2008-08-20

Investigation of the Inter-correlations Among Standardized Written and Performance-based Assessments of Measurement Content Knowledge Among Third Grade English Language Learners

Marcella D. Elliott
University of Miami, m.elliott@mycanes.miami.edu

Follow this and additional works at: https://scholarlyrepository.miami.edu/oa_dissertations

Recommended Citation
https://scholarlyrepository.miami.edu/oa_dissertations/148

This Open access is brought to you for free and open access by the Electronic Theses and Dissertations at Scholarly Repository. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of Scholarly Repository. For more information, please contact repository.library@miami.edu.
UNIVERSITY OF MIAMI

INVESTIGATION OF THE INTER-CORRELATIONS AMONG STANDARDIZED WRITTEN AND PERFORMANCE-BASED ASSESSMENTS OF MEASUREMENT CONTENT KNOWLEDGE AMONG THIRD GRADE ENGLISH LANGUAGE LEARNERS

By

Marcella D. Elliott

A DISSERTATION

Submitted to the Faculty of the University of Miami in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Coral Gables, Florida

May 2008
The purpose of this study was to provide an empirical test of the widely held belief that performance-based assessment provides a fairer picture of English-language learners’ mathematical skills and knowledge than does a standardized assessment. Specifically, I compared the performance of 94 third-graders on the measurement subscale of the Florida Comprehensive Assessment Test (FCAT) mathematics test to their performance on a set of measurement reasoning and applications that was drawn from their third-grade hands-on science curriculum. Then, I present examples within the non-standardized testing setting where students were provided with real-time language-based accommodations as recommended by the research literature. Finally, I looked at how well these students’ level of English language proficiency predicted performance on each of the two assessments. English proficiency level failed to predict FCAT performance. It did predict performance on the reasoning and applications tasks. These findings present a challenge to the conventional wisdom that performance-based assessments provide a less-biased picture of ELL’s mathematical knowledge than do standardized tests.
Dedication Page

This work is dedicated to my parents Cynthia (nee Davis) Elliott and the late George Albert Elliott. Through their hard work, dedication and undying love for their children, they taught me that all things are possible with the help of the Almighty God and the support of family and friends. They taught me what is important in life and how to live one’s dream. They gave me a passion for learning and the importance of dedication and commitment in the pursuit of one’s goals. I recognize how privileged I am to have grown up in their household. The values which they have passed on to me have shaped my life. I am honoured to be a part of their legacy and committed to pass it on to all with whom I come into contact.
Acknowledgements

The pursuit of this Ph.D. has been a journey that may not have been possible without the assistance and support of many. The support that I have received falls into three categories; personal, academic and community.

I am pleased to publicly acknowledge the personal support of my mother, Cynthia Elliott; my family members, Alphonso Elliott, Lawrence Elliott, Clovis Elliott and Jerome Elliott; the Elliott and Davis families; and my close friends Joan Lightbourn-Jones, Gilbert Elliston, Sharlene Johnson, Janice Bowleg, Jaime Maerten-Rivera, Mark Rivera and Margarette Mahotiere.

I also would like to acknowledge the academic support of my faculty advisor, Dr. Walter G. Secada, the members of my dissertation committee; Drs. Secada, Lee, Buxton and Penfield, the team members of the P-SELL project, and the National Science Foundation that provided funding for the research that I conducted under the umbrella of P-SELL. I also acknowledge the support I received from faculty, staff and fellow students at the University of Miami, particularly in the School of Education.

Lastly, I would like to acknowledge the support of my church and neighbourhood communities in The Bahamas and Doral, Florida, the faculty and staff of The College of The Bahamas and numerous friends throughout the Caribbean and around the world.

I am grateful for your support and kindness along this journey. You have given to me more than you may ever know. May the Almighty God continue to bless us all.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF FIGURES</th>
<th>vi</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>Chapter</td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>2 RELATED LITERATURE</td>
<td>4</td>
</tr>
<tr>
<td>3 METHODOLOGY</td>
<td>24</td>
</tr>
<tr>
<td>4 RESULTS</td>
<td>33</td>
</tr>
<tr>
<td>5 CONCLUSIONS, DISCUSSION, IMPLICATIONS</td>
<td>40</td>
</tr>
<tr>
<td>References</td>
<td>47</td>
</tr>
<tr>
<td>Appendix A</td>
<td>55</td>
</tr>
<tr>
<td>Appendix B</td>
<td>103</td>
</tr>
<tr>
<td>Appendix C</td>
<td>113</td>
</tr>
<tr>
<td>Appendix D</td>
<td>118</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 2.1 Unit Iteration ................................................................. 5
LIST OF TABLES

Table 2.1 Average difference in scores on NAEP 2000 ........................................ 9
Table 2.2 2000 Western Australia mean measurement scores by sub-group .......... 9
Table 2.3 2002 Mean Scores on the Measurement Strand by Sub-group ................. 9
Table 2.4 Alignment of Instruction with Assessments: 8th Grade Math-SCASS Study ... 14
Table 3.1 Student gender .................................................................................. 25
Table 3.2 Student ethnicity ................................................................................ 25
Table 3.3 Student ESOL Level ............................................................................. 26
Table 3.4 Student Combined ESOL Level .......................................................... 26
Table 4.1 FCAT (Standardized Test) Measurement Results ................................. 37
Table 4.2 Performance-based Assessment Results ............................................. 38
Table 4.3 Multiple Comparisons ........................................................................ 39
Chapter 1: Introduction

English language learners (ELLs) encounter difficulties in an environment where English is the language of instruction and assessment. This learning environment is challenging for students who must adapt to a new language and culture while being expected to perform at acceptable levels at school. Nationally, students generally perform 10% to 30% worse on arithmetic word problems than on comparable problems presented in numeric format (Carpenter et al., 1980) because language is an important component of student performance on assessments in mathematics (Abedi, Lord, Hofstetter, 1998; De Corte, Verschaffel, & DeWin, 1985, Garcia, 1991). In particular, ELLs score lower than English-proficient students do on standardized tests of mathematics achievement in elementary school, SAT and the GRE (Abedi & Lord, 2001). Although there is no evidence to suggest that the abilities of ELLs are different from non-ELLs, the achievement differences between ELLs and non-ELLs are pronounced (Cocking & Chipman, 1988).

There has been a heavy reliance on standardized testing as an indicator of student achievement and performance. Some researchers have long criticized this practice. According to Haertel (1999), Resnick, (1987a), Wiggins, (1993), Wigdor & Garner (1982), one identified "past mistake" has been the over-reliance on standardized multiple-choice tests; and one identified “solution” has been to add in performance assessments. Because of this, performance and/or authentic assessment were at the centre of state and national education reform initiatives in the 1990s.

With performance-based authentic assessment, students are thought to be afforded the opportunity to demonstrate what they know, can do, and/or are thinking in a manner
that is not so heavily dependent on language skills. One potential benefit of performance assessment is that it allows for real-time use of language-based accommodations. When ELLs are tested in subject matter areas, their performance is thought to be systematically under-estimated (i.e., biased) because of the language in which the test is presented. Language-based accommodations are seen as one way to mitigate the language-based bias of tests for the ELLs. In addition, performance assessments are thought to afford educators a more complete picture of what the students really know and are able to do against real-world tasks for which students have had opportunities to learn.

Rivera and Collum (2004) contended that, compared to English-proficient students, ELLs must direct more cognitive resources to processing the language of the test and therefore they have fewer resources to attend to the content of the test. Language-based accommodations are intended to minimize the cognitive resources that ELLs need to process the language, thereby affording more resources to be directed to the content of the test itself. Hence, they are thought to be a more appropriate way to measure ELLs’ progress (Lenski et al., 2006).

