Investigating the fit and functioning of a high school Algebra assessment for English Language Learners using the dichotomous Rasch model

Shannon O. Sampson and Kelly D. Bradley

University of Kentucky

Investigating the fit and functioning of a high school Algebra assessment for English Language Learners using the dichotomous Rasch model

Important decisions are made based on all sorts of assessments, but there are questions regarding whether they really assess what is intended for all students. An area of interest within the assessment field is connected to the population of English Language Learning (ELL) students. Given the growth of this group of students nationally, concerns are rising with regard to the validity of the results of content assessments with ELL students, as linguistic issues may obscure students' access to the construct being measured. When assessment tools are used for decision-making purposes, these decisions may be made on the basis of inaccurate information. Determining how well assessments function for all students assessed is a critical piece of validating instruments the used for these purposes.

The Problem

The problem investigated in this study was if a multiple-choice assessment constructed to measure algebra ability, with varying levels of linguistic complexity among the items, fit the requirements of the Rasch model (Rasch, 1960) and function similarly for ELL and English-proficient students. Items identified as poorly fitting the expectations of the Rasch model were examined to determine if linguistic features were contributing to the difference. While the intention may be for item difficulties and student abilities to dominate person responses (Wright & Stone, 1979, p.11), for students learning English, an assessment designed to measure algebra ability may be confounded by item linguistic complexity and student lack of English proficiency. This is recognized in *Standards for Educational and Psychological Testing* (AERA, APA, &

students. To increase the accountability of at-risk groups of students, NCLB further requires that schools, school districts, and states disaggregate, or separate out, the assessment results for several subgroups of students including ELL students.

In the past, ELL students were often excluded from large-scale assessments and accountability systems because it was posited that even though many ELL students have the content knowledge and/or the cognitive ability to perform successfully on assessment tasks, the assessment experience would be extremely frustrating and the picture of ELL students' content knowledge would likely not be valid (August and Hakuta, 1997; LaCelle-Peterson & Rivera, 1994; O'Malley & Valdez Pierce, 1994; Shepard, Taylor, & Betebenner, 1998). Still, many argue that when any group is systematically excluded from assessment system, a biased picture of education is presented, particularly if the group that is excluded tends to be lower-performing students (McGrew, Thurlow & Spiegel, 1993; Rivera, Stansfield, Scialdone & Sharkey, 2000). Full participation in an assessment system is a critical piece of monitoring if all students are to benefit from reforms that are implemented. As more ELL students are participating in assessments, it is the task of the assessment developers to design assessments that paint a true picture of what all students are learning.

Some research suggests assessments that limit language complexity may provide more accurate information about ELL algebra ability. A study of language accommodation on math assessments by Kiplinger, Haug and Abedi (2000) revealed that the performance of students on a mathematics assessment with high proportions of word problems was directly related to their proficiency in reading in English. Better performance of ELL students and other students who read less well resulted from the simplification of linguistic structures and the addition of a glossary for non-mathematics vocabulary. The study concluded linguistic simplification or

assessment, which is offset by a lack of research on this dimension in the assessment literature. They write that the language of assessments should be "as 'transparent' as possible, allowing the mathematical demand to be made clear and the subject competence to become evident for the great majority of the pupils taking the tests" (p. 125). More research in this area will begin to reveal whether simplified language in assessments does assess the intended construct of mathematics for most students.

Schifter (1997) observes that learning algebra is similar to learning a new language. With algebra, to students who are already familiar with properties and relationships among operations, the challenge is to learn the conventional symbol system. For a student who has not developed operation sense, learning the symbol system is a daunting task. In a sense, algebra students are also students of a mathematical language; the cognitive foundation they bring to the classroom would likely similarly affect their understanding of algebra.

