Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊

Volume 2 | Issue 1 Article 5

2021

A Latent Class IRT Approach to Defining and Measuring Language Proficiency

Tammy D. Tolar
University of Houston

David J. Francis
University of Houston

Paulina A. Kulesz
University of Houston

Karla K. Stuebing
University of Houston

Follow this and additional works at: https://www.ce-jeme.org/journal

 Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation

Tolar, Tammy D.; Francis, David J.; Kulesz, Paulina A.; and Stuebing, Karla K. (2021) "A Latent Class IRT Approach to Defining and Measuring Language Proficiency," *Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊*: Vol. 2 : Iss. 1 , Article 5.
Available at: https://www.ce-jeme.org/journal/vol2/iss1/5

This Article is brought to you for free and open access by Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊. It has been accepted for inclusion in Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊 by an authorized editor of Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊.

Chinese/English Journal of Educational Measurement and Evaluation
Vol 2, 49-73, March 2021

A Latent Class IRT Approach to Defining and Measuring Language Proficiency

Tammy D. Tolar [a], David J. Francis [a], Paulina A. Kulesz [a], and Karla K. Stuebing [a]

[a] University of Houston

Abstract

English language learner (EL) status has high stakes implications for determining when and how ELs should be evaluated for academic achievement. In the US, students designated as English learners are assessed annually for English language proficiency (ELP), a complex construct whose conceptualization has evolved in recent years to reflect more precisely the language demands of content area achievement as reflected in the standards of individual states and state language assessment consortia, such as WIDA and ELPA21. The goal of this paper was to examine the possible role for and utility of using content area assessments to validate language proficiency mastery criteria. Specifically, we applied mixture item response models to identify two classes of EL students: (1) ELs for whom English language arts and math achievement test items have similar difficulty and discrimination parameters as they do for non-ELs and (2) ELs for whom the test items function differently. We used latent class IRT methods to identify the two groups of ELs and to evaluate the effects of different subscales of ELP (reading, writing, listening, and speaking) on group membership. Only reading and writing were significant predictors of class membership. Cut-scores based on summary scores of ELP were imperfect predictors of class membership and indicated the need for finer differentiation within the top proficiency category. This study demonstrates the importance of linking definitions of ELP to the context for which ELP is used and suggests the possible value of psychometric analyses when language proficiency standards are linked to the language requirements for content area achievement.

1 A Latent Class IRT Approach to Defining and Measuring Language Proficiency

Around the world, school and governmental systems are confronted with the challenge of measuring the language proficiency of students, employees, and citizens. This challenge has increased in recent years as globalization leads to increasing numbers of individuals who speak a language other than the societal language, but whose academic and economic fortunes are tied to their proficiency in the societal language. In the US, language proficiency assessments are most commonly used in public schools to assess the linguistic competencies of language minority students to ensure their ability to benefit from instruction conducted in English. Common descriptors used to characterize *functional* linguistic competence (i.e., ready to participate in regular instruction without linguistic support) focus on the ability to fluently interact with native speakers, to understand the main ideas (both concrete and abstract) when presented in complex texts and speech, and to produce complex written and oral arguments.

Insofar as many *native* speakers of a language may struggle to achieve this level of linguistic competence because they lack declarative knowledge, or the ability to understand complex arguments involving abstract topics, measures of language proficiency can confound language competence with academic proficiency at higher levels of thinking. Put another way, there is marked variability in the verbal abilities of native speakers of a given language, and language proficiency assessments seek not to confound individual differences in verbal ability with individual differences in language proficiency. On the one hand, it is unrealistic to expect a person to display a level of language proficiency in a second language (L2) that they do not display in their first language (L1). Exceptions exist, of course, and certainly children acquiring L2 prior to full development of their L1 without continued emphasis on L1 development can achieve a level of competence in L2 that is unmatched in their L1. At the same time, there is clearly considerable variability in verbal ability among native speakers of a language, and this variability

CONTACT: David J. Francis. dfrancis@uh.edu. Texas Institute for Measurement, Evaluation, and Statistics, 4849 Calhoun Rd, Mail Rm 373, Houston, TX 77204.

must be considered in any attempt to establish standards of proficiency in a language among non-native speakers. The basic distinction that we are making is that language proficiency is language specific, but verbal ability is an aptitude for verbal reasoning without regard to a specific language, notwithstanding the fact that verbal ability must be measured in a specific language (usually a person's L1). Regardless of that fact, we can imagine a generalized ability to think and reason verbally, and this ability contributes to performance on achievement tests in reading language arts as well as in mathematics, science, and social studies. We could even argue that such verbal ability is causally implicated in achievement outcomes, although proving as much is challenging. Regardless, we can distinguish between language aptitude (verbal ability) and proficiency with a particular language, even if distinguishing between these constructs is challenging conceptually. It stands to reason, however, that if some native speakers of a language can fail to meet a state's proficiency standards in Reading Language Arts, then some non-native speakers might also fail to meet that standard despite being proficient in the language.

In this study, we evaluate the relationship between language proficiency (as typically assessed with English language proficiency tests) and language based declarative knowledge (as assessed with state achievement tests) through the novel application of modern statistical methods to language proficiency and content area achievement assessments in common use with school-aged children. Defining language proficiency for school contexts has important implications for students and schools. For students, language proficiency is an important determinant of content area achievement (Perfetti, 2007; Vellutino, 2003). Definitions of language proficiency also have important implications for schools (Wolf, Guzman-Orth, & Hauck, 2014) including determination of when a student can be assessed in their L2, or more importantly, when inferences based on achievement scores for L2 speakers have comparable validity for native speakers of the language. Only at that point does the resultant achievement score provide comparably meaningful information for EL and non-EL test-takers to teachers, students, and parents about the student's content area proficiency rather than their language proficiency. At that point, we might also expect that students can function independently in an L2 instructional context without the need for language supports that are not made available to native speakers of the societal language, although the ability for students to function independently in the classroom cannot be assumed [1]. For instance, it is possible that the language proficiency required to be successful in the classroom without language supports may differ substantially from the point where test items function comparably. Such a divergence would certainly be expected if test developers were successful in either minimizing the construct irrelevant language demands of tests, or in providing accommodations that mitigate such demands. However, it is also possible that the language demands for independent participation in instruction are different from those required to native-like functioning of test-items. Still, it seems reasonable that such a point on the test might provide a lower bound for language independence in the classroom, especially as developers become increasingly adept at minimizing or successfully accommodating construct irrelevant language related variance in test items.

1.1 Language-Achievement Connection

Early research on academic achievement and English language proficiency reported mixed findings concerning the relation between language proficiency and academic achievement (Gue & Holdaway, 1973; Hwang & Dizney, 1970; Light, Xu, & Mossop, 1987; Mulligan, 1966). This variability in results may have been a function of inconsistencies in defining and measuring English language proficiency as well as academic achievement (Graham, 1987). For example, English language proficiency definitions have varied along several dimensions including a single, global score versus skill specific scores involving listening, writing, reading, and vocabulary, as well as conversational versus academic language proficiency (Francis & Rivera, 2007).

More recent results are more conclusive as they focus on relations between academic achievement and skill specific scores rather than general/global language proficiency scores. Academic achievement is correlated with English reading performance among college students from different language and cultural backgrounds (Bayliss & Raymond, 2004; Dooey & Oliver, 2002; Kerstjens & Nery, 2000; Oliver, Vanderford, & Grote, 2012). English language proficiency also predicts academic achievement among younger students. For example, Ardasheva, Tretter, and Kinny (2012) investigated English proficiency as a predictor of academic success in reading and mathematics among middle school students and found that academic achievement depends on English proficiency level. Former

[1] We wish to thank an anonymous reviewer for sharing this insight.

English language learners with a high competence in English performed better on reading and mathematics state assessments when compared with the native English speakers and English language learners with lower English competence.

1.2 Assessing and Understanding Achievement Among English Language Learners

An *English language learner* (EL) is a student designated by the state as limited English proficient. As of the fall of 2017, approximately 5.0 million students (10.1% of all students) in the U.S. were designated as ELs (Hussar et al., 2020), an increase of 1.2 million students from Fall of 2000, when ELs comprised 8.1% of all students. In some states ($n = 11$ in Fall 2017), more than 10% of students are ELs (e.g., 19.2% in California, and 18% in Texas). They are a heterogeneous population in terms of time or age of arrival, prior school experience, parental education, degree of economic and social advantage/disadvantage, and home language (August & Shanahan, 2006).

ELs are required to be monitored in terms of language proficiency and content area achievement under federal education law. There is a persistent achievement gap between ELs and non-ELs as measured by the National Assessments of Educational Progress (NAEP), with non-ELs outscoring ELs by the same margin since 1998 in reading and 1996 in math (Kena et al., 2014). EL's also perform more poorly on state tests (U.S. Department of Education, 2011) although it should be noted that there is wide variability across states in levels of reported proficiency and in the level of parity between ELs and non-ELs. This variability may be a function of a variety of factors which vary by state including academic content, assessments of achievement and English language proficiency, achievement and language proficiency standards, exclusionary criteria, and test accommodation policies. Other factors, such as economic risk, exacerbate the difference between ELs and non-ELs to the extent that the factor negatively affects achievement while also being more prevalent in the EL population than in the overall population.

Even if the factors discussed above were controlled or eliminated, comparisons between ELs and other subgroups from state and federal accountability systems may bias results against ELs for at least three other reasons that merit consideration. First, the defining characteristic (viz. language proficiency) is causally linked to the outcomes of interest (viz., content area achievement). Acquisition of English is a consequence of effective instruction and mediates instructional effects on content mastery (Calderón, Hertz-Lazarowitz, & Slavin, 1998; Carlo et al., 2004). Second, state achievement tests may not be valid measures of achievement among ELs. Although typically designed to be unidimensional measures of achievement domains (e.g., reading/language arts, math), state achievement tests may be multidimensional measures among ELs. Some of the variability in test performance among ELs may reflect construct irrelevant variance associated with language proficiency. The same source of language variability does not exist among native English speakers. Thus, there are potentially two types of language related variability in content area achievement tests. One such source of language variability is construct relevant and is present for both ELs and non-ELs, whereas the second source of language variability is construct irrelevant and affects only the performance of ELs on content area achievement tests. It is this latter type of variability that test developers seek to control through test design and test accommodations, and that schools attempt to control through limiting participation in the content area achievement test. Finally, unlike all other demographic groupings (e.g., gender, ethnicity, learning disabilities), membership in the EL category is dynamic. Students are placed in the group when language proficiency is low and lose their membership as they acquire English. This dynamic nature of the designation of a student as EL overstates the difference in academic competence between EL and non-EL students; as ELs gain proficiency in English they no longer count in the EL category. This accounting problem has led Saunders and Marcelletti (2013) to refer to the EL achievement gap as "the gap that can't go away" because comparisons of achievement are between students not yet proficient in the language and those who are proficient in the language. While the difference reflects the role of language in content area achievement, it overstates the achievement differences between students who were ever designated as ELs and those who were never ELs (Umansky, Thompson, & Díaz, 2017).

