL (*n*=81) and 89.8% of non-ELL cases (*n*=53) were classified "proficient". This falls well below the expectation that of 100% "proficient" classifications for this High-Performing SBAA group.

Table 8: *ELPA classification for "High-Performing SBAA"[a] (n=174): Conjunctive*

	ELPA Proficient	ELPA Non-Proficient
ELL (*n*=115)	81 (70.4)[b]	34 (29.6)
Non-ELL (*n*=59)	53 (89.8)	6 (10.2)

[a]"High-Performing SBAA" cases scored "Proficient" or "Advanced" in all SBAA domains (Reading, Language Usage, Math, and Science). [b]All values in parentheses are percentages according to row.

In this second research question, some findings were as expected and some were not. SBAA performances indicate that 87% of ELL cases have reached FEP-eligibility for the content-based requirement on a test given in English, indicating the likely co-occurrence of ELPA FEP-eligibility. A standard of "Overall level of ELPA at 4 (Early Fluent) or 5 (Fluent)" was met by 79.3% of ELL cases, which also indicates the likely co-occurrence of ELPA FEP-eligibility at these rates. The convergence of these two indicators suggests that the ELPA may be a reasonably good predictor of "proficiency" as defined by SBAA FEP-eligible criteria. This is an expected finding. However, as reported in research question one, only 43.1% of ELL cases (*n*=377) were classified "proficient" according to the conjunctive scheme. Even among High-Performing SBAA cases expected to be 100% proficient, only 70.4% (*n*=81) were classified as such. These findings, and the large discrepancy between these indicators, were unexpected.

The number of probable proficient non-ELL cases should be 99% (according to cases meeting State A FEP criteria), an expectation in line with findings from an "Overall Level of ELPA at 4 or 5" standard met by 94.5% of these students. The convergence of these two indicators was expected. However, as reported in the previous research question, only 72.8% of non-ELL cases were classified "proficient" according to the conjunctive scheme. This finding was unexpected, and the extent of the discrepancy between indicators was surprising. Among

High-Performing SBAA non-ELL cases (expected to be 100% proficient) only 89.8% were classified "proficient" according to the conjunctive scheme. This is closer to expectations, likely due to the higher percent of SBAA "Advanced" cases in the non-ELL group versus the ELL group. However, overall, the number of non-ELL cases classified "proficient" was much lower than expected.

Alternate Classification Models

To answer the third research question, "*Which classification scheme (Conjunctive, Compensatory-1, Compensatory-2, or Mixed) provides the best convergence of FEP-eligible ELPA classifications with FEP-eligible SBAA performances using State A criteria ("Basic", "Proficient" or "Advanced" in Reading) for the full sample, and for those in a group of High-Performing SBAA cases ("Proficient" or "Advanced" in <u>all</u> domains: Reading, Language Usage, Math, and Science)?"* data were reclassified using three additional schemes from two models. Findings from each analysis are provided in the tables to follow (Tables 9-11) and a summary table (Table 13) provides information about all four models in terms of convergence with SBAA FEP-eligibility.

Compensatory Model – Scheme One

To explore a compensatory model to where low performance in one sub-domain can be compensated for by higher performance on another, two models were chosen. The first considers "proficient" as passing at least three of four sub-domain at or above currently set cut scores (Listening 21/25; Speaking 21/25; Reading 19/25; Writing 16/25). The results are found in Table 9.

Table 9: *ELPA Classification: Compensatory-1 (three or more sub-domain pass)*

	ELPA Proficient	ELPA Non-Proficient
ELL (*n*=875)	635 (72.6)[a]	240(27.4)
Non-ELL (*n*=92)	84 (91.3)	8 (8.7)

[a]*All values in parentheses are percentages according to row.*

Compensatory Model – Scheme Two

The second compensatory model was used to reclassify the data. In this model, "proficient" is considered as attaining an "Overall ELPA Level" of Level 4 (Early Fluent) or Level 5 (Fluent). These reported levels apply the following raw score cuts for the fifth grade form 2 test: Beginning = 0-30; Advanced Beginning = 31-49; Intermediate =50-75; Early Fluent=76-86; Fluent=87-100. Thus, under this scheme "proficient" includes all cases with raw scores of 76-100. The results are found in Table 10.

Table 10: *ELPA Classification: Compensatory-2 (Overall ELPA Level 4 or 5)*

	ELPA Proficient	ELPA Non-Proficient
ELL (*n*=875)	694 (79.3)[a]	181(20.7)
Non-ELL (*n*=92)	87 (94.6)	5 (5.4)

[a]*All values in parentheses are percentages according to row.*

Mixed Model

To explore the impact of taking standard error of measure into account while also applying a substantive argument in favor of a minimum passing level, a mixed model was created. In this model, cases are classified "proficient" if they achieved an overall raw score of 80% or higher, and scored within a 95% confidence interval of the established cut scores on all four sub-domains (lower-bound of 95% confidence interval: *Listening*=17; *Speaking*=17; *Reading*=15; *Writing*=12). The results are found in Table 11.

Table 11: *ELPA Classification: Mixed (Overall 80% or higher, all sub-domains at or above lower-bound of 95%CI)*

	ELPA Proficient	ELPA Non-Proficient
ELL (*n*=875)	591 (67.5)[a]	282 (32.2)
Non-ELL (*n*=92)	84 (91.3)	8 (8.7)

Note: CI=confidence interval.[a]*All values in parentheses are percentages according to row.*

As illustrated in Table 12, these findings show that the conjunctive model currently in use classifies more non-ELL cases as "non-proficient" than any other model (27.2% compared to 8.7% and 5.4%). This model also classifies the fewest ELL cases as "proficient" (43.1% compared to 67.5%, 72.6% and 79.3%). Furthermore, differences between FEP-eligible

percentages on SBAA and those generated by ELPA schemes (approximating of the number of students whose overall FEP decision would require a special, individualized classification review) illustrate how widely the conjunctive model differs from the others. Although 100% convergence was not expected at all performance levels within FEP-eligibility on the SBAA, the conjunctive model misses the mark by 43.8% for ELL, and 26.0% for non-ELL cases, which is in stark contrast to the Compensatory-2 scheme where the rate of incongruence is 7.6% for ELL and 5.4% for non-ELL.

