…AND THE BEAT GOES ON: FURTHER EVIDENCE TO SUPPORT THE NEED FOR ACCOMMODATIONS AND UNIVERSAL DESIGN IN HIGH STAKES TESTING OF ENGLISH LANGUAGE LEARNERS

Andrea Erichsrud

Christopher Johnstone

ABSTRACT

This article describes research that used the think-aloud method to elicit responses from students on released high stakes test items. Four students who were English language proficient and four students whose first language was Spanish completed a mini-test made up of four mathematics items. In the process of thinking aloud, the students revealed that design (formatting) issues in items can cause some students to struggle, that read aloud accommodations are still necessary for students who struggle with English, and that culturally irrelevant information may mislead or confuse students who are new to this country. The evidence from this study demonstrates that we need further research and activities at the state and district level to ensure that high stakes assessments are both accessible and valid for all students.

INTRODUCTION

Educational culture in the United States has, over the past decade, focused increasingly on demands for accountability for all students. The passage of the *No Child Left Behind Act* of 2001 laid out a framework of educational goals for schools and school districts, requiring them to demonstrate that all students are proficient on state-selected high stakes tests. To this end, districts and schools have responded to the requirements of federal law by evaluating the effectiveness of their own programs to ensure that students are making adequate yearly progress (AYP).

While there has been much discussion about the successes and failures of schools based on high stakes test results, there is a continued need to ensure that the tests themselves are accurate measures of student learning outcomes. The high stakes nature of large-scale assessments requires that careful meta-analyses of the tests themselves are conducted on an on-going basis. If the fate of schools is based on, among other things, the results of large-scale assessments, then the educational community has a responsibility to ensure that such assessments are of high quality and reflect the intended educational outcomes of states and districts.

The need for high quality tests is germane to a society that places high value on the results of such tests. Such a need is exemplified when considered within the context of populations who are typically identified as at risk of school failure, e.g., English language learners. English language learners are a growing and visible population in the school-aged population in our country and reside in urban and rural areas alike.

Butler and Stevens (2001) noted that the 4.2 million English language learners in this country are a heterogeneous group and, like other groups, may have a variety of problems related to high-stakes testing. For example, Abedi, Leon, and Mirocha (2001) and Abedi and Dietel (2004) have found that ELLs score considerably lower on standardized tests than their native speaking peers, especially in the areas of reading and writing. Abedi (2004) has suggested that results on content assessments typically correlate with students' language proficiency level (the lower the student's proficiency the lower they will likely score on content assessments).

Because ELLs, by definition, are not proficient in English, there is the possibility that some of the language found in high-stakes testing could be problematic for this population. Challenging language is often the construct tested in high stakes language arts tests (it is reasonable, and a legal requirement for an eighth grade language arts test to assess eighth grade language abilities). Problems arise, however, when overly complex language that is *unrelated* to the construct tested is introduced.

Messick (1989) expressed concern for "construct irrelevant variance" found in tests. Construct irrelevant variance exists when the construct, or the testing objective, is obscured by excess information in the test item that is not necessary for completion of the item (Messick, 1989). In the case of ELLs, construct irrelevant variance may arise when an item (or test question) has overly complex language that has nothing to do with the item tested, or cultural references that are not familiar to the students. Other examples of construct irrelevant variance are if a test contains extraneous clues, graphics, or text that are not focal points of the test question, and are not necessary in order to answer the test question correctly (Messick, 1989).

When construct irrelevant variance appears to systematically disadvantage particular populations (or when the constructs tested are systematically outside of the schema of particular populations), bias may be present. ELLs most frequently encounter bias in the areas of language and culture. Kopriva (2000), for example, noted that ELLs may not be able to demonstrate the depth of their knowledge and comprehension under the restrictions of large-scale tests that have been designed for mainstream students who share a common culture (i.e., the cultural components found in assessments often do not relate to ELLs). When second language learners are not familiar with the host culture, they are more apt to interpret questions differently than native speakers who are familiar with the host culture (Mohan, 1982).

Baily (2000) has posited that the language found in assessment materials is beyond the proficiency level of many of the ELLs. Specific vocabulary may also pose a problem for ELLs. In an experimentally designed study, Cunningham and Moore (1993) found that when the language of the test was modified, the students' performance increased. Their study showed that when everyday language and vocabulary was used in written comprehension questions, rather than academic language and vocabulary, the students' performance increased. ELLs may become frustrated when they have content mastered but are stymied by the language requirements of the test (Cunningham & Moore, 1993).