Research Questions

The mismatch between assessment practices and student understanding is what motivated this investigation of third grade students’ understanding of measurement using two types of assessments. The purpose of this study was to provide an empirical test of the widely held belief that performance-based assessments provide a fairer picture of English language learners’ mathematical skills and knowledge than does a traditional, standardized assessment. I investigated the relationship of third-graders’ English language proficiency levels to two forms of assessment in the content area of
measurement. The first assessment is the Florida’s Comprehensive Assessment Test (FCAT); and the second, a performance-based activity done via interview. I also looked at how some of the accommodations recommended by Abedi (2002) for ELLs are instantiated within a performance-based assessment setting.

I pursued the following research questions:

1. How well do two different kinds of assessments (standardized versus application-oriented performance-based interview) tap into the same construct of measurement?

2. What evidence is there that language-based accommodations are used in performance based interviews?

3. Is there evidence of language-based bias in one form of assessment over the other?

This investigation begins to address some concerns that researchers have in assessing ELLs, in general and in particular, in the area of measurement. This should provide an indication of what works best for the ELL in an effort to provide equitable education for all students.
Chapter 2: Related Literature

In this section, I will review the literature related to measurement and the assessment of ELLs using the commonly prescribed accommodations. This should highlight some of the challenges that ELLs face in learning and assessment and how these challenges are being addressed by researchers, educators and district policies.

Measurement

Measurement has been described as “the assignment of a numerical value to an attribute of an object” (NCTM, 2000). It involves relating numerical quantity with a spatial attribute such as a line, an area or a volume (Bishop, 1983). A purpose of measurement is to make comparisons, directly or indirectly, between two or more objects. Measurement might be a simple task for adults; however, it is not trivial for children. Measurement involves several mathematics concepts. Piaget et al. (1960) demonstrated that children must use two kinds of reasoning to measure objects: transitive reasoning and unit iteration. Transitive reasoning involves comparing one quantity to another. Unit iteration is the ability to think about a whole as consisting of equal parts (units). Measurement consists of identifying a unit of measure, subdividing the object by that unit and placing the unit end-to-end (iterating) along the side of the object being measured (Stephan & Clements, 2003).

Students’ Misconceptions and Challenges

Measuring is more advanced than simply counting. Inhelder, Sinclair, and Bovet (1974) illustrated this difference in the following example. Students were shown two equal rows of matches. (See Figure 2.1 below.)
The students argued that the row with six matches was longer because it had more matches. In another illustration, second grade students incorrectly measured the length of a side of a rectangle when given rulers that had the first two inches broken off (Barrett et al., 2003). Students read the length of the side as nine instead of seven. They did not take into consideration that the “zero point” was not zero but two.

Stephan, Cobb and Gravemeijer (2003) describe a similar misconception that students have when they measure by covering distances. When students were measuring the length of a rug with their feet, most did not count the first position of their foot as “one”, but they started counting only after they moved to the second step. They did not see their initial pace as covering distance. They considered that they had only covered a distance when they moved. With the teacher’s help, students were able to describe their methods and their interpretations of what measuring length is. When children are asked to count the number of steps in a stairwell, many do not know whether to count the first step or the last step and some count both. This misconception is common among young children.

Barrett (2003) found that children’s thinking about length fit one of three profiles. Students in the first profile are able to compare objects but do not use units. They may recognize that one pencil is longer than another is, but they cannot determine by how much. Students in the second profile use addition and some notion of units. They measure by adding the unit measures. However, they encounter difficulty if non-unit amounts are left over. Students in the third profile use multiplication and units in their measurement
tasks. These students are able to measure by multiplying the number of unit measures. Students who have only step-by-step procedural knowledge of measurement will not have a high degree of mastery of this process. Traditional measurement instruction is insufficient for helping students build these concepts. Clearly, students need both step-by-step procedural and deeper conceptual understandings to be able to measure properly.

Students at the elementary and middle school level, while competent in basic ruler skills, seem not to have acquired important concepts about how rulers work and the units of length. Bragg and Outhred (2004), in their study of 209 public school students in Australia, found that by the end of elementary school approximately sixty-three percent of students were still unable to identify the unit of measure for length on a ruler. They recommended that early measurement activities include explicit instruction in the relationships between informal units and the construction of scales on rulers.

Students also appear to have a poor understanding of the need to use identical units that leave no gaps when measuring. Curry et al. (2006) interviewed 96 students to investigate their understanding of length, area and volume measurement. They found clear differences in students’ understanding of length, area and volume. Volume was more difficult for students. Length and area measurement were similarly difficult. They attributed their findings to the fact that most students encounter more measurement of length and area in the daily lives, than they do of volume.

**Measurement Estimation**

Measurement estimation is important in its own right because it is often the only method available for solving problems (Levin, 1981). Usiskin (1986) commented that although estimation may be considered the “weaker sister” to computation, it is often the
“stronger sister” in real life because tools may not be readily available and estimation may be the only option that is available. In addition, successful estimation skills make it unlikely that students will mindlessly apply memorized procedures, but more likely, that they will develop a conceptual understanding of number sense. Decades ago, Piaget, Inhelder and Szeminska (1960) conducted a task that illustrated the importance of transitivity in estimation. Students were shown two towers, $A$ and $C$ and asked which was taller. They replied that $A$ was taller. Then tower $A$ was hidden and they were shown towers $C$ and $B$ to determine which is taller. Students correctly replied that $C$ was taller. The interviewer then asked them to say which is the taller of $A$ and $B$. The four year olds said that they could not tell because they had not seen $A$ and $B$ together. Older children who had transitive reasoning skills were able to deduce that if $A$ is taller than $C$ and $C$ is taller than $B$, then $A$ must be taller than $B$. Piaget et al. concluded that for students who cannot reason transitively, rulers are useless for comparing two lengths that are not placed next to each other.

**Assessment of Measurement**

*Types of Assessments*

Traditional standardized tests are written, timed, usually multiple-choice assessments that are relatively inexpensive to administer. They typically do not afford students the opportunity to demonstrate their knowledge through a realistic performance. Standardized tests generally reveal only whether the student can recognize, recall, or “plug in” what was learned out of the context in which it was learned. While many standardized tests now include short and extended responses, the measurement strand of the FCAT only includes multiple-choice items. These conventional tests typically ask
students to select or write correct responses, irrespective of reasons (Archibald & Newmann, 1988; Butler & Stevens, 2001; Lenski et al., 2006).

Authentic assessments provide information that is directly related to specific tasks on which students succeed or fail. Moreover, authentic tasks are supposed to be meaningful and significant, affording students the opportunity to demonstrate what they know and can do. Assessments that provide little substantive information and lack authenticity undermine the legitimacy not only of the numerical indicators, that are so heavily used, but also of the educational system (Archibald & Newmann, 1988).

Performance-based assessment, with all its concerns, is generally seen as a more authentic method of assessing what students know and can do (Abedi, 2002; Aguirre-Munoz, 2000; Lenski et al., 2006). Performance assessment methods require students to demonstrate what they know and can do by applying specific skills and competencies to a task. These tasks are often referred to as authentic assessments because they typically are designed to involve students in responding to a real-life or realistic task or problem.

**Differences in Achievement**

On large-scale national and state assessments, Hispanic and African American students have performed significantly lower than white students do (NAEP, 2003, 2007; FCAT, 2007). In some cases, this difference is as much as twenty-six to thirty points, depending on grade level. Although the performance gap among Hispanic and white students has been closing over time on tasks that assess basic procedural knowledge and skills, substantial performance differences remain on tasks that assess conceptual understanding, mathematical reasoning, and problem solving (Silver et al., 1996).
The figure below, adapted from Lubienski, Camburn and Shelley (2004), summarizes the differences in achievement between minority groups in the NAEP 2000 assessment. Overall, we see that Hispanic-White gaps are large, but not as large as the Black-White gaps.