Table 12: *Percentage (and count) of ELL and non-ELL cases FEP-eligible according to SBAA and ELPA (N=967)*

	FEP-eligible: SBAA[a]	FEP-eligible: ELPA by Classification Scheme[b]			
		Conjunctive	Compensatory-1	Compensatory-2	Mixed
ELL (*n*=875)	87% (761)	43.1% (377)	72.6% (635)	79.3% (694)	67.5% (591)
incongruence		*43.8% (384)*	*14.4% (126)*	*7.6% (67)*	*19.4% (170)*
Non-ELL (*n*=92)	99% (91)	72.8% (67)	91.3% (84)	94.6% (87)	91.3% (84)
incongruence		*26.0%(24)*	*8.7% (8)*	*5.4% (5)*	*8.7% (8)*

[a]State A SBAA FEP-eligibility criteria= "Basic" or above in Reading; [b]ELPA schemes: Conjunctive = all four sub-domain pass; Compensatory-1 = three or more sub-domain pass; Compensatory-2 = overall ELPA level four or five; Mixed = overall score 80% or higher plus all sub-domain scores at or above lower-bound of 95% confidence interval.
[c]All percentages are according to ELL/non-ELL row within each column.

Convergent Validity

In order to determine the extent to which the incongruence between SBAA FEP-eligible and ELPA FEP-eligible classifications was limited to cases at the lower levels of performance, a cross tabulation was created. The information provided in Table 13 shows the extent to which students scoring at all levels of SBAA would be considered "proficient" or "non-proficient" according to each ELPA classification scheme. Cases that are incongruent (ELPA "non-proficient" at SBAA FEP-eligible levels of "Basic" or above; and ELPA "proficient" at SBAA FEP-ineligible level of "Below Basic") are summed to determine the extent to which each model could trigger a special, individualized classification review for an overall FEP decision.

Table 13: *Concordance of FEP-eligible criteria according to SBAA levels and ELPA classification schemes*

	ELPA Classification Schemes							
	Conjunctive		Compensatory-1		Compensatory-2		Mixed	
	Prof	Non-Prof	Prof	Non-Prof	Prof	Non-Prof	Prof	Non-Prof
SBAA-*Reading*								
Advanced								
ELL (*n=59)*	50	9[a]	59	0[a]	58	1[a]	58	1[a]
Non-ELL (*n=47)*	42	5[a]	46	1[a]	47	0[a]	47	0[a]
Proficient								
ELL (*n=453)*	254	199[a]	400	53[a]	422	31[a]	384	69[a]
Non-ELL (*n=36)*	24	12[a]	33	3[a]	34	2[a]	33	3[a]
Basic								
ELL (*n=249)*	57	192[a]	140	109[a]	169	80[a]	117	132[a]
Non-ELL (*n=8)*	1	7[a]	5	3[a]	2	6[a]	4	4[a]
Below Basic								
ELL (*n=114)*	16[a]	98	36[a]	78	45[a]	69	32[a]	82
Non-ELL (*n=1)*	0[a]	1	[a]	1	0[a]	1	0[a]	1
Incongruent[a] Cases	Conjunctive		Compensatory-1		Compensatory-2		Mixed	
ELL (*n=875)*	16	400	36	162	45	112	32	202
Non-ELL (*n=92)*	0	24	0	7	0	8	0	7

[a]*Cases with incongruent FEP-criteria evidence which could trigger a special, individualized classification review.*
[b]*ELPA schemes: Conjunctive = all four sub-domain pass; Compensatory-1 = three or more sub-domain pass; Compensatory-2 = overall ELPA level four or five; Mixed = overall score 80% or higher plus all sub-domain scores at or above lower-bound of 95% confidence interval.*

Some findings were as expected. For cases scoring "Advanced" on the SBAA-*Reading*, the percentage classified ELPA "proficient" came close to 100% for ELL cases by three models (Compensatory-1 – 100%; Compensatory-2 – 98%; Mixed – 98%) but only 85% by Conjunctive. Similar results were found for the non-ELL cases (Compensatory-1 – 98%; Compensatory-2 – 100%; Mixed – 100%) with only 89% by Conjunctive. To a large degree regardless of ELL/non-ELL status, this indicates a rate of incongruence as high as 15% and as low as 0% for cases scoring "Advanced".

Most findings, however, were unexpected. Although 100% of cases scoring "Proficient" on the SBAA-*Reading* were predicted to be classified by the ELPA as "proficient", the models produced quite different results. For ELL cases, the Conjunctive scheme classified only 56% as "proficient", followed by the Mixed at 85%, the Compensatory-1 at 88%, and the

Compensatory-2 at 93%. Similarly for the non-ELL cases, with the Conjunctive scheme classifying only 67% as "proficient", followed by Compensatory-1 and Mixed at 92%, and Compensatory-2 at 94%. This indicates a rate of incongruence as high as 44% and as low as 6%.

Even more incongruent were the ELPA classifications for cases scoring "Basic" on the SBAA-*Reading*, predicted in large part to be classified ELPA "proficient". For ELL cases, the Conjunctive scheme classified only 23% as "proficient", followed by Mixed at 47%, Compensatory-1 at 56%, and Compensatory-2 at 68%. Similarly for non-ELL cases, the Conjunctive scheme classified only 13% as "proficient", followed by Mixed at 50%, Compensatory-1 at 62.5%, and Compensatory-2 at 75%. This suggests a rate of incongruence as high as 87% and as low as 25%.