Because of the issues that ELLs have historically encountered with high stakes testing, and because the potential for construct irrelevant variance and bias are possible on any test, there is a pressing need for districts and states to examine the tests they use to determine the educational progress of ELLs.

HIGH STAKES TESTING IN MINNESOTA

In Minnesota, the assessments used for system accountability, or to show that all students are making progress toward set academic standards, are the Minnesota Comprehensive Assessments (MCAs). The MCAs are standardized math and reading tests given annually to all 3^{rd}, 5^{th}, 10^{th}, and 11^{th} grade students in the public schools. In preparation for the MCAs, school districts often implement other standardized assessments as a means of obtaining formative data about students and making placement decisions.

In the district where this research took place, a computer-adaptive standardized test is used to prepare students for the types of items they will face. This district (and presumably others) have also implemented it to measure student progress throughout the year. The test used to prepare students for the MCAs was developed to be "self-leveling" or "adaptive," which means that the goal is to test every student at his or her academic proficiency level – this helps teachers find areas to remediate before statewide tests. There is no audio component for the math portion of the preparation test, although the school district recently decided to offer a read aloud accommodation to ELLs who may need additional help.

Because of test security concerns, it was impossible to gain access to actual MCA items. Because the school district where the research took place uses the computer-adapted test as a preparatory test for the MCAs, however, it was decided that the computerized test might provide good information about the potential problems with bias and construct irrelevant variance that the MCA (or any other high stakes test) may have. For the purpose of this research, we used released items from the computerized test, which were available on another state's website.

Although using released, non-MCA items is a less-than-ideal way to truly assess the potential problems with high-stakes assessments in Minnesota, it is a reasonable approach because the computerized test's validity studies have demonstrated that it aligns well with the state standards and test items on the MCAs. The released items from which research was conducted come from computerized test with a large item bank, this item bank has items that have a range of difficulty levels.

METHOD

This study attempted to sort out some of the thorny issues related to high stakes testing and ELLs. Although the studies mentioned above have examined issues of language and bias in tests for ELLs before, this research was intended to inform school, district, and state-level practitioners and policymakers about tests which (as close as possible) mirrored the assessments found in their jurisdictions. The research questions which guided the study were: 1) How do students of varied cultural backgrounds approach typical high stakes test items? 2) What errors do ELLs make in comparison to their native speaking peers at the same independent reading level on math items? And, 3) to what extent is culturally biased vocabulary and concepts present in the math test items selected for this study?

Research Participants

The participants consisted of eight students ranging in age from seven to nine years of age (in grades two to four). The students all resided in a suburban district in the state of Minnesota. Four of the students were native English speakers born in the United States. The other four were ELLs of varying English proficiency levels who were born in Mexico. The ELLs received ELL instruction for 30-45 minutes daily in a separate setting. All of the ELLs spoke Spanish as their first language. Two of the ELL participants were born in Mexico while the other two ELLs were both born in the United States. One has lived in the United States for 3 years, while the other has lived in the United States for one year.

Throughout the study, each ELL was compared to a native speaking peer who was working at the same independent reading and mathematics level in the mainstream classroom. Independent reading and mathematics levels for all the students were determined by using *Guided Reading* assessments (Fontas & Pinnell, 1996; Fontas & Pinnell, 2001) and teacher-developed computation tests. Results of locally-administered assessments (demonstrating similarities in students) are found in Table 1.

Table 1: Local Assessment Results

Student	*Years in US*	*Grade*	*ELL (Y/N)*	*Gender*	*Ind. Reading*	*Math Comp.*

1A	7	2	N	M	H	15/15
1B	7	2	Y	M	I	14/15
2A	7	2	N	M	G	14/15
2B	3	2	Y	F	G	15/15
3A	1	3	N	M	L	14/15
3B	8	3	Y	F	M	15/15
4A	9	4	N	F	P	15/15
4B	9	4	Y	F	P	15/15

Instruments

As stated above, students were tested using released items from a computerized adaptive test that a state uses for its accountability test. Items were selected based on RIT scores from the test that reflected the grade levels of students. A RIT (Rasch Unit) Scale is a measurement based on equal interval scoring that relates directly to a learning continuum or scale (NWEA, 2003; Van Horn, 2003). RIT scores range from 150 to 300 with scores of 150-190 typically found in the third grade and scores of 240-300 in the high school grade levels (NWEA, 2003). The learning continuum contains separate sections for each subject (mathematics, reading, and language usage). All items selected for this research had Rasch scores that ranged from 161 to 191.