Determining Linguistic Complexity of Algebra Items

Linguistic complexity is related to, but is more extensive than, readability, which has been researched for many years in education. Readability is affected by "students' previous experiences, achievement, and interests, and by text features such as word and sentence difficulty, organization of materials, and format" (Rakow & Gee, 1987, p. 28). Readability is an especially complex component for ELL students. Whereas many readability formulas are based on word counts, syllable counts, or sentence length, for an ELL student these features may not be good indicators of likelihood of comprehension. Various researchers have identified features beyond the typical readability formulas that may pose difficulties for ELL students (Abedi & Lord, 2001; Abedi, Lord & Plummer, 1997; Anstrom, 1999; Brown, 1999; Corasaniti Dale &

Method

Response Frame

The subjects for the study were drawn from the students enrolled in a large public high school located in a southern state. This school was selected for a number of reasons. First, it was reasonable to believe, based upon demographics and state assessment results, that the students within the school represented a wide range of mathematics ability, which would provide variance for assessment results for this study. Additionally, 5%, or approximately 90 students of a total of approximately 1830 total students in the school, were classified as ELL students, resulting in the school hosting the largest ELL population of the five high schools within the district.

All math teachers at the school were approached during the summer of 2004 regarding their willingness to participate in a concurrent study on their perceptions of math teacher quality (Bradley, 2004). That study included a student assessment, which was also used for the data collection for this research. Teachers were first contacted requesting their participation in the concurrent study via an e-mail sent prior to the first week of the school year. This was followed by an informational visit to their department meeting in mid-September 2004. Six of the thirteen (46%) teachers in the mathematics department agreed to administer the assessment. In addition to the math teachers, one English as a Second Language (ESL) teacher also administered the test. Math classes to which the assessment was administered were Algebra I part 1, Algebra I part 2, Algebra I, Algebra II, Algebra I-repeat, and Geometry. Four hundred forty-four students participated in the assessment. Of these, 51 students were identified as ELL students; they represented 56.7% of the total number of ELL students enrolled at the school at the time of the study. ELL students who did not participate were either enrolled in a mathematics class of one of the non-participating teachers or were absent on the day of the assessment administration. All

"exemplars of questions that probe students' knowledge of [a] specific content area," (NCES, 2005, para. 2), it was assumed these were quality items and thus, 17 released NAEP items were included on the assessment in the study. All items were connected to 8th grade algebraic ideas and state core content. Teachers were instructed to encourage students to give their best effort on the assessment, in order to increase the likelihood of the assessment results presenting an accurate reflection of student ability.

Algebra was selected as the mathematics strand for assessment for multiple reasons. It is generally considered to encompass a way of thinking essential to mathematics and should be available to all students (Moses, 1994). The NCTM Curriculum Standards (1989) stressed that understanding of Algebra is a necessary foundation for further work in mathematics. Pelavin and Kane (1990) also suggest that minority students who have successfully completed algebra and geometry are as likely to succeed at the college level as non-minority students. This study limited questions to the 8th grade state core content, assuming each student should be expected to know and should have had the opportunity to learn these concepts prior to entering high school.

Items selected for the assessment ranged from linguistically simple to complex, and from easy to difficult math content. Items were selected such that each cell in a linguistic-complexity (see Table 2 by content-difficulty matrix contained approximately 3 items. This was to ensure the assessment included a range of mathematical difficulty as well as linguistic complexity.

Table 2: Item difficulty by linguistic complexity matrix

		Linguistic Complexity		
		low (0 to 1 feature)	moderate (2 to 4 features)	high (5 or more features)
Item Difficulty	difficult	item 1 (NAEP) item 19 (NAEP) item 29 (MA)	item 18 (NAEP) item 30 (MA) item 3 (AZ)	item 2 (NAEP) item 9 (NAEP) item 14 (NAEP) item 20 (TX)

Data Analysis

The analysis began with the application of the Rasch model to the full data set including English-proficient and ELL students. Winsteps software, version 3.55 (Linacre, 2005) was used for the analysis. The data were obtained from a multiple choice assessment, so the dichotomous Rasch model was utilized, which is represented with $Log_e\left(\frac{P_{ni1}}{P_{ni0}}\right) = B_n - D_i$, where P_{ni1} and P_{ni0} are the probability that person *n* encountering item *i* is observed in category 1 or 0, B_n is the ability measure of person *n*, and D_i is the difficulty measure of item *i*.