For cases scoring "Below Basic" in SBAA-*Reading*, the models differ greatly in the percent they classify ELPA "proficient". For ELL cases, the Conjunctive scheme classifies the least (14%) followed by Mixed (28%), Compensatory-1 (32%), and Compensatory-2 (39%). This suggests a rate of incongruence as high as 39% and as low as 14%. The extent to which this represents *false-positive* misclassification is not interpretable, however, without examining the score profiles of each student as well as other sources of proficiency data.

When both sources of data are used for the FEP decision, the cases where FEP-eligibility is incongruent are the cases that would need a special, individualized classification review for a FEP decision. For our sample, the conjunctive model creates the highest amount of hand-classification cases at an astounding 47.5% of ELL (*n*=416) and 26.7% of non-ELL cases (*n*=24), with the majority being cases scoring at "Proficient" and "Advanced" on the SBAA. The mixed model cuts that amount nearly in half for ELL cases at 26.7% (*n*=234) and by about two-thirds for non-ELL cases at 7.6% (*n*=7). The schemes from the compensatory model cut that amount even further for ELL cases: Compensatory-1 at 22.6% (*n*=198) and

Compensatory-2 at 17.9% (*n*=157); and for non-ELL cases the rates stayed about the same: Compensatory-1 at 7.6% (*n*=7) and Compensatory-2 at 8.7% (*n*=8).

One final analysis was conducted to assess each classification scheme with a subset of "High-Performing SBAA" ELL and non-ELL cases previously described in the second research question. It was expected that 100% would be classified "ELPA Proficient". For ELL cases, the Compensatory-2 generated the most (98%), followed by Mixed (97%), Compensatory-1 (95%) and Conjunctive (70%). For non-ELL cases, three schemes met the 100% expectation (Compensatory-1, Compensatory-2, Mixed) followed by the Conjunctive (90%). This indicates that even for students with nearly indisputable foundational English language skills, the rate of incongruence ranges from 10-30% with the Conjunctive scheme.

Table 14: *Summary of classification schemes: "High-Performing SBAA"[a] Cases Only (n=174)*

	ELPA Classification Schemes[b]							
	Conjunctive		Compensatory-1		Compensatory-2		Mixed	
	Non-Prof	Prof	Non-Prof	Prof	Non-Prof	Prof	Non-Prof	Prof
ELL (*n=115)*	34	81	6	109	2	113	4	111
Non-ELL (*n*=59)	6	53	0	59	0	59	0	59

[a]"High-Performing SBAA" cases scored "Proficient" or "Advanced" in all SBAA domains (Reading, Language Usage, Math, and Science). [b]ELPA schemes: Conjunctive = all four sub-domain pass; Compensatory-1 = three or more sub-domain pass; Compensatory-2 = overall ELPA level four or five; Mixed = overall score 80% or higher plus all sub-domain scores at or above lower-bound of 95% confidence interval.

Taken together, the findings from the third research question demonstrated that the mixed model provided a marked improvement over the conjunctive model in matching SBAA FEP-eligible percentages, and in minimizing the rate of incongruence. However, findings did not confirm the hypothesis that a mixed model would produce the best results. Overall, the Compensatory-2 scheme (calculated by an overall ELPA level of 4 or 5) provided the best congruence in all areas. The expectation that all schemes would be 100% predictive of proficiency for "High-Performing SBAA" cases was also disconfirmed, although for non-ELL cases the mixed and compensatory schemes achieved this mark.

Discussion

This study explored evidence to determine the validity of inferences of classifications from the State A English Language Proficiency Assessment (ELPA) which uses a conjunctive classification model to determine FEP eligibility. A randomly-selected sample of native English speaker (non-ELL) students served as a language-proficient comparison group for the fifth grade population of English Language Learners (ELL) students and scores from the State A State Standards-Based Achievement Assessment (SBAA) were used as a source of convergent validity to examine FEP-eligible levels on the ELPA. All cases receiving special education and testing accommodations were excluded for the purposes of this study.

The first research question explored the ELL and non-ELL scores on the ELPA and SBAA in terms of mean differences and effect sizes. Non-ELL cases scored higher on average than ELL cases on all domains, as was expected. Effect sizes were small to moderate on the ELPA indicating that only some of the differences were of practical value, which aligns with the expectation that a portion of the ELL cases are as proficient in English as the non-ELL cases according to their performances on this measure. An unexpected finding was the smaller-than-predicted percentage of non-ELL cases who were classified "proficient" by the ELPA: 72.8% versus the predicted 100%. For ELL, 43.1% of cases were classified "proficient" despite high mean scores on all ELPA sub-domains. As expected, effect sizes on the SBAA were large indicating that the differences were meaningful substantively.

The second research question explored the extent to which students reaching FEP-eligible levels on the SBAA were classified by the ELPA as "proficient" (thus, FEP-eligible). Some findings were as expected, and some were not. Compared to rates of students reaching SBAA FEP-eligibility and students reaching ELPA overall levels of 4 (Early Fluent) or 5 (Fluent), the conjunctive model underestimated the percent proficient for both ELL and non-ELL cases. SBAA performances in *Reading* indicate that 87% of ELL and 98.9% of non-ELL

cases achieved FEP-eligible levels. A standard of "Overall level of ELPA at 4 (Early Fluent) or 5 (Fluent)" was met by 79.3% of ELL and 94.5% of non-ELL cases. This level of convergence of SBAA FEP-eligibility and overall scores on the ELPA was expected and suggests that the ELPA may be a reasonably good predictor of "English-language proficiency" as *sine qua non* for cases meeting SBAA FEP-eligible criteria. However only 43.1% of ELL and 72.8% of non-ELL cases were classified ELPA "proficient" (FEP-eligible) with the conjunctive model. Even among High-Performing SBAA cases expected to be 100% proficient, only 70.4% of ELL and 89.8% of non-ELL cases were classified ELPA "proficient"(FEP-eligible). The large discrepancy between these indicators was unexpected, especially for non-ELL cases, and emphasizes the value of the sampling design used for this study.