Heterogeneity in the EL subgroup with respect to language proficiency suggests that any group of ELs may include a mix of at least two populations of EL students: one for whom English proficiency is sufficient that content area achievement tests reflect only construct relevant variance and thus language exerts a comparable impact on achievement as for native English speakers and a second population for whom English proficiency affects the

validity of achievement measures in ways that are distinct from the effects of language on achievement for native English speakers. When the performance of EL students is compared to that of non-EL students the distribution of language proficiency within the EL students will affect the magnitude of the observed difference; interpretation of the observed difference is challenging when this distribution is ignored. Differences between ELs and non-ELs can shift over time, from grade to grade, school to school, or district to district simply because of differences in the distribution of language proficiency within the EL group. It is possible for School A to have superior outcomes for ELs at all levels of proficiency than School B, and yet School A shows a bigger gap in achievement for ELs simply because achievement is related to language proficiency and School A has more ELs at low levels of English proficiency compared to School B, a problem known as Simpson's paradox (Simpson, 1951).

1.3 Measuring Achievement Among EL Students

The validity of national and state achievement measures for ELs is a major source of concern for test developers, educators, researchers, and policy makers (Abedi, 2002; Abedi & Gándara, 2006; Francis & Rivera, 2007; National Research Council, 2000, 2002; Sireci, Han, & Wells, 2008). This issue of test validity is especially problematic for ELs because as described above language is confounded with achievement. Consequently any measure of achievement is inherently a measure of language proficiency (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). An important question raised by states and schools and debated among policy makers is when EL students should be tested on state achievement tests, and when should states be held accountable for the performance of their EL students. Ultimately, the answers to these questions hinge on the validity of the state assessment for making inferences about the achievement proficiency of EL students, which itself rests on the validity of item functioning for EL students in comparison with non-EL students. One approach for evaluating validity has been through studies of differential item functioning (DIF) between ELs and native English speakers (e.g., Mahoney, 2008; Martiniello, 2008; Ockey, 2007; Turkan & Liu, 2012; Young et al., 2008). However, results of these studies have been inconsistent in terms of both the identification and characterization of DIF, ranging from no DIF (Mahoney, 2008) to 25% of items exhibiting at least slight DIF (Martiniello, 2008; Turkan & Liu, 2012). Even when DIF is detected, DIF items do not always favor non-ELs (Turkan & Liu, 2012).

The difference in results across studies may be due to a variety of reasons including differences in outcome measures: e.g., state (Martiniello, 2008) versus national (Ockey, 2007) assessments; math (Martiniello, 2008) versus reading (Pomplun & Omar, 2001); and differences in the way DIF is evaluated: e.g., Mantel-Haenszel (Martiniello, 2008), IRT (Ockey, 2007), CFA (Pomplun & Omar, 2001), and Differential Bundle Functioning (Kim & Jang, 2009). A third potential source of the different results across studies is differences in the distribution of language proficiency within the EL groups being evaluated. As described earlier, ELs are a heterogeneous and evolving population. An obvious source of heterogeneity and change is English proficiency, and the distribution of proficiency within the EL group will affect any comparison with native English speakers.

Some studies have evaluated the effect of English proficiency on DIF by further subdividing the EL group, although the method of determining subgroup membership varies by study, e.g., different state definitions of English proficiency (Pomplun & Omar, 2001; Wolf & Leon, 2009) versus self-reported level of English proficiency (Snetzler & Qualls, 2000). Evidence from these studies suggests that the greater the difference in English proficiency between the focal and reference group, the greater the percentage of items exhibiting DIF. However, using cut-points on measures of English language proficiency (ELP) presumes a clear understanding of the relationship between the different forms of ELP (e.g., oral, written) and academic achievement. In addition, although ELP is the distinguishing feature of ELs, it is not the only source of heterogeneity among ELs. Ethnic and cultural background, years in the U.S., economic risk, and other factors, including some which are not yet known, may be sources of individual differences among ELs as well as sources of DIF. The degree and nature of DIF may differ depending on the mix of these factors among the focal EL group being evaluated. Reliance on a priori defined groups may reduce the power to identify and characterize DIF, and may obscure factors influencing DIF (Samuelsen, 2008). An alternative approach is needed to evaluate how ELP influences DIF and to identify other characteristics of ELs that influence DIF.

1.4 Differential Item Functioning: A Mixture Model Approach

A problem with the methods defined so far is their reliance on a priori groupings of EL students. However, advances in statistical modeling of test scores that combine latent class models with item response models offer a unique alternative to the examination of test validity and DIF as well as other questions related to academic achievement and development among EL students that does not depend on a priori grouping of students. These models are known collectively as mixture IRT models (Mislevy, Levy, Kroopnick, & Rutstein, 2008; Samuelsen, 2008).

The terms *mixture model and latent class model*[2] are sometimes used interchangeably (e.g., in Muthén, 2002) and refer to models that represent data composed of subpopulations with different probability distributions on relevant variables (e.g., different means, variances, covariances; see Lubke & Muthén, 2005; Muthén, 2001, 2008, for general overviews). Because the data include members from the unknown subpopulations, the distribution of the data reflects a mixture of the distributions in the unknown subpopulations. Thus, the distributions are "mixed" together in the data, and mixture/latent class analysis allows for the identification of the different subpopulations and estimation of distributions within each subpopulation as well as a basis for placing individual observations into each unknown subpopulation. The classes are latent because they are not directly observed and are not predefined but are "latent" in the aggregate distribution. Thus, these latent groups stand in stark contrast to known groups such as gender, EL status, or language proficiency categories based on ELP test performance. Latent classes may be defined by distribution parameters of observed variables (e.g., means of a set of continuous measures in latent profile analysis; Marsh, Lüdtke, Trautwein, & Morin, 2009) or latent variables (e.g., means of intercepts and slopes in growth mixture models; Muthén, Khoo, Francis & Kim Boscardin, 2000).

Within a structural equation modeling framework, a unidimensional IRT model is a latent factor model in which items (e.g., in our case items from a state achievement test) load onto a single latent ability factor (e.g., math achievement). Item difficulty and discrimination parameters that are freely estimated in a 2PL IRT model are mathematically related to item thresholds and loadings, respectively (Asparouhov & Muthén, 2016). An additional threshold representing a guessing parameter is estimated for each item if one computes a 3PL IRT model. A mixture IRT model is a latent class model in which the latent classes may differ on item parameters (see Mislevy et al., 2008; Samuelsen, 2008, for discussions of mixture IRT models and their application to evaluations of DIF). Classes may also differ on means and variances of the latent ability factor; however, in the context of measurement non-invariance (i.e., classes differing on item parameters) comparisons of factor means and variances are problematic because they cannot be assumed to represent the same construct across classes (i.e., there are latent factors such as language proficiency in the case of EL students influencing item performance that are not included in the model).

In educational research, Cohen and Bolt (2005) used mixture IRT models to demonstrate that female college students comprise a heterogeneous population because some types of math placement items function differently for some female students. Mixture IRT models have been also used to demonstrate that some students with learning disabilities (LD) look more like non-LD students in terms of item functioning. Choi, Alexeev, and Cohen (2015) also used mixture IRT modeling to evaluate internet access as a predictor of latent class membership among TIMSS 2007 math test takers. The two identified latent classes were distinguished based on a math ability level, with larger proportions of students from high performing countries being members of the high ability class. At the same time, increased access to the internet also corresponded to increased probability of being a member of the high ability class. These studies demonstrate that known groups (e.g., gender, LD status) may not be homogenous, nor are they uniformly prone to DIF on academic assessments. The Choi et al. (2015) study also provided support for not using known groups for DIF in educational research in so much as known groups "are not directly related to the issues of learning that educators care about" (p. 179, Samuelsen, 2008). For example, educators and policy makers may find it more useful to know that internet access influences educational equity rather than knowing that students from lower performing countries are more prone to DIF on international assessments.

Mixture IRT models not only may be used to address questions of test validity, but they may also be used to address substantive questions. These methods offer a unique alternative to evaluating factors that influence academic achievement among ELs because they allow

[2]Latent class analysis may also refer to the specific case of a mixture model in which observed categorical variables are latent class indicators.

us to directly ask whether and how ELP can be used to identify a subgroup of EL students for whom content area achievement test items function the same as for non-EL students. Answering this question has implications for the validity of ELP classifications, but also for the validity of content area achievement test interpretation. At the same time, answering the question is not trivial because we cannot assume that the characteristics that determine homogeneous groupings within the heterogeneous population of ELs are known a priori and thus would lend themselves to classical DIF approaches.

1.5 Study Rationale and Hypotheses

We hypothesize that EL students are composed of a mix of at least two groups of students, some for whom achievement test items function similarly to native English speakers and some for whom the achievement test items function differently. Although there may be more than two latent classes of ELs among the latter, we are primarily focused on identifying students for whom achievement test items perform similarly to non-EL students in terms of item functioning. We further hypothesize that a key variable for identifying these EL students is English language proficiency. More specifically, we expect some ways of measuring English language proficiency (e.g., oral vs. reading proficiency) are better at identifying these EL students than others.

2 Methods

2.1 Participants

Results reported here are based on a single cohort of 4th grade students tested in a northeastern state with a significant EL population. Participants included students not designated as EL (non-EL, $n = 65,415$) as well as EL students ($n = 4,533$). The state database included scores on the state's English proficiency exam for EL students and item level and summary scores for English Language Arts (ELA) and Mathematics (MATH) achievement tests. Among students who were designated ELs, we excluded students who had incomplete English proficiency data and students who were repeating the grade in the assessment year or whose grade information was in some other way discrepant (4.7% of students). For non-EL students, we excluded those whose first language was not English (6.3%; i.e., some students were never designated as EL, although English was not their first language).

From the combined sample of EL and non-EL students we also excluded students if they took either of the alternate achievement tests that are administered to students whose disability prevents them from taking the standard achievement tests, even when accommodations are provided (0.3% and 1.4% of EL and non-EL students, respectively). Students were excluded from analyses for a given achievement test (ELA or MATH) if they did not take the test or there was a discrepancy in the item level data for that test. ELA tests were missing or discrepant for 10.1% of EL[3] and 0.6% of non-EL students. MATH tests were missing or discrepant for 10.0% of EL and 0.4% of non-EL students.