**Table 2.1**

*Average difference in scores on NAEP 2000*

<table>
<thead>
<tr>
<th></th>
<th>4th Grade</th>
<th>8th Grade</th>
<th>12th Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispanic-White</td>
<td>24</td>
<td>33</td>
<td>25</td>
</tr>
<tr>
<td>Black-White</td>
<td>31</td>
<td>39</td>
<td>34</td>
</tr>
</tbody>
</table>

Similar results have been reported on the measurement strand in Western Australia, where aborigines and ELLs perform lower than their white counterparts do (Western Australia Department of Education, 2000, 2002). These results are summarized below.

**Table 2.2**

*2000 Western Australia mean scores on measurement strand by sub-group*

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>1630</td>
<td>481</td>
<td>105</td>
</tr>
<tr>
<td>Girls</td>
<td>838</td>
<td>486</td>
<td>104</td>
</tr>
<tr>
<td>Boys</td>
<td>792</td>
<td>477</td>
<td>105</td>
</tr>
<tr>
<td>ATSI</td>
<td>218</td>
<td>385</td>
<td>101</td>
</tr>
<tr>
<td>LBOTE</td>
<td>226</td>
<td>449</td>
<td>98</td>
</tr>
</tbody>
</table>

**Table 2.3**

*2002 Mean Scores on the Measurement Strand by Sub-group*

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>1071</td>
<td>528</td>
<td>99</td>
</tr>
<tr>
<td>Girls</td>
<td>528</td>
<td>525</td>
<td>97</td>
</tr>
<tr>
<td>Boys</td>
<td>543</td>
<td>531</td>
<td>100</td>
</tr>
<tr>
<td>ATSI</td>
<td>205</td>
<td>457</td>
<td>87</td>
</tr>
<tr>
<td>LBOTE</td>
<td>136</td>
<td>527</td>
<td>103</td>
</tr>
</tbody>
</table>

ATSI = Aborigine
LBOTE = Language Background Other Than English
Possible Reasons for Differences in Achievement

To explain the disparities, some have claimed that lower SES and minority students, which include ELLs, do not receive the same instruction as their white counterparts. Minorities generally receive predominantly rote-based mathematics instruction (Anyon, 1981; Ladson-Billings, 1997; Means & Knapp, 1991). In contrast, white students appear to experience more of the fundamental shifts called for in the NCTM Standards (1989). Even high-SES African-American students were less likely than low-SES white students to be permitted to use calculators in 8th grade (Lubienski, Camburn, & Shelley, 2004). In general, examinations of race/SES interactions have found instructional differences between black and white students that persist even after accounting for student SES (Lubienski et al., 2004).

McNeil (2000) described the effects of differential education for minority children and ELLs in mathematics. Their education emphasizes test-taking strategies and low levels of knowledge. The study of mathematics becomes “having students choose among four or five possible answers. They are not asked to explain their answers, so if students have alternative ways of working a problem, their reasoning is not made visible on the test nor are their reasons for selecting the ‘correct’ answer” (McNeil, 2000, p.241).

Another popular explanation for achievement differences is that traditional assessments do not reflect ELLs’ true understanding of content knowledge (Solano-Flores & Trumbull, 2003; Abedi, 2004; Lenski, 2006). While ELL students have difficulty writing in English, they can often speak about mathematics and science with a level of sophistication not reflected on written assessments alone (Lenski et al., 2006).
This inconsistency may be due, in part, to language-based demands of the written assessment that exceed oral expectations.

Performance-based authentic assessment provides a possible solution to concerns about curriculum coverage and the demands for writing since it is supposed to reveal students’ strengths and weaknesses, while mitigating the bias of language-based tests. Authentic assessment could also be used to improve teaching and learning since it is based on what is actually taught (Wiggins, 1989). Traditional assessment, by contrast, relies on indirect, cost-efficient, simplistic substitutes from which we think valid inferences can be made about student’s curriculum-based performance (Wiggins, 1989; Abedi, 2002). There is strong argument that in order to evaluate students’ problem-posing and problem-solving techniques in mathematics or doing experimental research in science, assessment should be based on those tasks. Assessment should also allow for different starting points of learning and for diverse ways of demonstrating competence.

**Test Validity**

A concern with all assessments, but in particular, performance assessments is the question of validity. According to Standards for Educational and Psychological Testing (1999), “validity is unitary concept which refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests (p. 9)”.

A fundamental notion in test validity is that test scores should not be related to factors that are irrelevant to the content (Messick, 1989). As noted over seventy-five years ago, any test that employs language is to some degree a measure of language skills (Abedi, 2006; LaCelle-Peterson & Rivera, 1994; Sanchez, 1934). As Abedi (2002) asked: “How valid are inferences about students’ knowledge when they are based on a test
administered in a language that students may not understand?” Recognizing the importance of this question, the 1999 edition of *Standards for Educational and Psychological Testing* devoted an entire section to testing English language learners, cautioning that test results should be weighed carefully because of the potential for confounding influences due to language deficiencies.

According to the *Standards* (1999), to evaluate test validity, one should begin with an explicit statement of the proposed interpretation of the test scores. This should include a clear description of the constructs or concepts that the test is intended to measure. Only then can one begin to discuss how valid a test might be, by measuring it against its intended purpose. The validation process is the development of a scientifically sound set of arguments to support the intended interpretation of the scores. While many may view this as the sole responsibility of the test developer, validation is the joint responsibility of both the test developer and the user. While the test developer is responsible for the development of the test, the user is indirectly responsible for providing evidence in support of its use. Student scores can be analyzed to give support to the argument for validity of the test.

**Sources of Evidence of Validity**

While there are several sources of evidence that can be used to support the argument for test validity (*Standards*, 1999), the ones highlighted here are those directly related to the assessments of ELLs.

*Content.* Evidence that is based on test content refers to the wording, themes, questions and format of the test, as well as guidelines for administration and scoring. This source of evidence can be evaluated by “experts” in the field who may supply test items
or vet them. Evidence about content can be used to start to address questions about
differences in scores across different subgroups. Upon examination of the test, experts
may highlight sources of irrelevant difficulty for certain subgroups, for example, how
broadly the content of a test is constructed (if it includes real-world, real-time
applications or if it is strictly limited to what is taught). Measurement in the traditional,
standardized test may be restricted to knowing how to read a measurement from a chart;
whereas a performance-based assessment might afford students the opportunity to
actually select the appropriate tool and do the measurements. In addition, a performance-
based assessment might highlight how deeply the content was covered by affording
students the opportunity to talk about what they are doing and how it relates to what they
learned at school or at home. When students are asked low-level questions, the answers
are not likely to reveal high-level thinking.

*Match to the intended construct.* Another source of evidence is based on the
internal structure of the test. This looks at the degree to which the relationships among
test items and test components conform to the intended construct. Standardized tests like
the SAT and the GRE make no claim about alignment to instruction, but the FCAT does.
Porter (2002) conducted content analyses of six states’ eight-grade assessments to
investigate the degree of alignment to instruction. The average within-state alignment
was .22 and the alignment of instruction to NAEP was slightly higher at .39. Table 2.4
adapted from Porter (2002), highlights this.
Table 2.4

*Alignment of Instruction With Assessments: Eighth-Grade Math—SCASS Study*

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Assessment</th>
<th>State</th>
<th>H</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>E</th>
<th>O</th>
<th>NAEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td></td>
<td>.35</td>
<td>.22</td>
<td>.19</td>
<td>.28</td>
<td>.21</td>
<td>.04</td>
<td>.38</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td></td>
<td>.34</td>
<td>.21</td>
<td>.18</td>
<td>.25</td>
<td>.20</td>
<td>.05</td>
<td>.38</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td></td>
<td>.42</td>
<td>.28</td>
<td>.21</td>
<td>.29</td>
<td>.25</td>
<td>.05</td>
<td>.39</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td>.36</td>
<td>.24</td>
<td>.19</td>
<td>.29</td>
<td>.22</td>
<td>.05</td>
<td>.40</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td></td>
<td>.35</td>
<td>.21</td>
<td>.16</td>
<td>.26</td>
<td>.20</td>
<td>.05</td>
<td>.38</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Average within-state alignment = .22; average between-state alignment = .23; average state-test-to-NAEP alignment = .39.