The third research question explored two alternative classification models by employing three alternative schemes: two compensatory (Compensatory 1 and 2), and one mixed. Using SBAA FEP-eligible criteria, performances in *Reading* were used to estimate FEP-eligibility on the ELPA, the conjunctive scheme created the highest rate of incongruent results for ELL, non-ELL, and High-Performing SBAA cases. Together, the findings from the third research question demonstrated that the mixed model provided a marked improvement over the conjunctive model in matching SBAA FEP-eligible percentages, and in minimizing the rate of incongruence. However, findings did not confirm the hypothesis that a mixed model would produce the best results. Overall, the Compensatory-2 scheme (calculated by an overall ELPA level of 4 or 5) provided the best congruence at all performance levels within FEP-eligibility on the SBAA for ELL and non-ELL cases. The expectation that all schemes would be 100% predictive of proficiency for "High-Performing SBAA" cases was also disconfirmed, although for non-ELL cases the mixed and compensatory schemes achieved this mark.

Taken as a whole, this study provides a unique investigation of the validity of inferences from FEP-eligible ('proficient") classifications on an ELPA. Prior studies which only utilized

mean differences to interpret the validity of inferences from an ELPA provide an incomplete investigation of validity. This is fully realized when one interprets the assignment of "proficient" and "non-proficient" status within the context of other sources of proficiency data, as was provided in the second and third research questions.

In the larger context of an overall FEP decision for each ELL, the chosen ELPA scheme is often the gatekeeper. For example, in a State or district with a conjunctive ("all evidence pass") model for the overall classification of FEP, cases with incongruent ELPA and SBAA data would not trigger a special, individualized classifications. It is not surprising to find States which prioritize the ELPA, not only being the namesake of the English-language proficiency construct but also the promise of a simpler, fairer FEP process. Indeed, at the inception of NCLB (2001a), researchers heralded the ELPA as superior to the SBAA for FEP decisions (e.g., Linquanti, 2001; Abedi, 2004) and with good reason as many States had up until that point relied *only* the SBAA for such decisions (especially States with small populations of ELL students). However, as English-language proficiency is *sine qua non* for reaching "Proficient" and "Advanced" levels on the SBAA, if not also "Basic" levels, FEP-eligible SBAA performances can reasonably be considered as evidence for convergent validity when ELPA classifications may be called into question.

From test development to test score use, it is the *inferences* which carry the burden of validity – not the assessment. As ELL status is not meant to be permanent (Linquanti, 2001), most ELL students can acquire English in 4-7 years at most when the conditions are right (Hakuta, Butler & Witt, 2000), and annual measurable achievement objectives expect growth of one proficiency level for each ELL student each year (NCLB, 2001c), it is time for the measurement community take a more serious look on the effect of classification models and schemes on classification inferences with these outcomes in mind. Although standard-setting procedures may be carefully followed in the setting of sub-domain cut scores, this practice

alone does not guarantee that the FEP-eligible ELPA standard is effectual. If a classification scheme currently in use is generating upwards of 26% misclassification, a larger problem emerges – namely, the backlog of students waiting to be reclassified in the next assessment cycle. Although the current data does not indicate that this backlog is being adequately resolved, the ability to estimate the likely number of ELPA proficient students likely depends on additional factors not measured in this study such as school attendance and program quality. That being said, an inflation of *false-negative* cases among fifth grade ELL students transitioning into sixth grade has serious ramifications for their long-term academic success (Kim, 2011; Walqui et al., 2010; Xiong & Zhou, 2006) and deserves careful, urgent action.

This study does not intend to recommend one best classification model for all States and all ELPA assessments. Rather, findings are intended to stimulate thinking and discussion of how we interpret the risk and cost of a poor-fitting classification model – and particular schemes used to calculate FEP-eligible performances – in relation to how classifications are interpreted for high-stakes decisions. Furthermore, findings are intended to open a discussion on the way incongruent sources of FEP-eligibility data are interpreted when making overall FEP decisions in terms of the allocation of educational services for ELL students. Beyond the typical psychometric discussion of *false-positives* and *false-negatives*, the implications of incongruent FEP-eligibility data are in need of a substantive, theoretical review based on evidence which could extend the conversation beyond the limitations of this study. Such a discussion would ideally provide both policy- and school-level guidance.

Validation studies such as this one, using a modest number of non-ELL students, are not expensive and could easily be considered a meaningful, necessary step in the judicious following of the *Standards* (AERA, APA, & NCME, 1999). Non-ELL and ELL students differ in important ways due to the unique, multifaceted processes of first and second language acquisition. Even so, when calibrating classification schemes it is practical to consider the use

of a “known-to-be-proficient” population for that process. For the inferences needed for high-stakes decisions, the ELPA classification system needs to be one that accurately and consistently indicates when ELL students have reached a level of English-language proficiency which can be adequately supported with the resources of the general education or gifted classroom. When interpreted in conjunction with ELPA performances, FEP-eligible SBAA performances of non-ELL and ELL students have provided one way to estimate that level. Furthermore, the ELPA performances of ELL and non-ELL students considered High-Performing in all domains on the SBAA (notably, not disproportionate to the full sample in regards to socio-economic status) provide another viable way to evaluate the effectiveness of a classification scheme.