Rasch models provide a direct estimate of the modeled error variance for each estimate of a person's ability and an item's difficulty, providing a quantification of the precision of every person measure and item difficulty which can be "used to describe the range within which each item's 'true' difficulty or person's 'true' ability falls" (Smith, 2004, p. 96). Winsteps reports both person reliability and item reliability. Person reliability is equivalent to traditional test reliability, computed as the true variance divided by the observed variance, and dependent upon the number of items in the assessment. Observed variance is the standard deviation of the person measures, squared, whereas the true variance is calculated by taking the standard errors of the person measures, squaring each, and summing the squares. This sum is divided by the count of entries, and is subtracted from the observed variance (Linacre, 2004, p.275-276). Person reliability can be described as the reproducibility of the person ordering. Winsteps reports both the model and the real reliability. Model reliability is an upper bound reliability value and the real reliability is the lower bound. The true reliability falls somewhere between the two scores. Item reliability is computed as the true item variance divided by the observed item variance. If it is low, an increased sample size with greater variance can improve it (Linacre, 2004).

content, and linguistic commonalities among items were sought. Use of the Opportunity to Learn Survey allowed for the investigation of whether or not students had been exposed to the content of the items.

Results

Nearly 450 students participated in the assessment; of those, approximately 12% were ELL students, and approximately 89% were English-proficient students. Over half of the students self-identified as White/Caucasian, followed by almost 19% African or African-American students, about 3.5% Hispanic and just over 5.5% Asian or Pacific Islander students. Of the ELL students, a majority of the students were Hispanic; the next largest subpopulation was Asian or Pacific Islanders, which was half the size of the largest group. Most ELL students in the study were enrolled in Algebra I part I. Just under half of those students were enrolled in a Level 1 ESL class, so their English proficiency was very limited. The other half of those students taking Algebra I Part I were enrolled in a Level 2 or higher ESL class.

Item-Person Map

The analysis began with a look at the item-person map (see Figure 4.3 below) to examine the spread of the students and the location of the ELL students in relation to their English-proficient peers. Negative logit scores indicate less ability in reference to students and less difficulty in reference to items. Conversely, positive logit values indicate more ability in reference to students and more difficulty in reference to items. Students located directly across from an item have a 50% probability of answering that item correctly. A student located one logit below an item has a 25% probability of answering correctly, and a student located one logit above an item has a 75% probability of answering correctly.

Forty ELL students (78%), represented with 1s, fell well below the mean difficulty measure of the items and most fell below the bulk of the items. Furthermore, many of ELL students were located at or below the second-easiest item. Seventeen ELL students would most likely only answer one item, 6, correctly. Based on the model assumptions, three students were less than 50% likely to even answer that item correctly. The most difficult item in the sample , Item 18, appears at the top of the map. The least difficult in the sample, item 6, appears at the bottom of the map. Most items were located near the mean, which is to be expected of items designed to determine student proficiency in algebra; however, the items are distributed along the continuum. While this map displays persons and items as measured along the same unit of value, the fit of the data to the model must be evaluated before considering the information presented to be reliable.

Reliability

Winsteps reports a "real" and "model" person separation reliability; the actual reliability falls somewhere between these two values (Linacre, 2004). Fox and Jones (1998) indicate that person reliability requires ability estimates well targeted by the pool of items designed to measure the construct, as well as a large-enough spread of ability that the measures demonstrate a hierarchy of ability (person separation) on this construct (as cited in Bond & Fox, 2001, p.32). The real person separation reliability was .84, suggesting the instrument was reasonably reliable.