Authentic assessment should be aligned with instruction (Stiggins, 1994; Valencia, 1990; Wiggins, 1989) since items are taken directly from what is taught.

A concern is whether tests are tapping into the same knowledge of measurement. Test items may look like they do. However, Test A and Test B may look similar, but may not give similar results. If they do not give similar results, then they are tapping into either different constructs or different dimension of the same construct.

The student’s response process can also provide evidence concerning the fit between the construct and the response enacted by the students (Standards, 1999). For example, we are more likely to assume that a task, which actually requires students to measure things, taps into measurement than a task that can be answered by guessing at the answer.

Capabilities irrelevant or ancillary to the construct may influence student performance. For example, a student may be able to measure in inches but not in cubits (an old system of measurement). In addition, the consequences of the test can also provide evidence relevant to test validity. Students and teachers feel the pressure of performing well, particularly if their promotion or salary is tied to the results. When there
is evidence of wide-spread cheating, the validity of the intended interpretation is threatened.

Use of tests. A sound argument for the valid use(s) for any test should include and integrate several of the aforementioned sources of evidence. The degree of validity for an intended interpretation of a test score should rely on all available evidence (Standards, 1999). This includes evidence of test construction, appropriate test administration and scoring, and accurate score scaling. The Standards document outlines twenty-four standards that ought to guide test development, administration and scoring. In the interest of brevity and relevance, a few are highlighted here that might be more applicable to the assessment of the ELL.

A rationale should be clearly outlined for each interpretation and use of the test scores. The population(s) for which the test is appropriate should be clearly described. If a test is used in a different way, it is incumbent on the user to justify the new case. Since ELLs are typically omitted from the development of a standardized test, the use of any test with ELLs needs to be justified.

While one objective of FCAT is to assess students’ knowledge of content, the test is also a test of language proficiency. This questions the validity of the test, particularly with regard to ELLs. In an environment where the home language is the language of instruction and the language of assessment, this concern is lessened.

When validation rests on “experts”, procedures for selecting experts should be clearly described. When validation is determined by appropriateness of test content, the procedures followed for item development should be described and justified. Since the demographic makeup of the country’s student population is always changing and since
the environment under which the tests are taken and the use of scores are constantly evolving, test validation should go through several iterations and should be a continuous process.

I did not study how tests might predict future achievement or measure the development of real-world abilities, since additional factors, such as home environment and educational opportunities are known to influence later life achievement. I focused on the construct being assessed (measurement, in this case) following Kane (2001), who recommended that we focus on the degree to which a test measures the intended content of student’s knowledge. In this case, I looked at measurement and its applications.

Test Bias

Bias is the systematic tendency for one group to perform better or worse than another group for reasons that are not intended by what is being measured. It is the degree to which a test systematically over-estimates or under-estimates true performance on a test for reasons that are unrelated to one’s knowledge of the test content. A test item is possibly biased when the probability of a correct answer is associated with group membership after controlling for the target trait intended to be measured (Camilli & Shepard, 1994). There are many possible sources of bias. For example, ELLs may be administered a mathematics test that is written beyond their ability to read. In this case, the students who performed poorly may do so not because they did not know mathematics, but because they could not read the test. In this case, the test contains a language-based bias against less proficient readers or ELLs. Another example is the cultural references or background information that ELLs need and might be lacking in order to correctly answer some items. Source of bias could also arise if the person who
administers the test and/or scores the test knows and/or likes the student. Test performance could be over-estimated because of that person’s subjectivity.

Identifying and Removing Bias

There are two general ways for identifying bias: judgmentally and empirically (Shepard, 1981). Judgmental reviews are conducted throughout the development of the test and concerned with the individuals (judges) representing the relevant subgroups, such as ELLs, gender, students with disabilities, etc. Judges are asked to identify such things as stereotypes, unfamiliar content, cultural barriers and complexities of language. This is typically referred to as a “sensitivity review” (Camilli, 2006; Penfield & Camilli, 2007).

The other method of detecting bias is an empirical review to statistically determine whether individuals perform differently, using techniques such as differential item functioning (DIF) analysis. DIF analysis has been used to determine if items are operating differently for subgroups of the population. DIF is present when persons in two separate groups have the same ability or total test score (Holland & Wainer, 1993), but have a different probability of correctly responding to a particular test question. Once DIF has been identified, representatives of the group should review the items exhibiting DIF to revise, remove or approve the item in question.

Supporters of standardized-test usage argue that externally developed and scored tests are more objective and hence safer for making decisions because local judgment is avoided (Wiggins, 1993). This type of test may remove one source of bias (e.g., prejudice), but may open itself to other sources of bias. Sources of bias in the assessment of ELLs include language, culture, background knowledge and SES factors.
Perhaps the most significant source of test bias affecting the ELLs is language. When students are expected to perform on a timed test and answer questions in a language in which they are not proficient, this creates an automatic, systematic disadvantage. ELLs struggle with the meaning and concepts of individual words; they have trouble putting the words together in order to extract a sentence’s complete meaning (TESOL, 2002). For example, a student may know the meaning of the word “inch”, as in “How many inches is your desk?” However, when “inch” is used in the phrase “the worm inched forward”, an ELL student may struggle with the meaning that is clear for a native English speaker, who is familiar with the American culture and its locutions.

*Validity and Bias with ELLs*

Critics have long argued that standardized tests are developed for and standardized against the performance of the white, middle-to-upper class groups and not minority groups (Darling-Hammond, 2000; Ladson-Billings, 1997). A number of researchers have pointed to the inherent flaws underlying the procedures and instruments used in the standardized testing of minority students and for their placement in special education (e.g., Harry & Klingner, 2006; Jitendra & Rohena-Diaz, 1996; Oller & Eilers, 2002; Ortiz & Garcia, 1995). According to Gonzalez, Brusca-Vega and Yawkey (1997), some of the problems include “a lack of appropriate norms, use of verbal standardized texts that are not valid and reliable for minority students” (p. 66).

Whenever a test is used as part of the data collection process, the concerns of validity, reliability and bias of that test are important. Just as we would not use a math test to assess verbal skills, we would not want to use a verbal skills test to assess
Accommodations

Accommodations are “supports provided to students for a given testing event either through modification of the test itself or through modifications of the testing procedure to help students access the content in English and better demonstrate what they know” (Butler & Stevens, 1997, p.5). Accommodations for ELLs are designed to reduce the language demand and can include: 1) extra testing time, 2) small group administration, 3) individual administration, 4) testing in a separate location, 5) breaks during testing, 6) linguistically modified version of the test, 7) use of a glossary, 8) bilingual version of the test, and 9) translation of the test into the student’s native language (Abedi, 2006).

Assessments that employ some of these accommodations are thought to go a long way in lessening test-irrelevant linguistic demands on ELL students, hence, providing a better picture of student performance (Abedi, 2006). Test accommodations generally fall into two categories: modifications to the test itself and modifications to the testing procedure. An example of the former is a linguistically simplified version of the test and an example of the latter is allowing students more test time.