Twenty-four percent of the excluded EL students were missing English proficiency test data. Comparing the remaining excluded students to included students, excluded students were lower in English proficiency than those retained in the study, $\chi^2(3) > 1277, p < .001$. Only 16% of excluded students were at the Transitioning (highest) level of English proficiency whereas 66% of included students were at the Transitioning level (see below for a description of the English proficiency levels). Among the EL students, excluded students were less likely to be on free or reduced lunch (75% vs. 83%, $\chi^2(1) > 22, p < .001$. There were also significant differences in the racial distribution between excluded students and the study sample, $\chi^2(4) > 36, p < .001$. Excluded students were less likely to be Asian (12% vs. 20%) and more likely to be Black (13% vs. 9%) or White (17% vs. 13%), but equally likely to be Hispanic (58%). There were no differences in special education status between excluded and non-excluded students, $\chi^2(1) > 2.38, p > .05$. Finally, there were no differences between excluded and non-excluded students in the gender distribution, $\chi^2(1) < 0.49, p > .05$.

[3]Not taking the state achievement tests was the primary reason for excluding EL students from the study (65% of the excluded cases). We do not have specific information about why students did not take the state achievement tests. However, 95% of the EL students who did not take the state achievement tests had been enrolled in the state's schools for only one year (vs. 4% of EL students who did take the state tests). In addition, 87% of the EL students who did not take the state tests were below the Transitioning English language proficiency level (vs 34% of students who did take the state tests). It is possible that the primary reason EL students not taking the test is a policy related to the number of years they had been in the US (or at least in the state's schools). It's also possible that the reason for not taking the state test is a function of their English proficiency level. The two are confounded and it is not possible to determine the specific criteria that influenced EL students not taking the state tests.

Given the extremely large sample sizes, any differences between the excluded and non-excluded samples of non-EL students were statistically significant. Notable differences between excluded and non-excluded students included that non-EL excluded students were more likely to be on free or reduced lunch (opposite of EL students, 57% vs. 23%), less likely to be white (32% vs. 83%), more likely to be Hispanic (40% vs. 5%) or Asian (17% vs. 3%), and more likely to have special education designation (31% vs. 17%).

In summary, approximately 15% of EL and 8% of non-EL students in the state databases were excluded from the analyses. Excluded EL students were lower on English proficiency, higher in SES, and were less likely to be Asian (12%) than students included in the analyses (20% Asian). Excluded non-EL students were more likely to be non-White and low SES. The final sample included 3,874 EL students (of those 3,853 and 3,855 had ELA and MATH data, respectively) and 60,312 non-EL students (of those 60,025 and 60,159 had ELA and MATH data, respectively). Table 1 shows demographics and English proficiency and achievement test scores by group (EL vs non-EL).

2.2 Measures

English Proficiency Assessment (EPA). The EPA consists of two components: (1) an assessment of reading and writing (EPA-RW) and (2) an assessment of oral language (EPA-O). The EPA-RW, administered to EL students during the spring, consisted of multiple choice, short-answer, and open-response questions. Students were required to read passages and write responses to questions referencing the passages as well as write compositions based on writing prompts. Item Response Theory (IRT) methods were used to calibrate items and equate forms. Scaled scores were computed separately for reading and writing as a function of students' estimated theta scores (i.e., IRT ability scores). EPA-RW scaled scores ranged from 1 to 30. All scaling and scoring were carried out as part of the state assessment program, as opposed to the research team.

The EPA-O is an observational assessment administered by qualified administrators. Each EL student was observed in a classroom setting during academic tasks and social interactions. Students were observed over a month-long period in a variety of classroom activities. A scoring matrix was used to assign scores for listening and speaking. Test administrators were required to participate in a minimum of 9 hours of training, a minimum of 1 hour of practice rating students, and score a minimum of 60% on a qualifying test. Listening was assigned a raw score from 0 to 5. Each subdomain of speaking was assigned a raw score from 0 to 5 resulting in a total raw speaking score ranging from 0 to 20.

Scaled Scores (SS) and General Performance Levels (PL). SS and PL were used by the state for reporting purposes. SS were computed as a function of estimated theta scores which were a function of raw composite scores of the reading/writing and oral language components. SS ranged from 300 to 400. Cut points on the SS scale that defined the four PLs (Beginning, Early Intermediate, Intermediate, and Transitioning) were determined by standard-setting panels. Cronbach's alpha coefficients for EPA composite scores ranged from .89 to .96, depending on testing session. Cohen's Kappa for PL's was .65.

English Language Arts (ELA) and Mathematics (MATH) Achievement Tests. ELA and MATH are the state assessments of achievement required for all public school students including students with disabilities and EL students. The achievement tests measure performance based on state curriculum learning standards. Most of the items on these tests are dichotomously scored. To simplify model estimation, only dichotomously scored ELA and MATH items, regardless of their response format, were used in the analyses.

ELA. Students read passages and answered 36 dichotomously scored multiple choice and 4 open response questions about the passages that were polytomously scored 0 to 4 based on a rubric. The analyses for the current study excluded the four polytomous open response items as estimation of their parameters would require implementation of similar though not identical IRT models relative to the dichotomously scored items. Cronbach's alpha coefficient for the multiple choice items was .86. Analyses were based on raw scores (item scores 0 or 1, sum scores 0-36).

MATH. Students solved 29 multiple choice, 5 short answer, and 5 open response computational and applied problems that included multiple forms of representation (e.g., symbolic, tables, graphs, and word problems). The multiple choice and short answer items were dichotomously scored while the open response items were scored 0 to 4, similarly to the ELA open response items. Open response items were excluded from the analyses for the same reasons as in the ELA measure. Cronbach's alpha coefficient for the multiple choice items was .84 and for the short answer and open response items was .78. Analyses were based on raw scores (item scores 0 or 1, sum scores 0-34).

Table 1
Demographics, English Proficiency and Achievement Scores by EL Status

Demographic/Measure	Non-EL	EL	*p*
		%	
Female	49	48	ns
Race			*
Asian	3	20	
African American	9	9	
Hispanic	5	58	
White	83	13	
Free or reduced price lunch	23	83	*
Special education	17	19	*
		M(SD)	
ELA achievement	78.1 (15.4)	57.9 (19.3)	*
Math achievement	76.7 (16.7)	61.3 (20.0)	
English proficiency			
Composite SS		378.4 (21.2)	
Reading SS		17.3 (3.1)	
Writing SS		17.9 (3.3)	
Listening RS		4.1 (0.9)	
Speaking RS		15.3 (3.6)	
Level		%	
Beginning		4	
Early intermediate		6	
Intermediate		24	
Transitioning		66	

Note. Total sample $n = 60{,}312$ non-ELs, 3,874 ELs; MATH achievement test $n = 60{,}025$ non-ELs, 3,853 ELs; ELA achievement test $n = 60{,}159$ non-ELs, 3,855 ELs. Non-EL are significantly different from EL students ($p < .05$) for all comparisons except gender. SS = Scale Score. RS = Raw Score (only raw scores were available for the listening and speaking subtests). Achievement scores are percent correct out of 36 (ELA) and 34 (MATH) items.

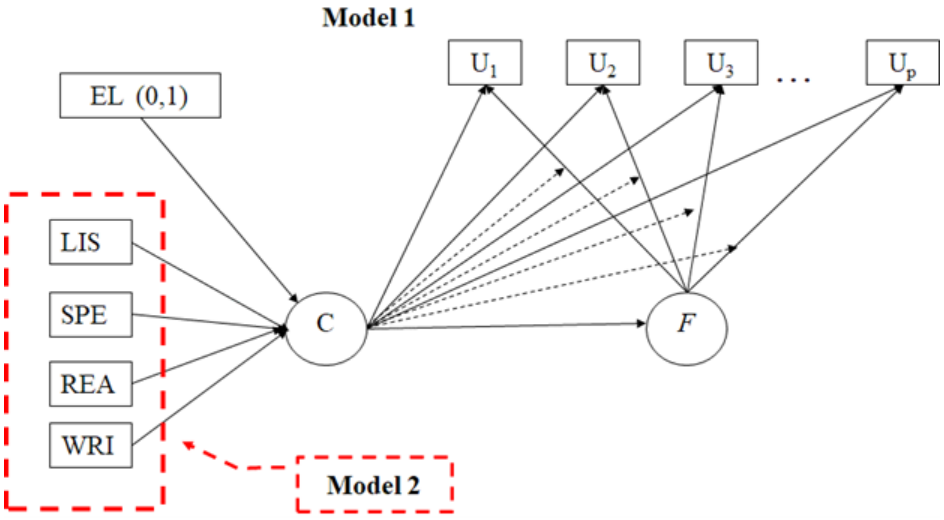
2.3 Analytic Procedures

We evaluated four Item Response Theory (IRT) Mixture Models for each of the achievement tests (ELA, MATH). The models are progressively more complex in terms of the number of parameters to be estimated.

Model 0a and 0b specified two known classes defined by EL status. Model 0a specified item parameters invariant between the non-EL and EL groups. In this model, the non-EL factor mean and variance are 0 and 1 and the EL mean and variance are freely estimated. In Model 0b item parameters were allowed to differ between non-EL and EL students. For model identification purposes, means and variances are fixed to 0 and 1 for both groups. Models 0a and 0b stipulate no latent classes, only known groups, where measurement is either invariant across groups (0a) or non-invariant across groups (0b).

Models 1 and 2. We evaluated two Mixture 2PL IRT Models for each of the achievement tests (ELA, MATH) as depicted in Figure 1. In estimated models, three latent classes were distinguished such that one class was technically a known class (non-EL students) because of the way the latent classes are specified (see below), while the other two classes were truly latent classes of EL students. The two latent classes of EL students

Figure 1
IRT Mixture Models Evaluated



Note. Model 1 includes all variables not enclosed in dashed box. Model 2 adds the variables in the dashed box to Model 1. The symbols are based on conventions adopted by Muthén and Muthén (1998–2017). Boxes labeled "U" represent observed ELA or MATH test item scores. Circle "F" represents latent ELA or MATH achievement scores estimated from the observed item scores. The circle "C" represents latent EL class (non-EL, EL_I, EL_NI). The box "EL (0,1)" represent observed EL status (0 = non-EL, 1 = EL). Boxes" LIS", "SPE", "REA", "WRI" represent observed ELP listening, speaking, reading, and writing scores. Solid arrows indicate regression relationships between variables (e.g., latent EL class, C, predicts latent ELA or MATH achievement and observed test item thresholds, F; F predicts test item performance, U_i). Dashed arrows indicate that the relationship (loadings) between observed item performance, U, and latent achievement, F, is random and may vary depending on latent class, C.

comprised EL students for whom the achievement test items functioned comparably between EL and non-EL students (EL_I, where I means *invariant*), and EL students for whom the achievement test items functioned differently compared to non-EL students (EL_NI, where NI means *not invariant*). During model estimation items were *invariant* between the EL_I latent class and the non-EL known-class, whereas items were *not invariant* between the EL_NI latent class and the non-EL known-class of students. Prediction of latent class membership involved a two-step process. In the first model (Model 1) only observed EL status predicted class membership. In other words, the model was specified such that observed EL status forced non-ELs into the known-class (Class 1) and ELs into the other two latent classes (EL_I, EL_NI). In this model, classification of ELs was based only on the latent class indicators (achievement test item parameters, latent achievement factor means and variances). In the second model (Model 2), the scores for the four English language proficiency skills (listening, speaking, reading, and writing) in addition to EL status were used to predict the class membership for the two latent EL classes, so classification of ELs into the two latent classes was based on ELP. A logit link function related observed EL status and ELP scores to the probability of class membership (see Appendix for Mplus code used to estimate the models with the parameterization of the logit link function). Logit regression coefficients were estimated for the log odds of students being in the EL_I class relative to the EL_NI class. Based on the model parameterization (see Appendix), the non-ELs had a 100% probability of being in the known-class (Class 1). The EL students had some probability between 0.00 and 1.00 of being in each of the other two classes (EL_I and EL_NI) depending on their ELP subtest scores and their achievement test performance.