Limitations

There are some limitations of these data to consider. While the non-ELL cases were a representative random sample of the entire population of non-ELL students, there was potential for selection bias due to randomization at the school level only and thus opportunity sampling at the student level. The ELL cases were a non-random sample as the entire population of ELL students was tested; however, there is potential for selection bias due to a reduced range of proficiency at both ends of the distribution as the groups chosen for this study did not include ELL-newcomer or reclassified FEP students. In addition, unmeasured factors (such as absenteeism, student motivation, or access to high-quality instruction) or the time between ELPA and SBAA assessments (ranging from 10 to potentially 74 days) may have had an effect on scores and thus the inferences made about those scores. Also, as this was only one grade level, generalization to other grade levels or test forms is limited without further investigation.

Future Research Directions

Identifying the likelihood that students who are actually proficient will be classified as non-proficient (or proficient) can shed light on the influence of measurement error in the classification process and alert decision-makers to the magnitude and frequency of such errors. This knowledge, for each grade level and cluster test, can lead to the development of more effective and efficient classification systems. Additionally, the timing of ELPA and SBAA data delivery is known to impact the relevance (i.e., are the data still current?) and prioritization of data for use in classification-based decisions. Examining how decision-makers interpret and prioritize concurrent data (i.e., ELPA and SBAA delivered in June) or non-concurrent data (i.e., ELPA data delivered in March and SBAA data delivered in August) for FEP and class-placement decisions could inform our understanding of how data delivery structures impact FEP classification rates.

Unfair score-based decisions can have profound consequences on test takers as well as society at large as these consequences could also raise concerns about the fairness of the test (Xi, 2010). Research could also be conducted to investigate the potential washback effect of incongruence between FEP-eligible criteria in terms of attitudes or perspectives on the accuracy and usefulness of ELPA and/or SBAA scores in the FEP decision process. In States and districts where incongruence occurs but does not trigger a special, individualized classification review, such as in a conjunctive FEP decision-making model, research is needed to determine to what extent the ELPA classification scheme prolongs ELD services for students already FEP-eligible on the SBAA. Findings from these studies could inform to what extent "choice of ELPA classification scheme" may contribute to the known negative effects of long-term ELL status (e.g., lack of school persistence, Kim, 2011). Furthermore, examining school-level factors not included in this study, such as quality of instruction and relative percentage of

ELL population, could provide insight into what causes certain schools and districts to have higher annual rates of FEP reclassifications.

As next generation ELP assessments now being developed by State Departments of Education and test vendors, it is important for researchers to come together with common standards of evaluation. The measurement community is now preparing to publish an update to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, *forthcoming*) which will hopefully provide additional guidelines for how classifications models can be verified and the level of transparency which can be expected from test vendors regarding proprietary schemes.

Concluding Remarks

In closing, it is fitting to evoke Bernard Spolsky (2008) who made the following observation in a reflection on the history of language testing:

> *When we realized over 100 years ago the inevitability of error in the measurement of human capacity (Edgeworth, 1890), we set out to try to reduce the size of the error, rather than trying to understand the risk of making decisions about the fate of human beings using erroneous data. As we learnt the complexity and the enormous variation of human communicative skills and knowledge, we ignored the challenge to produce interpretable but rich profiles and agreed to build unidimensional scales. Even as we discovered the intricate co-construction of normal conversation, we chose to take the abstract formalization of idealized but non-existent monolingual speakers of standard languages as a norm against which to measure real language use (McNamara & Roever, 2006). (p. 302)*

This study held even native English speakers against the standard set by these non-existent monolingual speakers. And the non-existent monolingual speakers won. Despite the intuitive appeal of an "all sub-domain pass" to measure English-language proficiency, the conjunctive model is simply not a logical fit for non-ELL or ELL students. The authentic, uneven language performances of even the highest performing non-ELL cases indicate that there is still careful thinking to be done about what constitutes true language proficiency, and how it can best be

measured. While English language proficiency assessments in all fifty States have the potential to produce rich, interpretable data for use in identifying FEP-eligible levels of English, there is still work to be done. The educational community has a role to play in ensuring inferences made from ELP assessments are part of an effective comprehensive system of proficiency classification. Inferences made from faulty classifications, especially during high-stakes decisions, can have a profound effect on English learners and their academic future. State and district leaders may consider providing expert teams to work with test vendors to better specify the choice and verification of their ELPA classification scheme using student performance data, as lack of model fit could carry a high cost to States in inflated measurement costs and poorly served long-term ELL students. States and districts may also consider providing professional development to address how inferences of ELPA scores are interpreted for high-stakes decisions along with carefully specified guidelines for interpreting incongruent sources of data. Above all, test developers and State assessment teams could examine how scores are interpreted for every use of an ELP assessment and ensure steps are taken to lessen the impact of misclassification.

Appendix A

Glossary

Classification Model	A decision rule or an approach for combining multiple indicators based on a theoretical stance related to the nature of the scores or indicators being combined and the purposes for which the classifications will be used
Composite Scheme	The algorithm, formula or specified combination chosen to convert multiple scores from one or more sources of data (e.g., subtests, tests, or other rated performances) into a dichotomous "master" or "non-master" determination
ELL	An "English Language Learner" is a student identified by the Home Language Survey as potentially eligible for English Language Learner services and whose subsequent scores on the English Learner Placement test indicated a Limited English Proficient status
ELPA	An English Language Proficiency Assessment which is administered to all Limited English Proficient students annually, and to reclassified Fluent English Proficient students in their two-year monitoring period
FEP	Fluent English Proficient – an ELL who met all the requirements of proficiency and was reclassified as FEP. These students must maintain proficiency levels for two years of ELPA and SBAA testing to continue to be considered FEP. At that point, these students will no longer take ELP tests and will no longer count as "ELL" students
FEP-eligible	"Fluent English Proficient"-eligible is the "proficient" cut-point or score determined from sources of English-language proficiency data such as ELPA scores and classifications, SBAA scores and levels, teacher evaluations, parental opinions, and/or a comparison of skills relative to native English speaker students
HLS	A "Home Language Survey" is a questionnaire given to all families enrolling students in school to determine which students may be eligible for English Language Learner services
LEP	A "Limited English Proficient" student is an ELL who has not yet met the requirements to be reclassified as Fluent English Proficient
Non-ELL	For the purposes of this study, a native English speaker student who was never given the Home Language Survey and never designated as an English Language Learner
SBAA	Standards-Based Achievement Assessment which is administered annually to students in grades 3-10 to measure achievement in multiple content domains (e.g., in State A: *Reading, Language Usage, Math* and *Science*)