Accommodations Can Provide Benefit to ELLs

In a study looking at the needs of Latino English language learners, Moschkovich (1999b) disputed the myth that ELLs cannot participate in mathematical discussions because they are just learning English. Her study found that ELLs are able to participate in discussions, even though they are learning English. Moschkovich’s results suggest that
instruction and assessment need not emphasize low-level language skills over opportunities to engage in discussion of mathematical ideas. Accommodations can mitigate the linguistic demand that ELLs face.

Abedi and his collaborators (Abedi & Lord, 2001; Abedi, Lord & Plummer, 1997; Abedi, Lord, Hofstetter & Baker, 2000) have found that ELLs have difficulty with linguistically complex test items and that reducing linguistic complexity of test items narrows the performance gap between ELL and non-ELL students in mathematics and science. The studies found a large performance gap between ELLs and non-ELLs in reading and writing, the areas that have a substantial amount of language demand. The gap was less in science and even less in mathematics problem solving, where the test items were less linguistically challenging. For mathematics computation, the performance gap virtually disappeared. These findings suggest that, by reducing the language demand on content-based assessments, the validity of assessments can be improved and should result in more equitable assessments for all students. Language demands can be greatly reduced by the use of accommodations during assessments (Archibald & Newmann, 1988; Butler & Stevens, 1997; Abedi, 2002).

Not All Accommodations Provide Benefit to All ELLs

An accommodation that is appropriate for ELLs at one level of English proficiency may not be appropriate or necessary at another level. For example, a bilingual test version may be more helpful to a younger, lower level ELL for whom English is the strange language. For another ELL of higher functioning in English, bilingual tests must only confuse them. Not all accommodations are beneficial to all ELLs and some offer no significant benefit at all.
Abedi et al. (1998) found that ELLs’ mathematics scores increased slightly when given accommodations; however, the gains were not statistically significant. Moreover, native English speakers also had increased scores when given the same accommodations. Johnson and Monroe (2004) examined the impact of simplified language on a state mathematics assessment. A total of 1,232 seventh graders participated. Two forms of a 20-item math test were constructed using released items from Washington state’s education website. The test consisted of 16 multiple-choice and four short-answer items. The primary modifications were shortening of test items, using the active voice and using high-frequency words. Students in general education and ELLs did not benefit from these accommodations. Their findings are consistent with studies conducted in other states on the use of simplified language (Rivera & Stansfield, 2001; Tindal et al., 2000).

Even when students receive the accommodation of their choice, Castellon-Wellington (1999) found that student performance was not significantly improved. Students chose whether to have extra time or to have someone else read the test items and instructions aloud. Aguirre-Munoz (2000) found that ELLs with the lowest English proficiency level benefited most from the Spanish-only accommodation: students with intermediate English proficiency benefited more from the modified English version. This suggests that different accommodations work based on level of student English proficiency.

Abedi et al. (2000b) used three types of accommodations – English dictionary, bilingual dictionary and a linguistically modified test – with NAEP items at the fourth grade. The results showed varied differences across the accommodations. There was a significant difference between the accommodated and unaccommodated conditions for
English dictionary and the bilingual dictionary. However, there was no significant difference in performance between the group that received the original items and the one that received the linguistically modified version.

A study conducted by Ockey (2007) suggested that, even though non-ELLs outperformed ELLs on the NAEP, the score difference could not be attributed to DIF against ELLs. Only one out of ten NAEP items and none out of ten linguistically-simplified items were found to exhibit DIF against ELLs.

Shaftel et al. (2006) examined the impact of language characteristics in items from the Kansas general mathematics assessment, a traditional multiple-choice test of approximately 200 items. Relationships were examined for test items and students at 4th, 7th and 10th grades. Language characteristics had moderate effects on item difficulty at the 4th grade, dropping to small-to-medium effects at the 10th grade. ELLs were not disproportionately affected by language; and difficult mathematics vocabulary had a consistent effect on performance for all students, at all grade levels.

**Conclusion**

The results from these studies suggest that other options should be explored. Accommodations seem to have been implemented with good intentions but with little empirical base to demonstrate whether using them makes any significant difference for the ELL (Butler & Stevens, 1997a; Spicuzza et al., 1996).

In summary, the limited empirical research on the uses and effects of accommodation in large-scale assessment of ELLs (Lee, 1999) suggests that accommodations may fail to correct an unfair disadvantage; or in some cases, they may overcompensate to give unfair advantage. In either case, accommodations can reduce the
validity of assessment results (Lee, 1999). In addition, different students at different levels of English proficiency have different needs (Butler & Stevens, 2001). Although some types of accommodations show promise for some groups of ELLs (e.g., bilingual tests for young, low-level ELLs), the use of the same test accommodations with all ELLs is not likely to be an effective solution.

Research Questions

Against this backdrop, I investigated the relationship of third-graders’ English language proficiency levels to two forms of assessment in the content area of measurement and looked at how some of the accommodations recommended by Abedi (2002) for ELLs are instantiated within a performance-based assessment setting.

1) How well do two different kinds of assessments (standardized versus application-oriented performance-based interview) tap into the same construct of measurement?

2) What evidence is there that language-based accommodations are used in performance based interviews?

3) Is there evidence of language-based bias in one form of assessment over the other?
Chapter 3: Methodology

The purpose of this study was to provide an empirical test of the widely-held belief that performance-based assessment provides a fairer picture of English-language learners’ mathematical skills and knowledge than does a traditional standardized assessment. Specifically, I compared the performance of 94 third-graders on the measurement subscale of the FCAT mathematics test to their performance on a set of measurement content knowledge, reasoning and applications that was drawn from individual interviews.

This study employed a single-subject repeated-measures design, where one group of students was administered two types of assessments on measurement. The first assessment was performance-based and the second, conducted several weeks later, was the traditional standardized state test known as the FCAT. Since the purpose of this study is to contrast two types of assessments, using a single-subject design can help to mitigate influences that can arise from different background environments, different school environments, and different intellectual and/or subject knowledge abilities. Unfortunately, we cannot rule out possible differences based on standardized test anxiety due to pressures tied to Florida’s high-stakes testing system and/or based on order of presentation, since in all cases, the performance assessment was administered first.

Participants

The research was conducted in a large urban school district in the southeast U.S. with a linguistically and culturally diverse student population. The ethnic makeup of the student population in the school district was 60% Hispanic, 28% Black (including Haitian...
and other Caribbean Islanders), 10% White Non-Hispanic, and 2% Asian or Native American. Across the school district, 72% of elementary students participated in free or reduced price lunch programs, and 24% were designated as limited English proficient (LEP) which is the term used by the state for ELLs in ESOL programs.

The research team interviewed 106 third grade students over three years. These students came from nine elementary schools, which were selected based on three criteria: (a) percentage of ELLs (predominantly Spanish or Haitian Creole speaking) above the district average at the elementary school level, (b) percentage of students on free or reduced price lunch programs above the district average at the elementary school level, and (c) school grades predominantly of “C” or “D” according to the state’s accountability plan. Students in this study were selected to be more-or-less equally distributed across gender, ethnicity and ESOL level. A larger proportion of students were in ESOL levels 3, 4 and 5 than levels 1 and 2. Tables of the descriptive statistics are given below.