In both models, achievement test items (thresholds and loadings) and latent achievement factor means and variances were latent class indicators. (i.e., the observed and latent variables in the measurement model were regressed on the latent class variable where the unordered categorical class variable was specified as a set of dummy variables; see Lubke & Muthén, 2005, for a detailed explanation of the regression equations that define factor mixture models). All item thresholds and loadings were forced to be the same for the non-EL and EL_I classes. One

item per achievement test (ELA and MATH) was specified as an anchor item and forced to be the same for EL_NI and the other two classes[4]. Specifying an anchor item allowed for scale comparability between the EL_NI and the other two classes. Factor means and variances were fixed to 0 and 1 for the non-EL group but allowed to differ for the EL_I and EL_NI latent classes. This model specification ensured that estimates of item parameters are not affected by differences in achievement across the three latent classes. Because item parameters of the achievement test (ELA, MATH) are constrained invariant between non-EL and EL_I students, the factors can be assumed to represent the same construct, and the distribution of ability in the EL_I class can be directly compared to the distribution of ability in the non-EL class. At the same time, because item parameters for all items except the anchor items are allowed to differ between non-EL and EL_NI students, it cannot be assumed that the E_NI achievement factor represents the same construct as the achievement factor for the other two groups. If it is the case that the same construct has been measured, the construct has not been measured in the same way, i.e., to the same degree of precision. Although we can compare the distributions of ability across the three classes, comparisons involving the EL_NI class must take into account the lack of measurement invariance.

As described above, each EL student had some non-zero probability of belonging to each of the two EL latent classes (EL_I, EL_NI). Although fit of the latent class models does not involve actual assignment of individuals to classes, comparisons of classes post-hoc on other attributes require that students are assigned to one class or the other. For the purpose of post-hoc comparisons, students were classified into that group for which their probability of group membership was highest. The average value of these probabilities is an important measure of the quality of the model fit, with average probabilities closer to 1.0 signifying better fits and clearer separation between the two latent classes. Simulation evidence suggests that the Bayesian information criteria (BIC) consistently selects the generating model when evaluating mixture IRT models (Li, Cohen, Kim, & Cho, 2009), but because this evidence is limited we used loglikelihood and Akaike's information criteria (AIC) in addition to BIC to evaluate model fit (in addition to average probabilities of class membership). Research on mixture models recommends the use of the bootstrap likelihood ratio test for deciding on the presence of latent classes and the number of classes (Nylund, Asparouhov, & Muthén, 2007).

The models were estimated using *Mplus* (Muthén & Muthén, 1998–2012). In mixture modeling, there is a risk of local solutions (i.e., final solution based on local maxima instead of the global maximum of the likelihood). To decrease the chance of local solutions, multiple sets of starting values should be used. We specified 500 random starting values and final optimizations for the 125 best initial maximum likelihood estimates. Robust maximum likelihood estimation was used for all models. Missing achievement test items ($< 1\%$ of all data for ELs and non ELs) were treated as incorrect (i.e., scored as 0). None of the EL students were missing English language proficiency scores.

3 Results [5]

3.1 Model Fit Comparisons

For both content areas, Model 1 was preferable to Model 0a and 0b, indicating that the specification of a single class of EL students for whom measurement differed from non-EL students provided an inferior fit to the data than allowing for two classes of EL students, one for whom measurement was invariant with non-EL students and one for whom measurement varied relative to non-El students. At the same time, across both content areas and for all fit criteria, Model 2 was the best fitting model (see Table 2 for fit statistics)[6].

The estimated average probabilities of class membership

[4]To determine the anchor items, IRT DIF analysis was conducted in IRTPRO (Cai, du Toit, & Thissen, 2011) with manifest EL and non-EL groups. The non DIF items with the highest scores (p-values) among ELs and non-ELs (i.e., the easiest non DIF item for each test) were used as anchor items.

[5]We report here on results for a single cohort of fourth grade, but have examined this question for two cohorts of grade 4 students, as well as two cohorts of students from grades 5, 6, 7, and 8 in both Math and English Language Arts. Results are remarkably consistent in terms of evidence for the existence of latent classes, the value of using English language proficiency scale scores for classification, and the near zero probability of membership in the EL_I class for students at the lowest three levels of ELP. Other results are available upon request from the authors.

[6]The model fit is based on 2PL IRT parameterization. We attempted to estimate 3PL models, but encountered problems with estimation for the ELA test. The model estimates for this test did not replicate, and the best fitting model produced implausible estimates (i.e., guessing parameters > 1). At the same time, we were able to replicate 3PL models for the MATH test. EL_I and EL_NI classification agreement for the MATH test between 3PL and 2PL models (Model 2) was very high (.96). Because our focus is primarily on EL classification and not on specific item parameters, and because we were able to estimate 2PL models for both MATH and ELA, we only report on the 2PL models.

Table 2
Fit Statistics

Model	Loglikelihood	AIC	BIC	BIC (sample-size adjusted)	Entropy	Average latent class probabilities	
						EL_I	EL_NI
			ELA				
0a	-1016055	2032261	2032941	2032702			
0b	-1015516	2031322	2032637	2032176			
1	-1000882	2002058	2003391	2002924	.973	.763	.764
2	**-1000064**	**2000429**	**2001798**	**2001318**	**.990**	**.920**	**.922**
			MATH				
0a	-996717	1993576	1994220	1993994			
0b	-995953	1992180	1993422	1992987			
1	-981227	1962732	1963992	1963551	.977	.799	.795
2	**-980623**	**1961533**	**1962829**	**1962375**	**.987**	**.888**	**.897**

Note. Model 0a and 0b include only known groups of non-EL and EL students. Model 0a includes two known groups with invariant measurement, whereas 0b includes two known groups without constraints for measurement invariance. Models 1 and 2 include a known non-EL class and two latent EL classes. For one EL class, item parameters are invariant (EL_I) relative to non-EL parameters, and for the other EL class, item parameters are non-invariant (EL_NI) with the exception of an anchor item. EL status alone (Model 1) or EL status and subscale scores (Model 2) predict EL class membership. Bolded values show the best fitting model (within subject) per the relevant fit statistic.

were quite high for Model 2 ranging from .89 to .92 across classes and achievement outcomes (versus .76 to .80 for Model 1). These results suggested that the inclusion of ELP subtest scores (listening, speaking, reading, and writing) as predictors of EL class membership resulted in good separation between EL_I and EL_NI classes. The subsequent analyses are based on Model 2.

3.2 Achievement Test Item Parameter Comparisons (Non-EL vs. EL_NI)

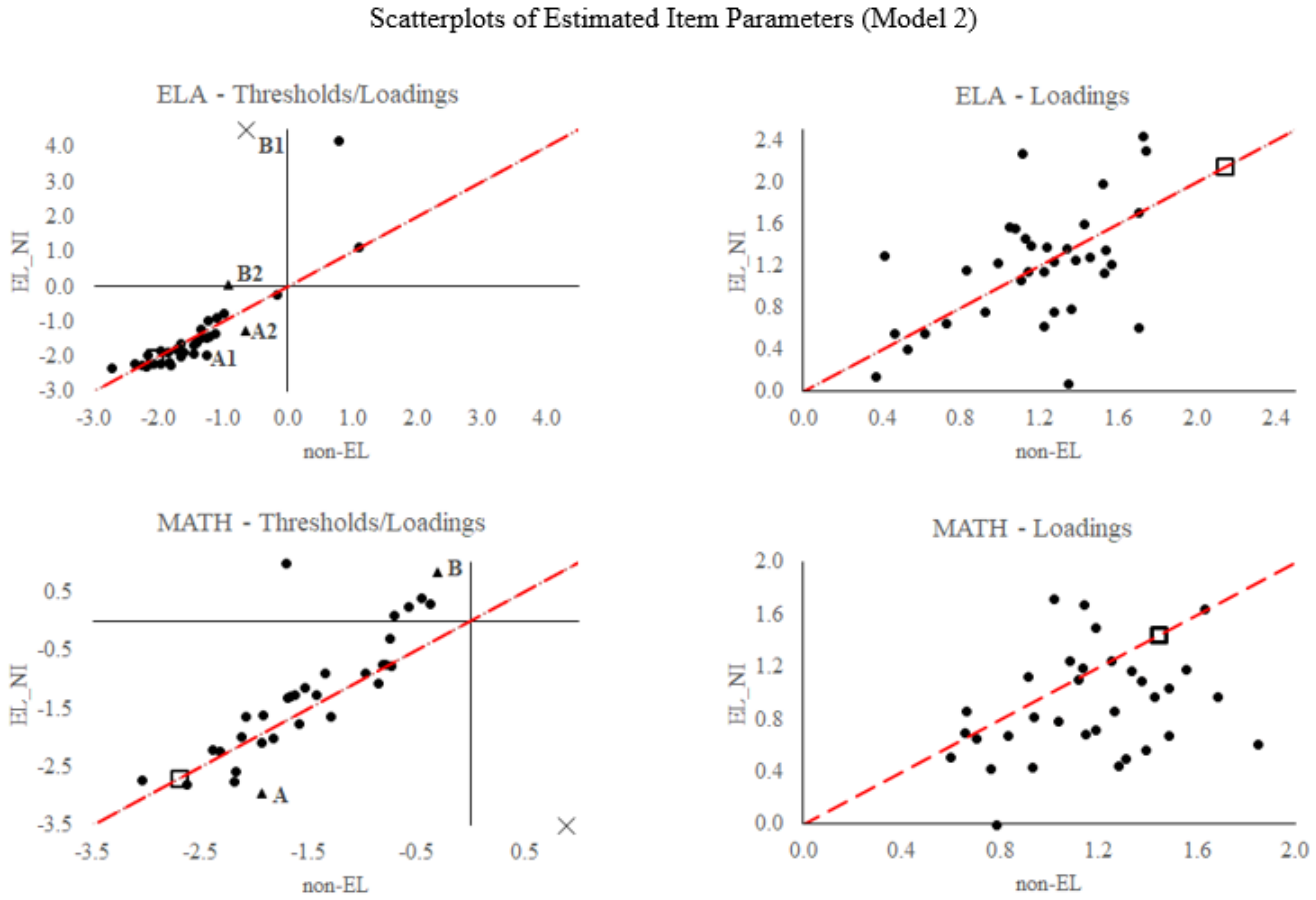
We compared non-EL item parameter estimates (from Model 2) to EL_NI estimates to evaluate how item parameters differ. Figure 2 shows scatterplots for the loadings (analogous to the 2PL IRT *discrimination* parameters) and thresholds divided by loadings (analogous to the 2PL IRT *difficulty* parameters). For MATH, it appears that the test items were generally less discriminating and more difficult for the EL_NI students than the non-EL students. There was also little correspondence in rank order of the discrimination parameters between the two groups ($r = .26$, ns). However, there was a high correspondence in rank order of the difficulty parameters ($r = .92$, excluding an extreme outlier that biased the correlation in the negative direction).