Appendix B

Table B1: *Descriptive statistics for English Language Proficiency Assessment [ELPA] and Standards-Based Achievement Assessment [SBAA]*

		N	Min	Max	Mean	S.E. Mean	Med.	SD	Var	Skewness (S.E.)	Kurtosis (S.E.)
ELPA - RAW											
Listening											
	ELL	875	6	25	21.37	0.10	22	2.97	8.80	-1.46 (.08)	3.31 (.17)
	Non-ELL	92	10	25	22.36	0.27	23	2.60	6.78	-2.09 (.25)	6.20 (.50)
Speaking											
	ELL	875	4	25	21.20	0.12	22	3.48	12.12	-1.49 (.08)	3.10 (.17)
	Non-ELL	92	18	25	23.21	0.20	24	1.90	3.59	-.92 (.25)	-.04 (.50)
Reading											
	ELL	875	4	25	21.04	0.12	22	3.63	13.19	-1.68 (.08)	3.43 (.17)
	Non-ELL	92	14	25	23.14	0.27	24	2.55	6.52	-1.86 (.24)	3.35 (.50)
Writing											
	ELL	875	3	25	18.13	0.12	19	3.50	12.25	-.87 (.08)	1.14 (.17)
	Non-ELL	92	12	25	21.54	0.26	22	2.51	6.32	-.94 (.25)	1.17 (.50)
Total Raw											
	ELL	875	26	98	81.75	0.36	84	10.77	116.05	-1.65 (.08)	4.21 (.17)
	Non-ELL	92	69	100	90.25	0.73	92	6.97	48.61	-1.11 (.25)	.78 (.50)

		N	Min	Max	Mean	S.E. Mean	Med.	SD	Var	Skewness (S.E.)	Kurtosis (S.E.)
ELPA - SCALE											
Listening											
	ELL	875	78	143	116.34	0.43	115	12.85	165.11	.41 (.08)	.05 (.17)
	Non-ELL	92	87	143	121.00	1.32	120	12.70	161.34	.10 (.25)	-.13 (.50)
Speaking											
	ELL	875	68	144	116.80	0.52	115	15.42	237.76	.18 (.08)	-.32 (.17)
	Non-ELL	92	102	144	127.24	1.49	129	14.29	204.18	-.04 (.25)	-1.46(.50)
Reading											
	ELL	875	76	146	119.29	0.46	118	13.54	183.29	.15 (.08)	.14 (.17)
	Non-ELL	92	99	146	131.12	1.53	131	14.65	214.66	-.41 (.25)	-1.09 (.50)
Writing											
	ELL	875	72	160	116.57	0.39	118	11.52	132.77	.04 (.08)	1.20 (.17)
	Non-ELL	92	99	160	130.85	1.40	130	13.38	179.01	.50 (.25)	.09 (.50)
Total Scale											
	ELL	875	361	487	431.56	0.63	432	18.74	351.01	-.28 (.08)	.79 (.17)
	Non-ELL	92	409	533	453.43	2.42	452	23.20	538.36	.75 (.25)	1.65 (.45)

		N	Min	Max	Mean	S.E. Mean	Med.	SD	Var	Skewness (S.E.)	Kurtosis (S.E.)
SBAA – SCALE											
Reading											
	ELL	875	179	237	205.57	0.27	206	8.06	65.04	.31 (.08)	.52 (.17)
	Non-ELL	92	195	257	218.99	1.23	219	11.77	138.58	.36 (.25)	.42 (.50)
Language Usage											
	ELL	875	182	258	206.81	0.29	207	8.52	72.66	.42 (.08)	1.56 (.17)
	Non-ELL	92	190	258	218.77	1.36	218	13.01	169.28	.22 (.25)	.37 (.50)
Math											
	ELL	873	188	243	210.10	0.28	209	8.13	66.15	.48 (.08)	.34 (.17)
	Non-ELL	92	193	263	221.77	1.40	223	13.38	179.06	.48 (.25)	.76 (.50)
Science											
	ELL	873	177	225	200.00	0.25	200	7.26	52.65	.29(.08)	.18 (.17)
	Non-ELL	89	190	253	210.42	1.16	210	10.91	118.97	.76 (.26)	1.68 (.51)

Note: N = 967; ELL (*n* = 875), non-ELL (*n* = 92). Missing data excluded pairwise.

Table B2: *Calculation of 95% Confidence Interval for State A English Language Proficiency Assessment [ELPA] for Fifth Grade English Language Learners*

(*N*=1,126)	Alpha	Max	Mean	SD	SEM	95% CI	Cut scores (95% CI)
Listening	0.76	25	21.0	3.3	1.66	±3.25	*21 (17.75, 24.25)*
Speaking	0.80	25	20.9	3.8	1.68	±3.29	*21 (17.71, 24.29)*
Reading	0.81	25	20.4	4.2	1.81	±3.54	*19 (15.46, 22.54)*
Writing	0.75	25	17.4	3.9	1.96	±3.84	*16 (12.16, 19.84)*
TOTAL[a]		100					*77 (66-91)*[b]

Note: Total population of fifth grade English Language Learners taking the Grade 3-5 Form 2 test was used for these calculations.[a]*Total at the cut scores, plus the estimated range of the 95% Confidence Interval.* [b]*Lower-bound scores rounded down and upper-bound scores rounded up prior to adding.*

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational assessment*, 8(3), 231-257.

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher,* 33(1), 4-14.