Table 3.1

<table>
<thead>
<tr>
<th>Student Gender</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>53</td>
<td>50.0</td>
<td>50.0</td>
</tr>
<tr>
<td>Female</td>
<td>53</td>
<td>50.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>106</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2

<table>
<thead>
<tr>
<th>Student Ethnicity</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>White, Non-Hispanic</td>
<td>2</td>
<td>1.9</td>
<td>1.9</td>
</tr>
<tr>
<td>Hispanic</td>
<td>54</td>
<td>50.9</td>
<td>52.8</td>
</tr>
<tr>
<td>Black, Non-Hispanic</td>
<td>50</td>
<td>47.2</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>106</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.3

*Student ESOL Level*

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESOL Level 1</td>
<td>8</td>
<td>7.5</td>
<td>7.5</td>
</tr>
<tr>
<td>ESOL Level 2</td>
<td>12</td>
<td>11.3</td>
<td>18.9</td>
</tr>
<tr>
<td>ESOL Level 3</td>
<td>17</td>
<td>16.0</td>
<td>34.9</td>
</tr>
<tr>
<td>ESOL Level 4</td>
<td>13</td>
<td>12.3</td>
<td>47.2</td>
</tr>
<tr>
<td>Exit ESOL in last two years</td>
<td>33</td>
<td>31.1</td>
<td>78.3</td>
</tr>
<tr>
<td>Never in ESOL or exited over two years ago</td>
<td>23</td>
<td>21.7</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>106</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Due to the relatively low number of students in ESOL levels 1, 2, 3 and 4, I grouped ESOL levels 1 and 2 together, 3 and 4 together. With that adjustment, the revised frequency table is given below.

Table 3.4

*Student Combined ESOL Level*

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESOL Level 1 &amp; 2</td>
<td>20</td>
<td>18.8</td>
<td>18.8</td>
</tr>
<tr>
<td>ESOL Level 3 &amp; 4</td>
<td>30</td>
<td>28.3</td>
<td>47.1</td>
</tr>
<tr>
<td>Exit ESOL in last two years (Level 5)</td>
<td>33</td>
<td>31.1</td>
<td>78.2</td>
</tr>
<tr>
<td>Never in ESOL or exited over two years ago (Level 6)</td>
<td>23</td>
<td>21.7</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>106</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

The focus of this study was not about the income levels of the families. However, the majority of the 106 students were from low SES background, as determined by their
receiving free or reduced-price lunch. Therefore, SES was not included as a demographic variable since almost all students selected were considered to be from fairly homogeneous SES backgrounds.

**Intervention**

The students attend class in schools that are participating in the larger National Science Foundation funded science research project (*P-SELL*) that promotes science to ELLs. Further information can be found at http://www.education.miami.edu/psell/Pages/description.htm. The schools have committed to a five-year professional development intervention project that uses a reform-based curriculum in lieu of the district’s more traditional science curriculum. The larger intervention was designed to promote science learning, along with English language and mathematics, for culturally and linguistically diverse students in grades three through five. It consisted of project-developed curriculum materials (student booklets and teachers’ guides), kits of supplies required to implement the inquiry-based curriculum, and teacher professional development on a range of issues relevant to science learning and English language development with ELLs. For a comprehensive description of the project, see the P-SELL website at http://www.education.miami.edu/psell/Pages/description.htm.

The measurement unit, taught in third grade, consisted of lessons on length, weight, volume, temperature, and time. The unit included activities to promote understanding of each measurement concept and practice collecting measurement data using tools for each type of measurement (ruler and meter/yard stick; kitchen scale, bathroom scale, pan balance, and triple beam balance; graduated cylinder and measuring
cup; thermometer; and clock). Students receive measurement instruction in two subject areas, mathematics and science. Mathematics at these schools is taught using a traditional textbook. The measurement unit in both subjects is taught before the statewide-standardized test is taken in March.

**Instruments**

The state-wide standardized written assessment is taken during March each year. The students’ results from the measurement strand from the standardized test were provided by the Mathematics and Science division of the public schools district office. Students are expected to select and use the appropriate units and to calculate and compare these measurable characteristics (FCAT Mathematics Test Item Specifications, 2005). There is a maximum possible score of eight points. The questions from this strand are all binary. The overall test has a reliability of 0.89, using Crönbach’s alpha as the measure of internal consistency reliability (FCAT, 2007). There is no report on the reliability of the measurement strand from FCAT. A sample test with solutions can be found at the FCAT website at http://fcat.fldoe.org/. The measurement strand on which we report is done by criterion-referenced scoring. The general and overall reporting of the FCAT scores are reported by norm-referenced scores, in addition to criterion-referenced scoring as outlined at http://fcat.fldoe.org/.

The performance-based assessment (see Appendix E) was designed to assess student reasoning and understanding of measurement during a non-threatening, one-on-one student interview, similar to what is recommended by Butler and Stevens (1997) and Abedi (2002) (see Appendix A). It was developed by members of the P-SELL project. The interview protocol is given in Appendix B. This assessment is a re-analysis of the P-
SELL reasoning data led by Buxton et al. (2006) and was reported by criterion-referenced scoring. This instrument has a reliability of 0.765, using Crönbach’s alpha as a measure of internal consistency reliability. The assessment was scored by a team of four persons (Buxton, 2006). The team was trained over a period of several weeks to provide understanding and consistency of the scoring process. To ensure inter-rater reliability, samples of the students’ work were given to another rater to re-score. When there was disagreement, the team met to reach a consensus on the score.

While both measurement assessments covered the topics of length, weight, area, volume and temperature, the standardized test also covers time and perimeter of regular and irregular figures.

*Data Collection*

Third-grade teachers in this study were assigned to one of 12 (6 ESOL-levels by 2 genders) conditions. In a few cases, all of a schools’ ESOL level 1 or level 1 and 2 students had been assigned to one or two teachers who specialized in the teaching of ESOL students. In that case, those teachers were assigned to that ESOL level. Within this constraint of teacher-student assignments, we otherwise randomly assigned teacher to ESOL x gender condition. Parental permission letters were sent home for all students in the block to which the teacher was assigned. If more than one signed permission letter was returned, we randomly selected the student who would participate in this study, with one exception. We assessed all ESOL level 1 students for whom we obtained parental permission. In years two and three, we assessed two students from the ESOL level 1 & 2 sub-group because we did not obtain any students from the sub-group in year one. Over a three-year period, this yielded a sample of 106 students participating in the study.
We collapsed our initial ESOL groups into four ESOL groups: (a) ESOL levels 1 and 2 (early emergent speakers of English), (b) ESOL levels 3 and 4 (students with medium level of English proficiency), (c) ESOL level 5 (students with higher level of English proficiency who exited ESOL programs within the past two years), and (d) non-ESOL students (those who had never been in ESOL programs or had exited from ESOL programs over two years ago). For the purposes of analysis, this last group was assigned an ESOL level 6. Within each ESOL group, male and female students were fairly equally distributed.

According to the state of Florida’s regulations, the designation of ESOL category is a factor of the student’s home language, how the school identifies them and the result from an English proficiency test. See http://www.flaglerschools.com/media/documents/9c3c899a-a74c-4db4-854f-4b44d6fb3f65.pdf. Students may be mis-categorized for any number of reasons. Moreover, a student who has exited the designation of ESOL may still be struggling with language concerns, especially academic language. With this in mind, all the students who exited ESOL within the last two years were counted as ESOL students (ESOL level 5).

Each interview took place in a secure, separate room, usually in the media centre. Each interviewer talked with the student, explaining the process before the interview began, to make the student more comfortable with the process. Students were told that the interview was not a test and it did not count toward the student’s grade in school reporting.

Interviews were typically conducted toward the end of the fall semester; a reasonable time after teachers had completed teaching the Measurement unit from the P-
SELL Curriculum. We completed the interviews over a four-week period. We chose interviews because it is through watching the students do the task and our discussions with them that we get some idea as to what they were thinking. In addition, videotaping allowed us to review and re-analyze the student’s performance and to observe any accommodations that the interviewer might have made during the interview.