Fit of the Data to the Model

Diagnosis of misfit followed Linacre's (2004) general rules: investigating outfit before infit and high values before low values. Linacre notes high outfit mean-squares may be the result of a few random responses by low performers, so person fit is evaluated before item fit. Person ability and item difficulty estimates are placed along the same metric and expressed in logits.

enrolled in the first-level ESL class. Seventy-three percent of the students were Hispanic or spoke Spanish as their first language, while the other 27% were Asian.

Analysis of Item Fit

Following Linacre's (2004) recommendation to review high values before low values, analysis of item fit began with the items with a high outfit mean-square value. Five items had mean-square fit values greater than 1.3, and were thus considered unproductive for measurement (Linacre, 2004). The expectation was that ELL students would answer certain linguistically complex items incorrectly even though they were likely to answer correctly based on their ability measure and the item difficulty measures. As it turned out, the ELL subgroup answered very few questions correctly; the majority even incorrectly answered items which were hypothetically easy in content with low linguistic complexity. Because of this finding, the investigation of misfit turned to why many ELL students consistently answered certain difficult items correctly.

As revealed in the item hierarchy, item 3 was among the most difficult items on the assessment. Thirty-one percent of the ELL students, as compared to 20% of the English-proficient students. This item had only two linguistically complex features—length and a compound sentence—making it one of the less complex items. Of the ELL students, 13 answered this question correctly, even though the Rasch expectation was an incorrect response. Ten of these students were enrolled in an Algebra I part 1 class and the Opportunity to Learn Survey indicated a teacher expectation that they had covered this concept during the present year, a strong case for recentness.

Almost no language was included in item 29; instead, getting it correct was related to knowing the symbol for absolute value. Only 22% of students correctly answered the question even though teachers reported that all classes except Algebra I part 1 had been introduced to this

expression into a mathematical expression. Finally, items 3, 4, and 24 are items which contain material the teachers reportedly taught to their students this year.

Investigation of Differential Item Functioning

An additional assumption of the Rasch model is the property of person-free measurement (also described as invariance) which indicates that when an assessment is administered to any group of students, the item difficulty measures remain the same, within measurement error. If item difficulty measures are significantly different across subgroups, that item is said to exhibit Differential Item Functioning (DIF). Prior to reviewing item invariance across subgroups, the two items with the highest mean-square values were revisited to determine if they distorted the person measures more than they contributed to measurement accuracy and precision by cross-plotting the person measures with and without the suspect items (in this case 3 and 29). The cross-plot displayed a well-defined diagonal, indicating a strong relationship between the measures. The items were not removed since the goal was to investigate idiosyncrasies.

Table 3 displays the difficulty measures for each item when separately calibrated for the English-proficient and ELL student subgroups. The items are sorted from least to most difficult for the English-proficient students according to the initial calibration. Items noted with a single asterisk were differentially more difficult for ELL students, and items noted with two asterisks were differentially easier for the ELL students.

Table 3 Relative difficulty measures for ELL and English-proficient subgroups

Item		ELL students		English-proficient students					
	Linguistic Complexity	Item difficulty measure	SE	Item difficulty measure	SE	DIF contrast	JOINT S.E.	t	Item d.f.
6	Mid	-2.13	0.34	-2.26	0.16	0.13	0.37	0.36	434

At the p=.05 level, items with a t value of less than -2 are said to favor the ELL students, and items with a t value of greater than 2 are said to favor the English-proficient students. Four items, 3, 4, 27, and 29, favored the ELL students, and five items 5, 11, 12, 17, and 25, favored the English-proficient students. All items that favored the ELL students had been flagged as misfitting, since many of the ELL students answered these items unexpectedly correctly. None of the items that favored the English-proficient students, those differently difficult for the ELL subgroup, had been flagged as misfitting. Hence, these items were reviewed to determine if the DIF could be attributed to linguistic complexity.

Item 5 contained only one linguistic feature, placing it among the easiest items in the linguistic hierarchy. Thirty-seven percent of students selecting the option indicated they did not know the mathematical meaning of ">." Two of the Algebra I part 1 teachers reported students had not yet been introduced to this material, and one of the Algebra I part 1 teachers reported she expected students to have learned the material in a previous class.