For ELA, it was less clear as to whether items favored non-ELs or EL_NI students, with relatively equal numbers of items favoring non-ELs and EL_NI students when evaluated with discrimination and difficulty parameters. There was a greater correspondence for ELA than MATH in rank order of the discrimination parameters between the two groups ($r = .61$) and difficulty parameters ($r = .87$, excluding one extreme outlier that biased the correlation substantially downward).

As further validation of the two EL classes, we conducted IRT likelihood ratio test (LR) DIF analysis (Thissen, Steinberg, & Wainer, 1993), comparing the non-EL students to the two EL classes (i.e., post hoc comparisons among the sample students by using model-estimated class probabilities to assign the EL students to EL_NI and EL_I groups). This approach was used for two reasons. First, model estimates of item parameters are idealized in the sense that they are based on probabilistic assignment of individuals to the invariant and not invariant class. It is reasonable to expect that item parameter estimates would differ somewhat if based on "manifest" groups of actual students categorized to a class based on a probability < 1.00 of being a member of the assigned class. Second, Model 2 does not directly test which items have meaningful DIF (i.e., statistically and practically significant). A post-hoc DIF analyses allowed us to identify those items and provide additional validity evidence for the mixture IRT models.

Figure 2

Scatterplots Relating Non-EL Item Parameters (Horizontal Axes) to EL_NI Item Parameters (Vertical Axes)



Note. Threshold (τ)/loading (λ) is analogous to the difficulty parameter (b), whereas the loading (λ) is analogous to the discrimination parameter (a) in the 2PL IRT model. Data points represented by open square symbol (□) are estimated parameters for anchor items. Data points labeled "A" on the Thresholds/Loadings scatterplots represent items with the most extreme and statistically significant DIF favoring EL_NI (i.e., items are easier for EL_NI students than non-ELs). Data points labeled "B" represent items with the most extreme and statistically significant DIF favoring non-ELs (see text for discussion of these items). The actual coordinates for ELA Threshold/Loading data point (×) are (-.66, 10.63), where the loading and threshold for EL_NI is 0.09 and 0.94. The actual coordinates for MATH Threshold/Loading data point (×) are (0.89, -37.94) where the loading and threshold for EL_NI is -.04 and .89. The small positive (.09) and small negative (-.04) loadings for these items imply that the items are not related to their respective constructs for EL_NI.

We conducted the post-hoc DIF analyses using IRTPRO (Cai et al., 2011). The IRT-LR DIF methods used in IRTPRO are analogous to the IRT factor models estimated in MPlus. We evaluated total chi-square values for each item and used the Bonferroni correction to adjust for the large number of tests.

For both MATH and ELA there were more items displaying statistically significant DIF for EL_NI than EL_I (MATH: 29 vs 5 items; ELA: 24 vs 12 items). The DIF items varied in whether they favored non-EL or EL students depending on the content area and the EL group being compared. Among EL_I students, all MATH DIF items favored the EL_I students in difficulty; however, all MATH items with DIF favored the non-ELs in discrimination. Conversely, among ELA items showing DIF, more favored the non-ELs in difficulty (55% of items) and more favored the EL_I students in discrimination (77%). For EL_NI students, among MATH items showing DIF more favored the non-ELs in difficulty (61%), but more favored EL_NI students in discrimination (55%). ELA items showing DIF were equally likely to favor non-EL and EL_NI students in difficulty (50%) and more likely to favor EL_NI in discrimination (56%).

To explore the content validity of the DIF analysis, we examined MATH and ELA items displaying the most extreme DIF among EL_NI students. Plotting the ratio of thresholds over factor loadings and the factor loadings, which is analogous to plotting IRT difficulty parameters and IRT discrimination parameters, for EL_NI against corresponding estimates for non-ELs in Figure 2 shows the more extreme, statistically significant DIF items favoring EL_NI students (points marked "A" in the plots) and items favoring non-EL students (points marked "B" in the plots). The MATH item favoring EL_NI students asked students to identify one of four figures that is not a quadrilateral. This item may be easily solved without reading the question stem by using pattern recognition, if the student focuses on the number of sides in each figure. It may also be the case that the use of "quadrilateral," a Spanish-language cognate, makes this item easier for many ELs. The MATH item favoring non-EL students asked students "Which of the following is a list of three fractions that are each equivalent to 0.50?". This item was one of the most difficult for both non-ELs (57% answered correctly) and EL_NI students (26% answered correctly). Although not entirely clear why this item was especially difficult for the EL_NI students relative to the non-EL students, the question construction is linguistically complex. Pattern recognition could be used with this item as well (i.e., ignoring the language and focusing on 0.50 in the question stem and matching it to the only answer option where all fractions are equivalent to .50). However, 44% of EL_NI selected the answer where 5 was the numerator for all three fractions (also the most selected answer after the correct answer among non-ELs, 24%). Most South American countries use commas instead of periods to represent decimal points, so it is possible EL_NI students who are predominantly Hispanic may not recognize 0.50 as equivalent to $\frac{1}{2}$, especially if they are primarily from South America.

A stronger item content argument could be made to support the direction of extreme DIF for the ELA items. The two ELA "A" items from Figure 2 favoring the EL_NI students required the students to read a short poem about a boy and his mother looking at photographs of family from Mexico. The poem contained Spanish language words for family relationships (e.g., tías) and celebrations (e.g., quinceañera). As mentioned above, the EL_NI students are predominantly Hispanic, so these questions may provide a linguistic advantage to EL_NI students despite any potential weakness in ELP. On the other hand, the ELA "B" items which favored the non-ELs, required students to read much longer passages (several pages each) about topics likely less culturally relevant to EL_NI students (e.g., one passage is a story about a girl who brings home a stray dog she calls Winn Dixie and tries to convince her father who is a Baptist preacher to keep the dog). Questions based on these passages also require somewhat abstract inferences, so the reading comprehension requirements are likely more demanding of English proficiency.

3.3 EL_I/NI Class Membership Parameter Estimates

Both reading and writing English proficiency were statistically significant predictors of class membership for both achievement tests. Reading proficiency z-values were 11.38 and 11.11, $p < .05$ for ELA and MATH; and writing proficiency z-values were 2.67 and 4.28, $p < .05$ for ELA and MATH, indicating that the two classes of EL students were clearly differentiated on the basis of their reading and writing. Speaking proficiency was a statistically significant predictor of class membership for ELA, $z = 3.04$, $p < .05$, but not for MATH, $z = 1.56$, $p > .05$. Listening English proficiency was not a significant predictor of EL_I/NI class membership for either ELA, $z = 0.79$, $p > .05$, or MATH, $z = 0.77$, $p > .05$.

To determine at a more practical level the effect that English proficiency skills have on class membership, we

Table 3
Odds Ratios for Class Membership (EL_I vs. EL_NI) as a Function of English Proficiency Subtest Performance

English proficiency subtest	ELA	MATH
Listening	1.1 [a]	1.1 [a]
Speaking	1.6	1.3 [a]
Reading	57.5	14.2
Writing	1.5	1.9

Note. The odds ratios are based on one *SD* differences in the predictor variables.
[a] The logit on which odds ratio is based is not statistically significant, $p > .05$. All other logits are statistically significant.

evaluated odds ratios as a function of EPA subtest *SD*s. We computed each odds ratio as a function of the estimated logistic regression coefficient (logit) and the *SD* for the subtest as follows:

$$Odds\ Ratio = e^{(b_i * SD_i)}$$

where b_i is the logit for subtest i and SD_i is the standard deviation for subtest i within the achievement test group (see Table 3 for EPA subtest SDs). Each odds ratio can be interpreted as the multiplicative change in the odds of a student being in the EL_I vs. EL_NI class for each additive *SD* increase in a given subtest (controlling for the other subtests). As Table 3 demonstrates, reading proficiency had a substantial effect on class membership for the ELA achievement test. For an increase of 1 SD in reading English proficiency the odds of a student being in the EL_I class were multiplied by 58. The effect of writing was much smaller than reading for ELA (odds were multiplied by 1.5 for each SD increase in reading English proficiency). The effect of reading English proficiency on class membership was not as strong (but still quite high) for the MATH achievement test, with the odds ratio = 14.3 (see Table 3). The effect of writing English proficiency for the MATH achievement test was similar to its effect on the ELA achievement test (odds ratio = 1.9).

It should be noted that the English proficiency subtest scores are correlated. Most of the correlations are moderate ($r = .33$ to $.43$) with the exceptions of the correlations between listening and speaking ($r = .86$) and reading and writing ($r = .64$). There is a risk of multicollinearity especially between listening and speaking scores. This could affect the p-values (i.e., listening was not a statistically significant predictor of EL classification for ELA and both listening and speaking were not statistically significant for MATH). However, this is not likely to affect the magnitude of the effects and by far, reading had the biggest effect on classification (at least 7 times the effect of the other proficiency scores).

3.4 Non-EL, EL_I, and EL_NI Characteristics

We conducted post-hoc comparisons of demographics, ELA and MATH achievement, and English language proficiency among the sample students based on their assignment to EL_NI and EL_I classes from the model estimated class probabilities. For both ELA and MATH, EL_I students were more likely to be Asian ($\sim$ 26% vs. 14%) and White ($\sim$ 17% vs 10%) and less likely to be Hispanic ($\sim$ 49% vs. 66%) than EL_NI students, $\chi^2(4) = 151.07$, $p < .05$. EL_I students were also less likely to be receiving free or reduced price lunch ($\sim$ 77% vs 88%, $\chi^2(1) = 82.62$, $p < .05$) or have special education designation ($\sim$ 10% vs. 28%, $\chi^2(1) = 197.00$, $p < .05$). There were no gender differences between the two groups ($\sim$ 49% vs 47%, $\chi^2(1) = 0.85$, $p > .05$).

Table 4 shows group sizes, achievement score means and standard deviations, and effect sizes comparing non-EL to EL_I and EL_I to EL_NI. Effect sizes (*ES*) were calculated by dividing mean differences by pooled standard deviations. There was a group effect on ELA achievement, $F(2,63875) = 4899.80$, $p < .05$, with EL_I students performing moderately below non-EL students, $ES =$ -0.33, $p < .05$, but substantially higher than EL_NI students, $ES = 2.29$, $p < .05$. When controlling for race, free/reduced price lunch status, and special education status, there still was an overall effect of class on ELA achievement, $F(2,63869) = 2706.12$, $p < .05$, but no difference in ELA performance between EL_I students and non-EL students, $p > .05$.