During the interviews the students were asked to participate in elicitation tasks on measurement, covering the topics of length, weight, volume and time, from a home and school context. Each interview lasted approximately thirty to forty-five minutes. Any length of time beyond that could have potentially exhausted the interviewee and make continuing fruitless. We allowed students as much time as they needed before moving on to the next question. The interviews were videotaped to allow for in-depth qualitative analysis. The interviews were conducted in a manner that was comfortable to each student.

**Data Analysis**

The data were analyzed using both qualitative and quantitative methods. In the first step of the analysis, a semantic map of verbal and procedural responses and actions was constructed for each of the 106 student interviews. This step created an overall verbal and diagrammatic picture of each interview. Construction of the semantic maps also served to identify three major types of statement categories: (1) unprompted responses, (2) prompted responses, and (3) clarifying questions. Figure 1 of Appendix C shows an example of a semantic response map. A score was assigned for each item to analyze it quantitatively. Concept maps were made using Inspiration® software. On few occasions, when students spoke in their home language, they were encouraged to speak in
English in an effort to ensure comparable data-gathering across all students. The team consisted of persons who spoke Spanish and Haitian-Creole. Therefore, we were able to translate student responses into English where necessary.

The data from the performance-based assessment were coded to determine if the students selected the appropriate tool and conducted the measurements accurately, using the appropriate units. The maximum possible score used in this assessment was 28 points. The scores were disaggregated to show the sub-categories of the ESOL level, gender and ethnicity contexts. The rubric for coding was developed by a team of the P-SELL project. The instrument has a reliability of 0.765 using Crönbach’s alpha as the measure. The rubric for this instrument is included in Appendix E.

To answer research question one, I correlated the two assessments to see if they are tapping into the same constructs of measurement. To answer research question two, I reviewed the video recordings of the interviews to determine if there was evidence of accommodations being employed by the interviewer and whether these accommodations might have affected student performance. Finally, to answer research question three, I conducted two one-way analyses of variances, with ESOL level as the independent variable and assessment source as the dependent variable. I did not use regression or correlation methods because ESOL levels do not form an equal-interval scale. I examined the analysis of variance (ANOVA) with post-hoc analysis to determine if there were any differences across the ESOL levels.
Chapter 4: Results

On the standardized test, scores ranged from the minimum (0 point) to the maximum (8 points). Similarly, on the performance assessment, scores ranged from the minimum (0 points) to the maximum (28 points). Scores by ESOL levels were comparable across the two assessments. Thus, both assessments seemed to be fair and attainable. There were students attaining the maximum score on each of measures. It is interesting to note that in the standardized test, there was an ESOL level 6 student whose score is zero. This might seem surprising considering that this student was classified English-proficient. Alternatively, this might speak to the limited content knowledge that the student possessed or to the concerns that usually accompany standardized tests. At all the lower ESOL categories on either assessment, all students were able to score some of the items correctly. In addition, some students from each ESOL level were able to score the maximum points.

Research Question 1 – Two Assessments, Same Construct?

The standardized test score is positively and significantly correlated with the performance-based assessment, with a correlation coefficient of 0.409 at $p < 0.001$. While on the face of it, both assessments seem to tap into the same construct of measurement, it seems likely, given the amount of unexplained variance of 84%, that they are tapping into different aspects of measurement. In addition, with only eight items, it is possible that the FCAT subscale had lower reliability than the performance assessment.
Research Question 2 – Evidence of Language-based Accommodations

In some cases, the accommodations proved helpful to the student and in other cases, the accommodations provided little or no benefit, particularly at the lower ESOL level students. Below are examples from performance assessment settings where students were provided with real-time language-based accommodations as recommended by the research literature (Abedi, 2002; Butler & Stevens, 1997; LaCelle-Peterson & Rivera, 2004).

Finding 1: Allowing Extra Time and Using Probing Techniques Provide Benefit

The following excerpts are illustrations of how accommodations employed by the interviewer benefited the ELL. In this first example, using accommodations like allowing extra time, rephrasing and explaining the question, speaking slowly and using probing techniques, can allow the interviewer to better understand what the student knows about measuring length. This student is at ESOL level 3&4.

I = interviewer   S = Student

I: Have you ever seen anyone in your family measure how long or how tall something is. (speaks slowly)
S: No. I don’t know. Maybe.
I: It’s OK. Take your time and think about it.
S: Oh, yes, my Dad.
I: What was your dad measuring?
S: He was measuring the computer desk to see if it fit in my bedroom.
I: OK, that’s nice. How did he do that?
S: I don’t know. What you mean?
I: Did he use a tool or something to measure the length of the desk? How did he measure the desk?
S: Oh, yeah. He use something like a ruler, but you pull it out of this round thing. It has numbers on it, just like a ruler.
I: How tall are you?
S: (Student pauses for about 20 seconds.) Tall.
I: OK, but how tall?
S: I don’t know.
I: Have you ever been to the doctor?
S: Yeah.
I: Did anyone at the doctor’s measure how tall you are?
Finding 2: Reading and Repeating Instructions Slowly Provide Benefit

In this second illustration, the interviewer accommodated for the student’s English language deficiencies by reading and repeating the instructions slowly. The student (ESOL level 1&2) was then able to make the appropriate selections. The student’s responses are written phonetically to illustrate how the student struggled with reading.

S: Oh, yes.
I: How did they do it?
S: You had to stand next to the wall with the ruler on it. Then they look at the line and say how tall you are and write it down.

Finding 3: Limiting ESOL 1&2 to English May Depress Their Performance on Authentic Assessments

To illustrate some of the linguistic challenges that ELLs face in an all-English setting, below is an excerpt of an interview of an ESOL level 1&2 student. It seems evident that the student’s limited command of the English language is a barrier to understanding his/her knowledge.
I = interviewer       S = Student

I: Have you ever seen anyone in your family measure how long or how tall something is. (speaks slowly)
S: Yes.
I: What were they measuring?
S: (Student pauses for about 20 seconds.) Measuring cup
I: Measuring cup. How tall are you?
S: Tall?
I: Yes how tall (interviewer gestures using hands) are you?
S: Eight.
I: OK. Have you ever seen anyone weigh anything to see how heavy it is?
S: No.
I: Who does most of the cooking in your family?
S: My daddy?
I: OK. Have you ever seen your daddy weigh anything while he was cooking?
S: Yes.
I: What was he weighing?
S: (long pause) The kitchen.
I: OK. Do you ever go to the food store or the market?
S: No.
I: You don’t go to the food store?
S: Oh, yes!
I: Have you ever seen anyone in the food store weigh anything?
S: Yes. Sandwich, chicken, coke.

Despite the fact that the interviewer spoke slowly or used probing questions that might have prompted the student responses, her English was very limited. It appears as if she recognized some words like “food” but mistakenly connected that word with “fast food”. We see this evidenced when she spoke of sandwich, chicken and coke. Clearly, since these are not items that are weighed in the food store, this may be viewed as a lack of comprehension.

**Finding 4: ESOL 3&4’s Struggle with Vocabulary**

In another interview, the student (ESOL 3&4) does not have recall of the vocabulary, but he is able to describe the tool that is used for measuring. Even though the student is able to correctly describe the tool, in the
performance assessment scoring, no points were awarded. In a traditional multiple-choice test, the student may have recognized the word from the test itself and selected the correct answer. An illustration from the student interview is given below.

I: Have you ever had a fever? Have you ever been hot?
S: Yes.
I: Anyone ever take your temperature?
S: No.
I: Ever been to the doctor?
S: Oh yeah, only at the doctor.
I: How did they do that? How did they take your temperature?
S: They use that thing that you put in your mouth and it shows how much your temperature is.
I: Do you know what that thing is called?
S: No.