Item 11 contained many linguistically complex features including complex noun phrases, conditional structure, a word with more than one meaning and a comparative structure. The item could have been written language-free, but as presented it included multiple linguistic features that potentially created barriers to understanding for the ELL students. Most importantly, if a student did not know that "less than" is the English equivalent for the symbol "<," he or she would likely answer the question incorrectly even if he or she understood the concept of inequalities. While the majority of the English-proficient students (77%) selected the correct option, only 43% of the ELL students did so. The teachers of the ELL students reported they expected the content to have been taught a previous year and did not cover the content in their classes. Considering that, the DIF could be attributed to the linguistic features of the item and

Finally, item 25 was considered long and moderately complex. It contained passive voice, a compound sentence, complex noun phrases and a conditional structure. For the majority of ELL students, teachers reported introducing the content during the current school year. Only 27% of ELL students chose the correct option, compared to 60% of English-proficient students. The answer most selected by ELL students indicated students simply added the listed numbers.

DISCUSSION

One would expect most students to have a high probability of answering correctly for most of the items, since the items selected for the assessment were linked to the standards for eighth grade students in the state in which the study occurred. The ability of the ELL students largely fell below the mean item difficulty; however, indicating the students had a low probability of answering correctly on most items. While disappointing, the finding is reflective of the performance of Latin American students on statewide math assessments (Holloway, 2004).

Although the overall performance of the ELL students was generally poor, not all the results were discouraging. Proportionately more ELL students than English-proficient students fit the model poorly. The review of ELL student fit revealed many students answered correctly on certain difficult items, including 3, 27 and 29.

In four of the five misfitting items, students were required to translate a verbal expression into a mathematical expression. This included item 3, in which 31% of the ELL students answered correctly, compared to 20% of English-proficient students with the correct selection. One might conclude the ELL students just made lucky guesses, since the correct answer was “very unexpected” for most of them. Percentages on the other distracters were more comparable for the two groups, however, suggesting this may have been more than guessing. Item 3 was one that may have been susceptible to a careless mistake by a student who read through it quickly

Opportunity to Learn Survey revealed that these items contained material that many of the teachers with concentrated ELL students in class did not expect their students to know. On the other hand, differentially easier items for the ELL students (3, 24, and 4) contained material reportedly presented this year to the classes with large numbers of ELL students. For these items, student opportunity to learn the content as well as how recently it was taught may have made a difference. While this observation was not consistent for all items, in light of the possible relationship between item difficulty and student opportunity to learn, it is especially interesting to note that teachers reported varying expectations regarding if and when the material had been taught to the students. For item 12, for example, in the Algebra I part 1 classes, one teacher had taught the material this year, one expected students to have been taught the material a previous year and one reported the students had probably never been taught the material. Depending on a student's series of teacher, he or she could have been presented the material two years in a row or could have missed the opportunity to learn the material altogether. Because opportunity to learn may have played a role in student performance on items, these inconsistent expectations were troublesome.

CONCLUSION

This study provides a methodology for evaluating the quality of an assessment, from large-scale at the state or national level to teacher-created assessments at the classroom level, for use across diverse populations that could be used by researchers and classroom teachers alike. While IRT has been utilized for analysis of assessment results, particularly with large-scale assessments, this study utilizes Rasch measurement to identify and attempt to explain the misfit of many ELL students and certain items. As the ELL population continues to grow, schools, districts and states are increasingly faced with the challenge of responsibly measuring student

mainstream mathematics classes, as findings suggest mathematics instruction may be far more important in determining ELL students' success on algebra items than limited linguistic complexity. Creating student opportunity to work with given concepts in the context of the English language may be as important as concerns over accommodations and assessment modifications.