Table 4
Post-Hoc Achievement and English Language Proficiency Comparisons by Class

Measure	non-EL		EL_I		EL_NI		*ES* EL_I vs.	
							non-EL	EL_NI
				ELA				
n	60,025		1,834		2,019			
	M	*SD*	*M*	*SD*	*M*	*SD*		
ELA percent correct	78.1	15.4	73.1	11.4	44.1	13.8	-0.33	2.29
English proficiency SS			392.0	7.2	366.0	22.1		1.55
Reading SS			19.4	2.3	15.3	2.3		1.80
Writing SS			19.5	2.8	16.5	3.1		1.02
Listening RS			4.4	0.7	3.7	0.9		0.81
Speaking RS			16.8	2.7	14.0	3.7		0.87
Level			%		%			
Beginning			0.0		7			
Early intermediate			0.1		12			
Intermediate			4.3		42			
Transitioning			95.7		39			
				MATH				
n	60,159		1,721		2,134			
	M	*SD*	*M*	*SD*	*M*	*SD*		
MATH percent correct	776.7	16.7	77.1	12.2	48.6	15.4	0.03 [a]	2.03
English proficiency SS			392.0	7.4	367.5	22.3		1.41
Reading SS			19.5	2.4	15.5	2.4		1.41
Writing SS			19.7	2.8	16.5	3.0		1.67
Listening RS			4.4	0.7	3.8	0.9		1.08
Speaking RS			16.7	2.8	14.2	3.7		0.76
Level			%		%			
Beginning			0.0		6.5			
Early intermediate			0.1		11.4			
Intermediate			4.8		38.9			
Transitioning			95.1		43.2			

[a] No statistical difference between groups, $p > .05$, based on ANOVA with Tukey post-hoc comparisons. All other group comparisons are statistically significant.

There was also a significant difference in ELA performance between EL_I and EL_NI, even when controlling for these factors, $p < .05$.

Not surprisingly, the three groups differ on MATH achievement, $F(2, 64011) = 2980.90$, $p < .05$. However, EL_I students did not differ from non-EL students in math performance, $ES = 0.03$, $p > .05$, but again performed substantially higher than EL_NI students, $ES = 2.03$, $p < .05$. When controlling for race, free/reduced price lunch status, and special education status, the group effect on MATH achievement remained significant, $F(2, 64005) = 1565.96$, $p < .05$. Controlling for these other factors, EL_I students' MATH scores were significantly higher than both non-EL and EL_NI students, $p < .05$.

In comparing the two EL groups on English language proficiency, the EL_I students had significantly higher ELP than EL_NI students in all comparisons (i.e., composite score, reading, writing, listening, and speaking), even when controlling for race, free/reduced price lunch status, and special education status (see Table 4). An important consideration is how students in the two classes compare on the four-group proficiency level (PL) classification based

on the language proficiency assessment. Over 95% of EL_I students were at the highest (transitioning) performance level in English language proficiency. No student whose English proficiency was at the beginning level of ELP was in the EL_I group, and less than 1% of EL_I students were at the early intermediate level of proficiency. While almost all students classified into the latent class of EL_I come from the highest language proficiency level, 43% of EL_NI students also come from that PL category. Thus, having a language proficiency scale score that places one into the transitioning level of ELP is a necessary, but not sufficient condition to be classified into the EL_I latent class. When EL_I and EL_NI students within PL category 4 are compared to one another, the students in the EL_I class tend to have higher reading and writing scores than those PL category 4 students classified into the EL_NI class.

3.5 Consistency of Classification Across Subject Areas

Finally, there was somewhat high consistency in classification of EL students across content domains. The classification agreement across ELA and MATH was 84%. The consistency of classification was higher among students with lower English proficiency levels: Beginning (n = 137; 100%), Early Intermediate (n = 243; 98.8%), Intermediate (n = 911, 88.7%), Transitioning (n = 2,543; 80.7%).

4 Discussion

The purpose of this study was to evaluate the hypothesis concerning the existence of two latent classes of ELs, one for whom achievement tests function the same as for native English speakers and one for whom achievement tests function differently as evidenced by invariance or non-invariance in item parameters. We also hypothesized that English language proficiency predicts class membership. Both hypotheses are supported by the results of the study. We did find evidence for two classes of EL students as described above. For both ELA and MATH assessments, models that included English language proficiency, fit better than models that did not include ELP as a predictor of EL class.

These results are consistent with the few other studies that demonstrate the heterogeneous nature of groups defined by a priori classifications (e.g., gender, LD status; Cho, Lee, & Kingston, 2012; Cohen & Bolt, 2005). In the case of EL, heterogeneity is partly a function of English language proficiency. Francis and Rivera (2007) propose that emphasis should be placed on academic as opposed to conversational language to ensure ELs have the best opportunity for academic success. The results of this paper are consistent with the theory, suggesting that this distinction is important in evaluating whether students have the language skills necessary for valid assessment of achievement. Whereas a measure of reading English language proficiency is a strong predictor of EL_I vs EL_NI classification, measures of oral English language skills are not strong predictors of the EL class. Reading English language proficiency skills are assessed with content and structures similar to academic assessments of English Language Arts. Students are required to read text and answer comprehension questions. Oral language skills are assessed in a classroom context, and the scoring rubric is based on proficiency in interpersonal and classroom discussions which are likely more consistent with conversational than academic language. An alternative explanation is that the language tasks on the assessment of reading English language proficiency are more aligned with the target language uses (TLUs, Bachman and Palmer, 1996, as cited in Francis & Rivera, 2007, p. 18) than the assessment of oral English language skills. In this case, the TLUs are language tasks contained in the assessments of academic achievement.

English proficiency is the most obvious source of change and heterogeneity among EL students, but not the only source. Although virtually all students in the EL_I latent class were at the highest proficiency level, a high level of English proficiency did not guarantee EL_I membership. Forty-three percent of EL_NI students were at the highest level of English proficiency. Proficiency level is determined by the composite score; however, the results suggest that some ELP subscales (reading and writing) are better indicators of EL class than others (listening and speaking). If ELP is going to be the primary mechanism for decisions about EL students' readiness for taking state achievement tests, then these results suggest cut-points based on ELP should be based on assessments that define ELP consistent with the academic requirements.

In terms of other sources of heterogeneity influencing classification, it should be noted that the EL_I students were more likely to be Asian, less likely to be receiving free or reduced/price lunch, and less likely to have special education status than EL_NI students. However, controlling for these factors did not erase performance differences (favoring EL_I) in ELA or MATH. On the other hand, although non-EL students were much less likely to be non-White or low SES, when controlling for these factors, EL_I students performed similar to non-ELs on the ELA test

and higher than non-ELs on the MATH test.

Exploratory evaluations of the content of DIF items suggested that other cultural factors (e.g., country of origin) may be sources of DIF that are obscured by treating EL students as a homogenous group (i.e., in traditional DIF analyses). Evaluating these other potential sources of DIF is beyond the scope of this study but may be a fruitful path for future research. Research that uses mixture IRT methods to evaluate DIF among EL students may also be used to identify malleable factors that may be addressed through interventions to improve academic outcomes among EL students (e.g., identifying how prior knowledge acquired in non-English learning contexts, such as how decimals are represented, inhibits performance in English language learning contexts).

Implications for Language Proficiency Assessment and Identification. The results of this study may be used to inform several issues in language proficiency assessment and evaluating academic achievement among ELs; including the validity of language proficiency assessments for their intended purpose, the validity of methods for identifying students who need instructional support due to their L2 proficiency and disentangling language proficiency from verbal ability and academic content knowledge.

Validity of Language Proficiency Assessments. This study demonstrated that reading and writing assessments are better indicators than oral language assessments of whether an EL student's performance on an achievement test is a function of his or her content knowledge and ability (comparable to a native English speaker) or a function of L2 language proficiency as well as achievement. It should be noted that our results are based on written (vs oral) achievement tests, and that different findings might be expected if the achievement tests included tasks that involved significant listening or speaking components to them. When evaluating the validity of language proficiency assessments, both the content of the assessment (e.g., reading, writing) as well as the intended purpose (e.g., determining readiness for demonstrating academic content knowledge on state assessments) should be included in the evaluation. Both of these aspects of language proficiency assessment have been discussed in some detail among researchers and policy makers (e.g., Wolf et al., 2014). For example, although the domains of reading, writing, listening and speaking are typically evaluated with ELP assessments, it has been suggested that there should be a move toward defining ELP in terms of "integrated language skills" such as collaborative, interpretive, and productive in order to be consistent with the higher language demands of standards such as the Common Core (Wolf et al., 2014). Another approach is to assess "prerequisite ELP" as language that is common to all content domains versus "content and disciplinary ELP" which is specific to content domains. The validity of these approaches may be evaluated using the methods demonstrated in this study by operationalizing both the assessment of ELP and the criterion for academic competence then using a person-centered approach (e.g., latent class analysis) to classify and characterize students as a function of their performance on both the ELP and academic achievement assessments.

Validity of Methods for Identification. The ELP assessment content is one element that must be considered in evaluating its validity. A second element is the level of performance that indicates "competency" (e.g., cut-scores). The levels of language proficiency defined by the ELP assessment evaluated in this study (Beginning, Early Intermediate, Intermediate, and Transitioning) were established through a traditional standards-setting process, and served as a strong, albeit, imperfect indicator of class membership. Almost all students classified into the invariant class scored in the highest proficiency category. At the same time a substantial proportion of EL students at the highest level of ELP were classified into the non-invariant class, indicating that for many students in the highest category of proficiency, the content area achievement test items function differently than for non-EL students. For these students, performance on the achievement tests was partially a function of L2 language proficiency as well as content area knowledge, and possibly other factors not captured by ELP.

It should also be noted that, although classification of EL students was relatively consistent across content areas, the degree of consistency varied by level of ELP. Whereas as classification agreement was perfect or nearly perfect at the two lowest levels, it dropped to 80% at the highest level of ELP. However, we believe this inconsistency lends credibility to the approach as a method for isolating a threshold for English proficiency. Certainly, we would predict more consistency across domains in the classification of students at low levels of proficiency and maximum instability around the cut-point, where errors of measurement would lead to misclassification, but also where differences in competencies across Reading Language Arts and Math could be sufficient to result in

different classifications across domains. We would expect these to be greatest for students whose language proficiency is very near the cut-score.