Procedure

Each student was given a paper copy of four mathematics test items. The test items were downloaded from a pool of released items found on the Idaho State Department of Education website that released the items. The test items were chosen because they had RIT scores between 161 and 191 (second and third grade difficulty level) and based on the first author's previous experience with similar items from the actual test that seemed to cause frustration for ELLs. The items were printed to represent the format in which the student would see the test item on his or her computer screen during the actual test as best as possible without access to the actual items.

Research was conducted one student and one question at a time. A copy of the first item was given to the student and the student was asked to start by reading the question aloud. The student was then asked to work through the problem and to verbalize his or her thoughts as he or she was working to solve the problem in either English or Spanish (based on student preference). This same process was repeated for each of the four questions. While the students were participating in the think-aloud protocol, their responses were videotaped to ensure that all utterances were captured. One student was not comfortable with the idea of videotaping, so this student's responses were recorded

longhand. Each think-aloud session took between 20-40 minutes, however, there was no time limit put on the student. The same process was repeated for all eight participants.

Throughout the think-aloud activities, a standard script was used. The only cue given to students was to "keep talking." According to Ericsson and Simon (1995), think-aloud activities reveal the most information when facilitators use as few cues as possible. Therefore, the standard script and the occasional reminder to "think out loud" were the only instructions given to students. On occasion, students were unable to complete an item. If the researcher felt additional information was needed, a word in the item was read aloud in order to continue the protocol.

Data Analysis

The data was initially analyzed using a think-aloud coding guide, based on a guide originally developed by researchers at the National Center on Educational Outcomes. The first author of this article watched each videotaped session and then completed each of the areas of the coding guide. The coding guide was divided into the following three categories: Reading of Test Item, Problem Solving of the Test Item, and Questioning. Each of the main categories had very specific sub-categories that were used as a checklist, as well as space to record notes, researcher questions, and student responses.

The coding guide was also used to analyze the reading process and fluency of the student, as well as the problem solving process and product, including a general description of the student's problem solving process. This process was followed for seven of the eight students. One student did not participate in the video taping process, so the responses for that student were written down, rather than video taped. To ensure inter-rater reliability during coding, another ESL teacher also reviewed all of the videos and used the assessment guide to analyze the student responses. There were no discrepancies between the two raters' coding guides.

Next, the coding guides and videotapes were used to transcribe the think-aloud dialogues for each of the students. Transcription was completed one test item at a time for a total of four test items. In total, all thirty-two think-aloud dialogues were transcribed. Dialogues and coding guides were then compared between the ELLs and the native-English speaking students. Patterns as well as discrepancies between the two groups of students in each test item were noted, as well as illustrative quotes that would help readers understand the subtleties of assessment issues.

RESULTS

Transcriptions and coding guides yielded important information concerning the issues that ELLs face on high stakes tests. Although it is impossible to include all transcription and student data in this section, each item is summarized below, preceded by a brief description of the item itself. These items are freely accessible to the public on the Idaho State Department of Education website.

Test Item 1

The first item had a RIT of 161, which would be considered below second grade level. (NWEA, 2002). The learning continuum goal listed for this item was measurement (Idaho Department of Education, 2004). The item consisted of a graphic representation of a calendar that asked students to use calendar skills. This item included multiple-choice questions such as 'What is the first day in December?'.

One of the native English-speaking students had difficulty with this item. This student had difficulty with the item because he paid no attention to the graphic of the calendar. In his mind, he understood the question to be asking him to pick the first day of the week, rather than asking which day of the week was the first in December. His comprehension of the question was at its literal level. The remaining three native English speakers all answered the question correctly and did not appear distracted by the graphics present in the calendar.

Three of the ELLs had difficulty with this question. Two of the ELLs, students 1B and 2B, picked Sunday, an incorrect answer choice because, according to their responses, they saw the sun in the first box of the calendar, which made them think of Sunday. The other ELL, student 3B, chose Sunday because it was the first day of the week seen on the calendar.