Abedi, J., (Ed.) (2007). *English language proficiency assessment in the nation: Current status and future practice.* Davis, CA: UC Davis School of Education.

Abedi, J. (2008a). Measuring students' level of English Proficiency: Educational significance and assessment requirements. *Educational Assessment, 13*(2/3), 193-214.

Abedi, J. (2008b). Classification system for English Language Learners: Issues and Recommendations. *Educational Measurement: Issues and Practice,* 27(3), 17-31.

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME] (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bachman, L.F. & Palmer, A. (2010). *Language assessment in practice.* New York: Oxford University Press.

Bailey, A.L. (2007). *The language demands of school: Putting academic English to the test.* New Haven, CT: Yale University Press.

Bailey, A.L. & Butler, F.A. (2003). *An evidentiary framework for operationalizing academic language for broad application to K-12 education: A design document* (CSE Tech. Rep. No. 611). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing.

Bailey, A.L., & Butler, F.A. (2004). Ethical considerations in the assessment of the language and content knowledge of U.S. school-age English learners. *Language Assessment Quarterly,* 1(2-3), 177-193.

Bailey, A.L. & Reynolds Kelly, K. (2010) *The use and validity of home language surveys in state English language proficiency assessment systems: A review and issues perspective.* EAG EVEA project deliverable. Retrieved from http://www.eveaproject.com

Brennan, R.L. (Ed.). (2006). *Educational measurement, fourth edition.* Westport, CT: American Council on Education and Praeger Publishers.

Chester, M.D. (2003). Multiple measures and high-stakes decisions: a framework for combining measures. *Educational Measurement: Issues in Practice,* 22 (2), 32-41.

Clark, L. A. & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3), 309-319.

Clauser, B.E., Clyman, S.G., Margolis, M.J. & Ross, L.P. (1996). Are fully compensatory models appropriate for setting standards on performance assessments of clinical skills? *Academic Medicine*, *71*(1:supplement), S90-S92.

Crane, E.W., Barrat, V.X., and Huang, M. (2011). *The relationship between English proficiency and content knowledge for English language learner students in grades 10 and 11 in Utah.* (Issues & Answers Report, REL 2011 – No. 110). Washington, DC: U.S. Department of Education, Institute of Educational Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West. Retrieved from http://ies.ed.gov/ncee/edlabs.

Davidson, F., Kim, J.T., Lee, H., Li, J., & López, A.A. (2007). English language testing: Evidence from the evolution of test specifications. In A.L. Bailey (Ed.), *The language*

demands of school: Putting academic English to the test (pp. 157-170). New Haven, CT: Yale University Press.

Douglas, K.M. (2007). *A general method for estimating the classification reliability of complex decisions based on configural combinations of multiple assessment scores*. Unpublished dissertation, University of Maryland.

Douglas, K.M. & Mislevy, R.L. (2010). Estimating classification accuracy for complex decision rules based on multiple scores. *Journal of Education and Behavioral Statistics,* 35 (3), 280-306.

Ercikan, K., & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels: Guidelines for assessment design. *Applied Measurement in Education, 3,* 269-294.

Forte, E. & Faulkner-Bond, M. (2010). *The administrator's guide to federal programs for English learners.* Washington, D.C: Thompson.

Forte, E., Perie, M., & Paek, P. (2012). *Exploring the relationship between English language proficiency and English language arts.* EAG EVEA project deliverable. Retrieved from http://www.eveaproject.com.

Florez, I. R. (2012). Examining the validity of the Arizona English Language Learners Assessment cut scores. *Language Policy, 11(1)*, 33-45.

Gándara, P., Rumberger, R., Maxwell-Jolly, J. & Callahan, R. (2003, October 7). English Learners in California Schools: Unequal resources, unequal outcomes. *Education Policy Analysis Archives, 11*(36). Retrieved on October 23, 2011 from http://epaa.asu.edu/epaa/v11n36/.

Garcia, E. E., Lawton, K. & Diniz de Figueiredo, E. H. (2010). *Assessment of young English language learners in Arizona: Questioning the validity of the state measure of English*

proficiency. Los Angeles: Civil Rights Project, University of California. Retrieved from: http://civilrightsproject.ucla.edu/research.

Hakuta, K., Butler, Y. G., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* Policy report, University of California Linguistic Minority Research Institute, UC Berkeley.

Kane, M. T. (2006). Validation. In Brennan, R. L. (Ed.), *Educational measurement,* fourth edition (pp. 17-64). Westport, CT: American Council on Education and Praeger Publishers.

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

Kane, M. & Case, S.M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education*, 17(3), 221-240.

Karatonis, A. & Sireci, S. G. (2006). The Bookmark Standard-Setting Method: A literature review. *Educational Measurement: Issues and Practice*, *25*(1), 4-12.

Kim, J. (2011). *Relationships among and between ELL status, demographic characteristics, enrollment history, and school persistence* (CRESST Report 810). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kim, J., & Herman, J.L. (2009). A three-state study of English learner progress. *Educational Assessment, 14(3-4)*, 212-231.

Kim, J. & Herman, J. L. (2010). *When to exit ELL students: monitoring success and failure in mainstream classrooms after ELLs' reclassification*. (CRESST Report 779). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Linquanti, R. (2001). *The reclassification dilemma: Challenges and choices in fostering meaningful accountability for English learners* (UC Language Minority Research Institute Policy Report 2001-01). Davis, CA: University of California, Davis.

Linquanti, R. & George, C. (2007). Establishing and utilizing an NCLB Title III accountability system: California's approach and findings to date. In J. Abedi (Ed.), *English language proficiency assessment in the nation: Current status and future practice* (pp. 105-118). Davis, CA: UC Davis School of Education.

Livingston, S.A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32(2), 179-197.

Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. *Educational Measurement: Issues and Practice,* 27(3), 32-42.

Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing, 24(4)*, 489-515.