The latent class methods demonstrated in this study could be used to augment standard setting procedures in developing proficiency cut-scores, and possibly proficiency categories by linking ELP proficiency to content area achievement as reflected by test functioning without the restrictive assumption of presuming that English language proficiency implies proficiency in content area achievement. Utilization of these methods may improve the validity of EL classifications within the context of the ELP and achievement evaluation methods used within states and could be evaluated through regression discontinuity designs as has been done with other reclassification rules for English learners (Cimpian, Thompson, & Makowski, 2017).

Language Proficiency Versus Verbal Ability and Achievement. The achievement test item analysis described in this study was used to characterize items that exhibited DIF for EL-NI students. Applying this process in a more systematic way may help to identify elements of test items that assess language proficiency as opposed to domain knowledge. Evidence produced from this type of item level analysis may also help with definitions of ELP such as prerequisite and disciplinary ELP as described above.

4.1 Practical Implications

The results of this study indicate that teachers and administrators should be cautious in interpreting results of state achievement tests for EL students, but especially for students at lower levels of English proficiency, as well as for many ELs at the highest levels of English proficiency prior to reclassification, especially those with lower scores on the reading and writing portions of the language proficiency test. For example, for the state from which the students in this study were sampled, test results among students below transitioning levels of English language proficiency are almost certainly being influenced by lack of language proficiency independent of actual verbal ability or domain knowledge. Even among students at the transitioning level of English proficiency, a substantial percentage of these students' test performance appeared to be influenced by language proficiency in an undesired way. To better evaluate whether a student's achievement test scores are reflections of domain knowledge (vs. language proficiency), English reading proficiency may be the best source of evidence (vs. oral language skills or composite scores). Finally, among students who demonstrate relatively high English proficiency, especially reading English proficiency, lower achievement test performance relative to non-EL peers may not be a function of their EL status, but demographic and other student factors that affect learning (e.g., SES, special education status, opportunity to learn, and access to proper instruction) that are also likely to affect non-EL student performance. These factors should be considered in determining methods for remediation among these students.

It should be noted that the focus of this study was on the effect of EL classification on state test performance, specifically, item functioning among the two groups of students. We did not evaluate how measures of English language proficiency influence learning in different contexts (e.g., dual language or full English immersion classrooms). It is possible, for example, that oral English proficiency skills have more influence on an EL students' readiness for instruction in English. The results of this study should not be generalized to evaluation of a student's readiness for instruction without language supports.

4.2 Limitations

As already described, ELs are heterogeneous, differing from one another on many dimensions of interest and relevance to educators. Whereas English language proficiency is the most obvious such source of heterogeneity, there are others. This study was not designed to model or examine these other sources. More complete models would include other factors that may influence EL class membership including cultural background, years in the US, instructional factors, etc. Our intent was to investigate the possibility of using latent classes to distinguish two subgroups of ELs, one for whom content area achievement tests function as they do for non-ELs and to ascertain the extent to which membership in this class was strongly linked to language proficiency. Given the strong support for at least two latent classes and the link to language proficiency, including one for whom content tests function comparably, a more comprehensive examination of factors related to classification seems warranted. A case could be made that at least two classes of non-ELs may exist as well due to racial, ethnic, and social class differences in experience with academic vocabulary that can affect learning and test-taking. This study was not able to examine this possibility. Another contextual factor that could influence the number of classes and class membership

is language-related supports or accommodations provided for the assessments. That information was not available to us, and as was the case with demographic information, not examined in this study.

The EL students excluded from the study were substantially lower in English language proficiency, less likely to be Asian, and less likely to be receiving free or reduced/price lunch or have special education status. It is unclear how the exclusion of these students influenced the results. Based on the demographic representation and ELP of the two classes identified in this study, it is possible that the excluded student could have had relatively higher domain abilities whose content area skills had been obscured by a lack of language proficiency (i.e., students who would have been classified as EL_NI students, but whose performance was higher than average for that class of students). However, further research is needed to more fully explore this possibility.

It cannot be determined with any degree of certainty whether the lack of any effect of oral proficiency in English on latent class membership was due to the nature of the latent classes, the constructs of academic achievement and oral proficiency, or the specific measures of these constructs. Reliability was not reported at the subtest level of the language proficiency test, so it is possible that more reliable measures of oral language skills would have resulted in stronger effects of oral language proficiency on EL classification. However, it seems plausible that the content area achievement tests are not sensitive to individual differences in oral proficiency in the same way that they are sensitive to proficiency with written English given that the content tests are paper and pencil based tests with minimal demands for processing or producing spoken language. To the extent that academic content standards call for processing or producing spoken language and the assessment was designed to measure those standards, then different findings might have been expected.

We excluded polytomous and open response items from the analysis. It is possible these items were not uniformly distributed across topics or standards, so excluding them may have caused the achievement constructs represented in this study to differ from those represented by the full tests. It is also possible that language demands differ between the retained and excluded items.

Finally, the brief item analysis as presented in this study was descriptive and useful primarily as information for designing future studies that more systematically examine questions related to item features that may influence class membership. Ultimately, models that incorporate item features and student characteristics will provide the most information for identifying EL students for whom strong inferences can be made about achievement from state assessments versus EL students for whom information from state assessments should be used with caution.

4.3 Conclusion

Although English language proficiency is the sole defining characteristic of EL status, EL students are a heterogeneous group whose learning and achievement are influenced by many of the same factors that influence non-EL students. Cut-points on tests of English language proficiency are imperfect methods for determining whether or not an EL students' achievement test performance reflects actual domain knowledge. The present study demonstrated that advanced psychometric models could be used to augment and/or validate the standard setting process used to establish levels of English proficiency used to classify EL students. For example, in addition to the invariant and not-invariant classes posed in the current study, one might posit a class of learners for whom the content area achievement test provides no useful information regarding achievement (i.e., a class of learners operating at chance performance). The value of administering a standards-based achievement test to students in this class is questionable. If evidence for such a class of students were found, and objective means for classifying students into this group were devised, then students could be ethically and objectively excluded from accountability testing while still holding schools and administrators accountable for the number of students meeting this designation, and for the length of time that a student held such a designation. Testing practice could be improved for students at all levels of English proficiency if these methods were carefully incorporated into the test development and standards setting validation process, but only through careful consideration of the links between English language proficiency standards, ELP assessment, and the language demands of content area achievement.

Acknowledgments

This work was supported in part by funding from the Eunice Kennedy Shriver National Institute for Child Health and Human Development, P50 HD052117, Texas Center for Learning Disabilities. The attitudes and opinions expressed herein are those of the authors and do not reflect the position of the funding agency, representatives,

agencies, officials, students, or parents of the state from which the data were collected, or the Federal Government of the United States.

References

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

Abedi, J., & Gándara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice*, *25*(4), 36–46.

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

Ardasheva, Y., Tretter, T. R., & Kinny, M. (2012). English language learners and academic achievement: Revisiting the threshold hypothesis. *Language Learning*, *62*, 769–812.

Asparouhov, T., & Muthén, B. (2016). *IRT in Mplus* (Vol. 2; Tech. Rep.). Retrieved from https://www.statmodel.com.

August, D., & Shanahan, T. (2006). *Developing literacy in second-language learners: Report of the National Literacy Panel on language-minority children and youth*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Bayliss, D., & Raymond, P. M. (2004). The link between academic success and L2 proficiency in the context of two professional programs. *Canadian Modern Language Review*, *61*(1), 29–51.

Cai, L., du Toit, S., & Thissen, D. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT Modeling [Computer software]. *Chicago, IL: Scientific Software International*.

Calderón, M., Hertz-Lazarowitz, R., & Slavin, R. (1998). Effects of bilingual cooperative integrated reading and composition on students making the transition from Spanish to English reading. *Elementary School Journal*, *99*(2), 153–165.

Carlo, M. S., August, D., McLaughlin, B., Snow, C. E., Dressler, C., Lippman, D. N., … White, C. E. (2004). Closing the gap: Addressing needs of English-language learners in bilingual and mainstream classrooms. *Reading Research Quarterly*, *39*, 188–215.

Cho, H.-J., Lee, J., & Kingston, N. (2012). Examining the effectiveness of test accommodation using DIF and a mixture IRT model. *Applied Measurement in Education*, *25*, 281–304.

Choi, Y.-J., Alexeev, N., & Cohen, A. S. (2015). Differential item functioning analysis using a mixture 3-parameter logistic model with a covariate on the TIMSS 2007 mathematics test. *International Journal of Testing*, *15*(3), 239–253.

Cimpian, J. R., Thompson, K. D., & Makowski, M. (2017). Evaluating English learner reclassification policy effects across districts. *American Educational Research Journal, Centennial Issue*, *54*(S1), 255S–278S.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*, 133–148.

Dooey, P., & Oliver, R. (2002). An investigation into the predictive validity of the IELTS test as an indicator of future academic success. *Prospect*, *17*, 36–54.

Francis, D. J., & Rivera, M. O. (2007). Principles underlying English language proficiency tests and academic accountability for ELLs. In J. Abedi (Ed.), *English language proficiency assessment in the nation: Current status and future practice* (pp. 13–32). Davis, CA: University of California, Davis, School of Education.

Graham, J. G. (1987). English language proficiency and the prediction of academic success. *TESOL Quarterly*, *21*(3), 505–521.

Gue, L. R., & Holdaway, E. A. (1973). English proficiency tests as predictors of success in graduate studies in education. *Language Learning*, *23*, 89–103.

Hussar, B., Zhang, J., Hein, S., Wang, K., Roberts, A., Cui, J., … Dilig, R. (2020). *The condition of education* (NCES 2020-144). U.S. Department of Education. Washington, DC: Retrieved [July 19, 2020] from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2020144.

Hwang, K.-Y., & Dizney, H. F. (1970). Predictive validity of the test of English as a foreign language for Chinese graduate students at an American university. *Educational & Psychological Measurement*, *30*, 475–477.

Kena, G., Aud, S., Johnson, F., Wang, X., Zhang, J., Rathbun, A., … Kristapovich, P. (2014). *The condition of education* (NCES 2014-083).

Washington, DC: Retrieved from http://nces.ed.gov/pubsearch.

Kerstjens, M., & Nery, C. (2000). Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance. *English Language Testing System Reports*, *3*, 85–108.

Kim, Y.-H., & Jang, E. E. (2009). Differential functioning of reading subskills on the OSSLT for L1 and ELL students: A multidimensionality model-based DBF/DIF approach. *Language Learning*, *59*, 825–865.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, *33*, 353–373.

Light, R. L., Xu, M., & Mossop, J. (1987). English proficiency and academic performance of international students. *TESOL Quarterly*, *21*(2), 251–261.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*, 21–39.

Mahoney, K. (2008). Linguistic influences on differential item functioning for second language learners on the National Assessment of Educational Progress. *International Journal of Testing*, *8*, 14–33.

Marsh, H. W., Lüdtke, O., Trautwein, U., & Morin, A. J. (2009). Classical latent profile analysis of academic self-concept dimensions: Synergy of person-and variable-centered approaches to theoretical models of self-concept. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(2), 191–225.