Student 4B was the only ELL to answer this question correctly. She was aware of the weather graphics in the calendar, as she mentioned in her verbal responses that the graphics were obviously meant to be distracting. However, the graphics did not seem to distract her or keep her from choosing the correct answer.

Test Item 2

The second question had a RIT of 180, which is a second grade level question. The learning continuum goal listed for the test item was geometry. There were no directions listed for the students beyond the actual test question and answer choices. The item consisted of the question: *Which are polygons?* The answer choices were all in text format. There were no graphics with this question. Below is a representation of what the

students saw during the study, as well as how they would see the item on the computer screen in a real testing situation.

Which are cubes?

1. boxes
2. basketballs
3. keys
4. soup cans
5. magazines

None of the native English-speaking students displayed any difficulty in understanding this question. They all answered the question correctly. However, one of the students, student 2A, had difficulty reading the word "cubes." It was decided to read the word "cubes" aloud to this particular student, because he would have otherwise been unable to complete the activity. In a real test, the student may have guessed or skipped this particular item.

Three out of four ELLs displayed difficulty with part of all of item 2. Student 2B was distracted because she did not know what a soup can was. She knew what soup was (her mother frequently made homemade soup), but she apparently had no previous experience with canned soup and therefore had difficulty visualizing the shape of a can of soup. Although this student was unable to determine if a soup can was a cube, she demonstrated she was familiar with the mathematical concept of a cube. This student drew a picture of a cube while working on the item. In this case, she was simply not familiar with the cultural concept of canned soup.

Students 1B and 4B were not able to read the word "cube." The researcher read the word for these students in order for them to continue with the problem solving process. According to current computerized test procedures, this accommodation would not be available if the student was taking the test independently on a computer. The final ELL student answered the question correctly and seemed to have a good grasp on the concept of a cylinder as well as types of real life objects that could be classified as cylinders.

Test Item 3

The third question had a RIT range of 191, a third grade level question. The learning continuum goal listed for the question was problem solving. There were no directions listed for this item. The item consisted of the test question and five answer choices. There were no graphics for this question. Below is the question the students saw during

the study It is also how they would see the item on the computer screen in a real testing situation.

Randy needs 30 tokens to get a CD. He has 10 tokens and a friend is giving him 5. How many more will he need before he can get the CD?

1. 35
2. 15
3. 25
4. 10
5. 5

Test Item #3 presented the least difficulty for both native English speakers as well as ELLs. All of the native English-speakers answered the question correctly. None of them struggled with reading any of the text present in the test item. One of the students, student 2A, made reading errors in the test item, but those errors did not interfere with his ability to comprehend to test item.

In addition, all of the ELLs answered this item correctly. There was one concern with comprehension of the concept of tokens for student 2A. She was not familiar with the process of collecting tokens in order to exchange them for prizes. However, this did not seem to hinder her ability to complete the problem solving process. She answered the item correctly without the knowledge of token collecting. Whether or not she would have been able to answer the question without encouragement to "keep talking" and "keep thinking" is unknown.

Test Item 4

The fourth item had a RIT of 190, which is considered at third grade level. The learning continuum goal listed for this item was Number Sense and Numeration. There were no directions listed for the test item. The item consisted of the question and five answer choices. The question read as follows: *In what place is the letter r in the word scooter?* The answer choices were ordinal numbers displayed in text format as follows: A. first, B. third, C. fifth, D. seventh, E. ninth. Below is a representation of what the students saw during the study, as well as how they would see the item on the computer screen in a real testing situation.

In what place is the letter r in the word scooter?

1. first
2. third
3. fifth

4. seventh
5. ninth

All of the students in the study completed test item #4 correctly. One of the native English-speaking students, student 1A, initially struggled with the *th* ending for the ordinal numbers, but she was able to correct her mistake and it didn't keep her from completing the problem solving process correctly.

Many of the concerns for the ELLs were in the area of reading. Student 2B and 3B had some difficulty reading some of the words in the test question as well as in the answer choices. Student 2B had difficulty reading the word "scooter." Eventually, the word "scooter" was read aloud by the researcher in order to facilitate completion of the activity. The student was able to work through the reading of the ordinal numbers independently and was able to choose the correct answer. Student 3B had difficulty reading the ordinal numbers that were present in the answer choices. This student made the mistake of using a short vowel sound rather than the long vowel sound. However, he made the correction himself, and his reading did not affect his ability to answer the item correctly.