MacSwan, J. & Rolstad, K. (2006). How language proficiency tests mislead us about ability: Implications for English language learner placement in special education. *Teachers College Board,* 108(11), 2304-2328.

Mahoney, K. S., & MacSwan, J. (2005). Reexamining identification and reclassification of English language learners: A critical discussion of select state practices. *Bilingual Research Journal*, 29 (1), 31-42.

Mislevy, R. J., Wilson, M. R., Ercikan, K., & Chudowsky, N. (2002). Psychometric principles in student assessment. In T. Kellaghan and D. Stufflebeam (Eds.), *International handbook of education evaluation* (pp. 489-532). Boston: Kluwer Academic.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed.), *Setting Performance Standards:*

Concepts, methods, and perspectives (pp. 249-281). Mahway, NJ: Lawrence Erlbaum Associates.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington DC; American Council on Education/Macmillan.

Mosier, C.I. (1943). On the reliability of a weighted composite. *Psychometrika,* 8 (3), 161-168.

National Research Council. (2011). *Allocating Federal Funds for State Programs for English Language Learners.* Panel to Review Alternative Data Sources for the Limited-English Proficiency Allocation Formula under Title III, Part A, Elementary and Secondary Education Act. Committee on National Statistics and Board on Testing and Assessment. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

No Child Left Behind Act of 2001 (NCLB) (2001a). Conference Report to Accompany H.R., 1, Rep. No. 107-334, House of Representatives, 107th Congress, 1st Session, December 13. Pub. L. No. 107-110, 115 Stat. 1425.

No Child Left Behind Act of 2001 (NCLB) (2001b). Title I: Improving the academic achievement of the disadvantaged. 107th Congress, 1st Session, December 13. (Printed version Washington, DC: George Washington University, National Clearinghouse for Bilingual Education.)

No Child Left Behind Act of 2001 (NCLB) (2001c). Title III: Language instruction for limited English proficient and immigrant students. 107th Congress, 1st Session, December 13. (Printed version Washington, DC: George Washington University, National Clearinghouse for Bilingual Education.)

Parker, C. E., Louis, J., & O'Dwyer, L. (2009). *New measures of English language proficiency and their relationship to performance on large-scale content assessments* (Issues & Answers Report, REL 2009-No. 066). Washington, DC: U.S. Department of Education,

Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. Retrieved from http://ies.ed.gov/ncee/edlabs.

Porter, S.G., & Vega, J. (2007). Overview of Existing English Language Proficiency Tests. In J. Abedi (Ed.) *English language proficiency assessment in the nation: Current status and future practice* (pp. 93-104). Davis, CA: UC Davis School of Education.

Ragan, A., & Lesaux, N. (2006). Federal, state, and district level English language learner program entry and exit requirements: Effects on the education of language minority learners. *Education Policy Analysis Archives*, *14*(20). Retrieved on October 6, 2011 from http://epaa.asu.edu/epaa/v14n20/.

Robinson, J. P. (2011). Evaluating criteria for English learner reclassification: A causal-effects approach using a binding-score regression discontinuity design with instrumental variables. *Educational Evaluation and Policy Analysis*, 33(3), 267-292.

Shepard, L.A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-8, 13.

Shepard, L.A. (1993). Evaluating test validity. *Review of Research in Education*, *19*, 405-450.

Sireci, S., Han, K. T., & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment*, 13(3), 108-131.

Solórzano, R. W. (2008). High stakes testing: Issues, implications, and remedies for English language learners. *Review of Educational Research*, 78 (2), 260-329.

Spolsky, B. (2008). Language testing at 25: Maturity and responsibility? *Language Testing*, 25 (3), 297-305.

Stephenson, A., Jiao, H., & Wall, N. (2004). *A performance comparison of native and non-native speakers of English on an English language proficiency test.* San Antonio, TX: Harcourt Assessment, Inc. Retrieved from http://www.pearsonassessments.com.

Stephenson, A., Johnson, D. F., Jorgensen, M. A. & Young, M. J. (2003). *Assessing English Language Proficiency: Using Valid Results to Optimize Instruction.* San Antonio, TX: Harcourt Assessment, Inc. Retrieved from http://www.pearsonassessments.com.

Stone, C., Weissman, A., & Lane, S. (2005). The consistency of student proficiency classifications under competing IRT models. *Educational Assessment,* 10(2), 125-146.

The Stanford English Language Proficiency Test (2003). San Antonio, TX: Harcourt Assessment, Inc.

U.S. Department of Education (2006). *Building partnerships to help English language learners.* Retrieved from: http://www2.ed.gov/print/nclb/methods/english/lepfactsheet.html.

Walqui, A., Koelsch, N., Hamburger, L., et al. (2010). *What are we doing to middle school English Learners? Findings and recommendations for change from a study of California EL programs* (Narrative Summary). San Francisco: WestEd.

Wang, J., Niemi, D., & Wang, H. (2007). *Impact of different performance assessment cut scores on student promotion.* CSE (CRESST) Report 719, 2007.

Wang, M.W., & Stanley, J.C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, 40, 663-705.

Wolf, M. K., Herman, J. L., Dietel, R. (2010). *Improving the validity of English language learner assessment systems* (CRESST Policy Brief No. 10 – Full Report). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

Wolf, M.K., Herman, J.L., Bachman, L.F., Bailey, A.L. & Griffin, N. (2008). *Recommendations for assessing English language learners: English language proficiency measures and*

accommodation uses, recommendation report (part 3 of 3). CRESST Report 737, 2008.

Xiong, Y.S., & Zhou, M. (2006). Structuring inequality: How California selectively tests, classifies, and tracks language minority students. *California Policy Options*, UCLA School of Public Affairs, UCLA. Retrieved from: http://escholarship.org/uc/item/98d66346.

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing*, 27 (1), 119-140.