Martiniello, M. (2008). Language and the performance of English language learners in math word problems. *Harvard Educational Review*, *78*, 333–368.

Mislevy, R. J., Levy, R., Kroopnick, M., & Rutstein, D. (2008). Evidentiary foundations of mixture item response theory models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Latent Variable Mixture Models* (pp. 149–176). Charlotte, NC: Information Age Publishing, Inc.

Mulligan, A. C. (1966). Evaluating foreign credentials. *College and University*, *41*, 307–313.

Muthén, B. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. F. Schumacher (Eds.), *New developments and techniques in structural equation modeling* (pp. 1–33). Mahwah, NJ: Lawrence Erlbaum Associates.

Muthén, B. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*(1), 81–117.

Muthén, B. (2008). Latent variable hybrids: Overview of old and new models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Latent Variable Mixture Models* (pp. 1–24). Charlotte, NC: Information Age Publishing, Inc.

Muthén, B., Khoo, S.-T., Francis, D., & Kim Boscardin, C. (2000). Analysis of reading skills development from kindergarten through first grade: An application of growth mixture modeling to sequential processes. *Multilevel modeling: Methodological advances, issues, and applications*, 71–89.

Muthén, L., & Muthén, B. (1998–2012). Mplus [computer software]. Los Angeles, CA: Muthen & Muthen. Retrieved from www.StatModel.com.

Muthén, L., & Muthén, B. (1998–2017). Mplus user's guide (version 8). Los Angeles, CA: Muthen & Muthen.

National Research Council. (2000). *Testing English-language learners in U.S. schools: Report and workshop summary.* Washington, DC: National Acadamy Press Retrieved from http://www.nap.edu/catalog/9998.html.

National Research Council. (2002). *Reporting test results for students with disabilities and English-language learners.* Washington, DC: National Acadamy Press Retrieved from http://www.nap.edu/catalog/10410.html.

Nylund, K., Asparouhov, T., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569.

Ockey, G. J. (2007). Investigating the validity of math word problems for English language learners with DIF. *Language Assessment Quarterly*, *4*, 149–164.

Oliver, R., Vanderford, S., & Grote, E. (2012). Evidence of English language proficiency and academic achievement of non-English-speaking background students. *Higher Education Research & Development*, *31*(4), 541–555. doi: 10.1080/07294360.2011.653958.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383.

Pomplun, M., & Omar, M. H. (2001). The factorial

invariance of a test of reading comprehension across groups of limited English proficient students. *Applied Measurement in Education*, *14*, 261–283.

Samuelsen, K. M. (2008). Examining differential item functioning from a latent mixture perspective. In G. R. Hancock & K. M. Samuelsen (Eds.), *Latent Variable Mixture Models* (pp. 177–198). Charlotte, NC: Information Age Publishing, Inc.

Saunders, W. M., & Marcelletti, D. J. (2013). The gap that can't go away: The catch-22 of reclassification in monitoring the progress of english learners. *Educational Evaluation and Policy Analysis*, *35*(2), 139–156. doi: 10.3102/0162373712461849

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society (Series B)*, 13:238–241.

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

Snetzler, S., & Qualls, A. L. (2000). Examination of differential item functioning on a standardized achievement battery with limited English proficient students. *Educational & Psychological Measurement*, *60*, 564–577.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models.

Turkan, S., & Liu, O. L. (2012). Differential performance by English language learners on an inquiry-based science assessment. *International Journal of Science Education*, *34*, 2343–2369.

Umansky, I. M., Thompson, K. D., & Díaz, G. (2017). Using an Ever-EL framework to examine special education disproportionality among English learner students. *Exceptional Children*, *84*(1), 76–96.

U.S. Department of Education. (2011). *ED data express: Data about elementary & secondary schools in the U.S.* Retrieved Febreary 11, 2013, from U.S. Department of Education http://www.eddataexpress .ed.gov/state-tables-main.cfm.

Vellutino, F. (2003). Individual differences as sources of variability in reading comprehension in elementary children. In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 55–81). New York: The Guilford Press.

Wolf, M. K., Guzman-Orth, D., & Hauck, M. C. (2014). Next-generation summative English language proficiency assessments for English learners: Priorities for policy and research. Retrieved from http://www.k12center.org/rsc/pdf/wolf_elp.pdf.

Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment*, *14*, 139–159.

Young, J. W., Cho, Y., Ling, G., Cline, F., Steinberg, J., & Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. *Educational Assessment*, *13*, 170–192.

Appendix: Mplus Code for Model 2 (MATH)

```
TITLE: Latent Class IRT
DATA: FILE IS g406math170211.csv;

VARIABLE:
  NAMES ARE lis spe rea wri ss everlep m1-m39 id;
  USEVARIABLES ARE lis spe rea wri everlep
  m1-m9 m11 m12 m14-m16 m18-m26 m28-m30
m32-m39;
  CLASSES = c(3);
  CATEGORICAL ARE m1-m9 m11 m12 m14-m16
m18-m26 m28-m30 m32-m39;
  IDVARIABLE = id;

ANALYSIS:
  TYPE = MIXTURE;
  ESTIMATOR = MLR;
  ALGORITHM = INTEGRATION;
  STARTS = 500 125;
  PROCESSORS = 3 (STARTS);

MODEL:
  %OVERALL%
  c#1 on everlep@-60 lis@0
        spe@0 rea@0 wri@0;
  [c#1@30];
  c#2 on lis spe rea wri;
```

Multinomial logistic regression (Mplus default) is used to regress class onto English language proficiency subtest scores (lis spe rea wri).

```
      mf by m1-m9*;
      mf by m11-m12*;
      mf by m14-m16*;
      mf by m18-m26*;
      mf by m28-m30*;
      mf by m32-m39*;
      mf@1;

%c#1%
      mf by m1-m9* (nlep1-nlep9);
      mf by m11-m12* (nlep11-nlep12);
      mf by m14-m16* (nlep14-nlep16);
      mf by m18-m26* (nlep18-nlep26);
      mf by m28-m30* (nlep28-nlep30);
      mf by m32-m39* (nlep32-nlep39);
      [m1$1-m9$1] (nlpt1-nlpt9);
      [m11$1-m12$1] (nlpt11-nlpt12);
      [m14$1-m16$1] (nlpt14-nlpt16);
      [m18$1-m26$1] (nlpt18-nlpt26);
      [m28$1-m30$1] (nlpt28-nlpt30);
      [m32$1-m39$1] (nlpt32-nlpt39);

      [mf@0];
      mf@1;

%c#2%
      mf by m1-m9* (nlep1-nlep9);
      mf by m11-m12* (nlep11-nlep12);
      mf by m14-m16* (nlep14-nlep16);
      mf by m18-m26* (nlep18-nlep26);
      mf by m28-m30* (nlep28-nlep30);
      mf by m32-m39* (nlep32-nlep39);
      [m1$1-m9$1] (nlpt1-nlpt9);
      [m11$1-m12$1] (nlpt11-nlpt12);
      [m14$1-m16$1] (nlpt14-nlpt16);
      [m18$1-m26$1] (nlpt18-nlpt26);
      [m28$1-m30$1] (nlpt28-nlpt30);
      [m32$1-m39$1] (nlpt32-nlpt39);
      [mf];
      mf;

%c#3%
      mf by m3 (nlep3);
      mf by m1-m2;
      mf by m4-m9;
      mf by m11-m12;
      mf by m14-m16;
      mf by m18-m26;
      mf by m28-m30;
      mf by m32-m39;

      [m3$1] (nlpt3);
      [m1$1-m2$1];
      [m4$1-m9$1];
      [m11$1-m12$1];
      [m14$1-m16$1];
      [m18$1-m26$1];
      [m28$1-m30$1];
      [m32$1-m39$1];
      [mf];

      mf;

OUTPUT: PATTERNS TECH1;
```

```
SAVEDATA: FILE IS g406m2plSubAnchM3d170527.sav;
SAVE = FSCORES CPROBABILITIES ;
```

Parameterization for regressing class onto English language proficiency scores. The Mplus code parameterizes the logit link function in the multinomial logistic regression to restrict the non-EL students to one class of students, while assigning EL students to one of two other classes based on their performance on MEPA and the achievement test items:

Model 1: $P(C=r|EL)=\frac{e^{(b_{r0}+b_{r1}EL)}}{e^{(b_{10}+b_{11}EL)}+e^{(b_{20}+b_{21}EL)}+e^{(b_{30}+b_{31}EL)}}$

r	b_{r0}	b_{r1}
1 $(\tau,\lambda_{\text{non-EL}})$	30	-60
2 $(\tau,\lambda_{\text{non-EL}})$	$b_2 0$	0
3 (τ,λ)	0	0

Model 2: $P(C=r|EL,LIS,SPE,REA,WRI)=\frac{e^{(b_{r0}+b_{r1}EL+b_{r2}LIS+b_{r3}SPE+b_{r4}REA+b_{r5}WRI)}}{e^{(b_{10}+b_{11}EL+b_{12}LIS+b_{13}SPE+b_{14}REA+b_{15}WRI)}+e^{(b_{20}+b_{21}EL+b_{22}LIS+b_{23}SPE+b_{24}REA+b_{25}WRI)}+e^{(b_{30}+b_{31}EL+b_{32}LIS+b_{33}SPE+b_{34}REA+b_{35}WRI)}}$

r	b_{r0}	b_{r1}	b_{r2}	b_{r3}	b_{r4}	b_{r5}
1 $(\tau,\lambda_{\text{non-EL}})$	30	-60	0	0	0	0
2 $(\tau,\lambda_{\text{non-EL}})$	b_{20}	0	b_{22}	b_{23}	b_{24}	b_{25}
3 (τ,λ)	0	0	0	0	0	0

where $P(C=r|EL,LIS,SPE,REA,WRI)$ is the probability, P, of belonging to class, r, given EL status (0 = non-EL, 1 = EL) and performance on EPA subscale scores (LIS, SPE, REA, WRI). b_{r0} are the intercepts for predicting class membership, and b_{r1-5} are the regressions coefficients for EL status and EPA subscale scores: LIS, SPE, REA, WRI. τ, λ are the threshold's and loadings relating the achievement factor (ELA, MATH) to performance on each achievement test item. Classes 1 and 2 share the same subscript, non-EL, for these parameters because the item parameters are invariant between these two classes. Class 3 item parameters are allowed to differ from the other two classes. Class 3 is the reference class in estimating regression coefficients for class membership. The values 30 and -60 for b_{r0} and b_{r1} are arbitrary but of great enough magnitude to force the probability of Class 1 membership to 1 if the student belongs to the non-EL group (i.e., EL = 0) and 0 if the student belongs to the EL group (i.e, EL = 1). The regression coefficients for Class 2 can be interpreted as follows: $e^{b_{2p}}$ is the multiplicative increase in odds (odds ratio) of belonging to Class 2 (vs. Class 3) for each unit increase in p where p is the EPA scale or subscale score.