Student 2A lacked knowledge of what a scooter was. Her inability to read the word "scooter" acted as a distraction, but the student was able to complete the problem solving process despite her lack of vocabulary. It is unknown whether the student would have given up or guessed in a real testing situation.

Student 3B had some difficulty understanding the format of the question. He did not understand that the *r* was simply an underlined version of the letter "r." The student looked at the letter and asked, "What is this?" He made a comment that he thought it was a "decoration, or symbol" However, after reading the question and answer choices, he was able to figure out that it was the letter "r," and was able to complete the problem solving process correctly.

Results that emerged from the items above indicate that there was not wholesale bias issues present in the items. Issues of some sort, however, were present in all items. The think-aloud methods used provided insights into the meta-issues that students face when completing test items. Such issues may lead to incorrect answers or simply increased aggravation and frustration because of students' lack of familiarity with the cultural or linguistic loads of items. For example, Item 1 had construct-irrelevant information (weather graphics), which distracted ELL students more than English proficient students. Item 2 contained a cultural icon that was unfamiliar to one of the ELLs (a soup can) and tested a mathematical concept that was appropriate, but difficult for some to read. Item 3 presented few, if any problems for students. One ELL struggled with a culturally-bound concept (tokens for prizes), but was able to overcome confusion and answer the question correctly. It is unknown how this concept may have affected

the larger ELL population. Finally, one ELL did not have knowledge of what a scooter was, but was still able to complete the item. Another ELL struggled with the underlined letter "r" in text.

DISCUSSION

This study was a qualitative study with a small sample size. Although data were rich and descriptive, readers should note that findings from this study may not generalize across settings. Furthermore, the experiences of the students who participated in this study may not generalize to the experiences of all ELLs. Teaching and assessing ELLs takes careful consideration because of the heterogeneity of students both within and between populations whose first language is other than English. Further study is needed to quantify the relative impact of item design on students who are learning English as a second language. Although research done at a statewide or national level will be more instructive in concretizing the impact of item design on specific populations, this study did have several noteworthy findings.

Overall, this study found that there are subtleties in every test item that may introduce construct irrelevant variance or bias. Although evidence from think-aloud studies did not find that there were wholesale errors made on any item because of design flaws, design issues did have effects on student performance and comfort levels during the research. Based on the results of this research, we reiterate what research has told us already, that bias and construct irrelevant variance are sometimes found in assessment items, and that bias and construct irrelevant bias appears to differentially affect English language learners.

Based on the results of this study, three major themes arose. Such themes are important for stakeholders to consider when preparing and taking high stakes testing. The first overall theme is that construct irrelevant variance may differentially affect English language learners. Based on prior research, we know that ELLs struggle with reading and comprehending test items. Such learners may then look for assistance from graphic information (as demonstrated in item 1). If the graphic information in an item is misleading or irrelevant, students may be led astray. Therefore, careful consideration must be made to the added value of visuals for students. What test developers may think are interesting visuals may be distracting to students.

Related to the effects of reading difficulties is the need for read aloud accommodations for ELLs. Students clearly understood what a cylinder was, but had difficulty decoding the word. In mathematics tests, the constructs tested are mathematics. Therefore, tests such as those drawn from for this study should allow read aloud accommodations for ELLs who struggle with reading. By eliminating extraneous factors (such as reading

ability), schools and districts can get a more valid understanding of students' mathematical, science, and social studies ability.

Finally, although the effects were minimal, cultural bias was found in three of the items. Soup cans, tokens, and scooters were all unfamiliar terms to students. The items described in the previous sentence seem like they would be common to all students, but think-aloud data demonstrate this may not be the case. This presents a challenging dilemma for test developers and illustrates the need for sensitivity reviews that can assess test items for their appropriateness. The effects on problems solving were varied, but including such terms may produce a level of discomfort or confusion for ELLs that students familiar with such terms do not face. Minor discomfort when taking tests is not a well-researched issue, but when stakes are high, all facets of testing are important to consider.