There is much discussion regarding the appropriateness and use of assessments with ELL students, regarding if and when ELL students should participate, and the consequences of making decisions based on these assessments. As noted in Lamprianou and Boyle (2004):

> Reasons that relate to examinee characteristics such as language deficiencies, gender, anxiety, motivation and race have been regularly proposed as potentially causing misfit. …However, not enough research has been done using empirical test data, and too few of the hypotheses mentioned… have been researched systematically. (p. 240)

This study adds to the body of literature regarding the role language may play in the assessment of mathematics content for ELL students. It also incorporates the importance of student opportunity to learn the material, which has not been considered in the previous research regarding the effect of linguistic complexity on ELL student performance. Finally, it expands the use of the Rasch model in analyzing mathematics assessments for subgroups of students divided by level of English-language proficiency.

August, D., & Hakuta, K. (1997). *Improving schooling for language minority students: A research agenda*. Washington, DC: National Academy Press.

Bond, T., & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bradley, K. (2004) Western Kentucky University; PI; KY EPSB. *Pilot Study: Assessing Quality-A Macro versus Micro Approach to High School Mathematics Education*. ($36,452); 7/1/04 – 6/30/05.

Brown, P. J. (1999). *Findings of the 1999 plain language field test*. University of Delaware, Newark, DE: Delaware Education Research and Development Center. Retrieved December 12, 2005, from http://www.doe.state.de.us/aab/Atch%204%20Findings%20of%20the%201999%20Plain%20Language%20Field%20Test.pdf.

Corasaniti Dale, T., & Cuevas, G. J. (1992). Integrating mathematics and language learning. In P.A. Richard-Amato & M.A. Snow (Eds.), *The multicultural classroom: Readings for content-area teachers* (pp. 330-348). White Plains, NY: Longman.

DeAvila, E. A., & Duncan, S. E. (1994). *Language assessment scales: Scoring and interpretation manual, English*. Monterey, CA: CTB Macmillan/McGraw Hill.

De Corte, E., Verschaffel, L., & DeWin, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. *Journal of Educational Psychology*, 77, 460–470.

Holloway, J.H. (2004). Research link: Closing the minority achievement gap in math. *Educational Leadership*, 61(8), 84-86.

Linacre, J. M., & Wright, B. D. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), p. 370.

McGrew, K. S., Thurlow, M. L., & Spiegel, A. N. (1993). An investigation of the exclusion of students with disabilities in national collection programs. *Educational Evaluation and Policy Analysis*, 15, 339-352.

Moses, R. (1994). Remarks on the struggle for citizenship and math/science literacy. *Journal of Mathematical Behavior*, 13(1), 107-111.

Munro, J. (1979). Language abilities and math performance. *Reading Teacher*, 32(8), 900–915.

National Center for Education Statistics (NCES). (n.d.). *NAEP questions: tool help.* Retrieved December 12, 2005, from http://nces.ed.gov/nationsreportcard/itmrls/startadvancedsearch.asp (select Tool Help).

National Council of Teachers of Mathematics (NCTM). (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

No Child Left Behind Act of 2001. 107th Congress of the United States of America. Retrieved December 12, 2005, from http://ed.gov/legislation/ESEA02/107-110.pdf.

Noonan, J. (1990). Readability problems presented by mathematics text. *Early Child Development and Care*, 54, 57–81.

O'Malley, J. M., & Valdez Pierce, L. (1994). State assessment policies, practices, and language minority students. *Educational Assessment*, 2(3), 213-255.

Orr, E. W. (1987). *Twice as less: Black English and the performance of Black students in mathematics and science*. New York: Norton.

Smith, E. (2004). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. In E. Smith & R. Smith (Eds.), *Introduction to Rasch measurement* (pp. 93-122). Maple Grove: JAM Press.

Spanos, G., Rhodes, N. C., Dale, T. C., & Crandall, J. (1988). Linguistic features of mathematical problem solving: Insights and applications. In R. R. Cocking & J.P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 221-240). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wright, B. D., & Stone, M. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.