The results from this study reflect the current thinking on assessments and reiterate the importance of considering construct irrelevant variance, bias, and the need for accommodations in testing. While this article was meant to add to the larger body of knowledge in assessment, its primary purpose was to inform district-level leaders about possible issues in statewide assessments for ELLs and encourage replication studies by others.

Teachers and administrators interested in improving the conditions in high stakes assessment may wish to take an active role in ensuring that construct irrelevant content and bias are removed from tests and that appropriate accommodations are allowed. Teachers may contact their state department of education representatives to find out when "sensitivity review panels" are conducted for high stakes tests. These panels allow constituents to analyze items for possible bias or other distracting content. Teachers knowledgeable of ELL issues would be helpful additions to such panels.

Additionally, district administrators have choices concerning the types of assessments they choose to prepare students high stakes tests. When selecting tests, administrators can be sure that accommodations can be built into the test structure. Accommodations such as those that allow tests to be read to students are a valid approach to testing ELLs in mathematics. Principals and administrators will need to consult state policies on accommodations for high stakes tests themselves.

Overall, work the test items in this research did not have flagrant issues, but each item had subtle biases and other construct irrelevant variances. Therefore, there is a continued need to research and address such issues. Research and action is the responsibility of all stakeholders, including parents, university personnel, administrators, and teachers. As educators, expect our students to perform at their highest level. As

consumers of assessments, we need to hold the same high expectations of the tools used to measure students.

Authors

Andrea Erichsrud is an ELL teacher in the Fridley Public School District. She has been teaching for 10 years, 8 of which have been spent working with second language learners. She is particularly interested in testing and assessment issues as they relate to ELLs.

Christopher Johnstone is an Assistant Professor of Special Education at Augsburg College and a Research Associate at the National Center on Educational Outcomes. His research focuses on developing accessible assessments for all students, including students with disabilities and English language learners. Recent projects include studies on Universal Design, accommodations, accessible reading assessments, and state special education reporting.

References

Abedi, J. (2004). Will you explain the question? *Principal Leadership*, *4*(7), 27-31.

Abedi, J. & Dietel, R. (2004) Challenges in the No Child Left Behind Act for English-language learners. *Phi Delta Kappan, 85*(10) 782-785.

Abedi, J., Leon, S., & Mirocha, J. (2001). Validity of standardized achievement tests for English language learners. Paper presented at the American Educational Research Association Conference, Seattle, WA.

Baily, A. (2000). Language analysis of standardized achievement tests: Considerations in the assessment of English language learners. In *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (pp 85-105). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Teaching (CRESST).

Butler, F., & Stevens, R. (2001). Standardized assessment of the content knowledge of English language learners k-12: Current trends and old dilemmas. *Language Testing, 18*(4) 409-427.

Cunningham, J. & Moore, D., (1993). The contribution of understanding academic vocabulary to answering comprehension questions. *Journal of Reading Behavior, 25*, 171-80.

Ericsson, K., & Simon, H. (1995). *Protocol analysis: Verbal reports as data, 2nd edition*. Cambridge, MA: MIT Press.

Fontas, I. & Pinnell, G., (1996). *Guided reading: Good first teaching for all children.* Portsmouth, NH: Heinemann.

Fontas, I. & Pinnell, G., (2001). *Guiding readers and writers*. Portsmouth, NH: Heinemann.

Idaho Department of Education (2004). Organization Webpage. Retrieved January 12, 2004 from http://www.sde.state.id.us/dept/standards.asp.

Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners.* Washington, DC: The Council of Chief State School Officers.

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

Mohan, B. (1982). What are we really testing? In P. Richard-Amato, & M. Ann Snow (Eds.), *The Multicultural Classroom: Readings for Content-Area Teachers* (pp. 258-269). Reading, MA: Addison-Wesley Publishing Company.

Northwest Evaluation Association. (2002). *Monitoring Growth in Student Achievement.* [Brochure]. Lake Oswego, OR: NWEA.

Northwest Evaluation Association. (2003). *Teacher's handbook for measures of academic progress*. [Workshop Brochure]. Portland, OR: NWEA.

Northwest Evaluation Association (2004). Organization Webpage. Retrieved November 1, 2004 from http://www.nwea.org/about/history.asp.

Van Horn, R. (2003). Computer adaptive test and computer-based tests. *Phi Delta Kappan*, *84*(8), 630-631.