Research Question 3 – Language-based Bias in Assessments?

The ANOVA results found no significant difference in scores across ESOL levels in the standardized test. Again, this may be due to the reliability of the FCAT. However, there were significant differences across ESOL levels in the performance-based assessment. The ANOVA summary tables are given below.

Table 4.1

FCAT (Standardized Test) Measurement Results

<table>
<thead>
<tr>
<th>Combined ESOL levels</th>
<th>N</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESOL levels 1 &amp; 2</td>
<td>18</td>
<td>4.94</td>
<td>2.01</td>
</tr>
<tr>
<td>ESOL levels 3 &amp; 4</td>
<td>25</td>
<td>5.40</td>
<td>1.47</td>
</tr>
<tr>
<td>ESOL level 5</td>
<td>29</td>
<td>5.62</td>
<td>1.72</td>
</tr>
<tr>
<td>ESOL level 6</td>
<td>22</td>
<td>6.09</td>
<td>1.82</td>
</tr>
<tr>
<td>Total</td>
<td>94</td>
<td>5.54</td>
<td>1.76</td>
</tr>
</tbody>
</table>
ANOVA Results

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>13.740</td>
<td>3</td>
<td>4.58</td>
<td>1.507</td>
<td>.218</td>
</tr>
<tr>
<td>Within Groups</td>
<td>273.590</td>
<td>90</td>
<td>3.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>287.330</td>
<td>93</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2

Performance-based Assessment Results

Descriptives

<table>
<thead>
<tr>
<th>Combined ESOL levels</th>
<th>N</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESOL levels 1 &amp; 2</td>
<td>18</td>
<td>16.17</td>
<td>5.80</td>
</tr>
<tr>
<td>ESOL levels 3 &amp; 4</td>
<td>25</td>
<td>19.88</td>
<td>4.73</td>
</tr>
<tr>
<td>ESOL level 5</td>
<td>29</td>
<td>21.21</td>
<td>3.22</td>
</tr>
<tr>
<td>ESOL level 6</td>
<td>22</td>
<td>24.00</td>
<td>3.43</td>
</tr>
<tr>
<td>Total</td>
<td>94</td>
<td>20.54</td>
<td>4.95</td>
</tr>
</tbody>
</table>

ANOVA Results

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>631.431</td>
<td>3</td>
<td>210.477</td>
<td>11.481</td>
<td>.000</td>
</tr>
<tr>
<td>Within Groups</td>
<td>1649.899</td>
<td>90</td>
<td>18.332</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2281.330</td>
<td>93</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In both assessments, we see a steady improvement in the average achievement as the ESOL levels increase. Also note that larger variance occurs with the students at the ESOL level of 1&2. The post-hoc analyses found that there are significance differences between ESOL level 1&2 and all the other ESOL levels; and between ESOL level 3&4 and ESOL level 6. See Table 4.3 below.
Table 4.3

Multiple Comparisons

Dependent Variable: Student's TOTAL Measurement Score

Tukey HSD

<table>
<thead>
<tr>
<th>Combined ESOL levels</th>
<th>Combined ESOL levels</th>
<th>Mean Difference</th>
<th>Std. Error</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESOL levels 1 &amp; 2</td>
<td>ESOL levels 3 &amp; 4</td>
<td>-3.71(*)</td>
<td>1.324</td>
<td>.031</td>
</tr>
<tr>
<td>ESOL level 5</td>
<td>ESOL level 6</td>
<td>-7.83(*)</td>
<td>1.361</td>
<td>.000</td>
</tr>
<tr>
<td>ESOL levels 3 &amp; 4</td>
<td>ESOL levels 1 &amp; 2</td>
<td>3.71(*)</td>
<td>1.324</td>
<td>.031</td>
</tr>
<tr>
<td></td>
<td>ESOL level 5</td>
<td>-1.33</td>
<td>1.169</td>
<td>.669</td>
</tr>
<tr>
<td></td>
<td>ESOL level 6</td>
<td>-4.12(*)</td>
<td>1.252</td>
<td>.008</td>
</tr>
<tr>
<td>ESOL level 5</td>
<td>ESOL levels 1 &amp; 2</td>
<td>5.04(*)</td>
<td>1.285</td>
<td>.001</td>
</tr>
<tr>
<td></td>
<td>ESOL levels 3 &amp; 4</td>
<td>1.33</td>
<td>1.169</td>
<td>.669</td>
</tr>
<tr>
<td></td>
<td>ESOL level 6</td>
<td>-2.79</td>
<td>1.211</td>
<td>.104</td>
</tr>
<tr>
<td>ESOL level 6</td>
<td>ESOL levels 1 &amp; 2</td>
<td>7.83(*)</td>
<td>1.361</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>ESOL levels 3 &amp; 4</td>
<td>4.12(*)</td>
<td>1.252</td>
<td>.008</td>
</tr>
<tr>
<td></td>
<td>ESOL level 5</td>
<td>2.79</td>
<td>1.211</td>
<td>.104</td>
</tr>
</tbody>
</table>

Based on observed means.
* The mean difference is significant at the .05 level.
Chapter 5: Conclusions, Discussion, Implications

Conclusions/Discussion

Research Question 1—Two Assessments, Same Construct

While there was positive correlation between the two assessments (0.409), there is a large unexplained variance of 84%. If both assessments are assessing the same construct of measurement, they are tapping into different aspects of it. A stronger correlation would have more closely matched the two assessments. This result is not surprising in light of the fact that the performance-based assessment was not designed to match the FCAT test in the content area of measurement, but as a measure of ELLs’ reasoning skills on the topic of measurement.

The performance assessment is based on the interview designed to assess student reasoning and understanding. The low correlation to FCAT raises questions about the utility of FCAT to provide information related to ELLs’ knowledge and reasoning/understanding of the measurement. The average correlation between scores from the FCAT and SAT-9, as per the Stanford Achievement Test and http://harcourttassessment.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=E132C, was very high at 0.96 (Greene & Winters, 2003). Since both tests are highly correlated and the SAT-9 makes no claim about alignment to instruction, one might wonder about the alignment of FCAT to instruction. In addition, while the overall reliability for the FCAT is 0.89, the measurement strand of eight items does not likely hold the same level of reliability. Since we do not know the reliability of the measurement subscale, we cannot rule out the possibility that the low correlation is due to less than optimal subscale reliability.
Research Question 2– Evidence of Language-based Accommodations

Several accommodations were made during the performance-based interviews. These included 1) allowing extra time, 2) rephrasing and restating the question or instructions, 3) speaking slowly, 4) articulating carefully and 5) assessing occurring in a quiet private room. The accommodations used influenced the performance of students differently, according to the various ESOL levels.

The results show that firstly, while ESOL level 3&4 students may struggle with recalling the vocabulary, he/she is able to describe the measuring tools used and explain the process. Secondly, allowing extra time and using probing techniques provide benefit for ESOL level 3&4 students, enabling them to recall prior knowledge, which might have otherwise eluded them. While these accommodations may have provided benefit to students at ESOL levels 3&4 and higher, they provided little or no benefit to ESOL level 1&2. It is at this level that students need language assistance, either through translations or the use of a bilingual dictionary. This is revealed by the third finding that shows that limiting ESOL level 1&2 students to communicating in English may depress their performance on authentic assessments, since these students are early emergent speakers of English. The last finding is that reading and repeating instructions slowly provide benefit to the ESOL level 1&2 student, since it removes the written component of the language.

Research Question 3– Language-based Bias in Assessments?

English proficiency level failed to predict FCAT performance. However, it did predict performance on the reasoning and applications tasks. The results show that the
standardized assessment results yield no significant differences across ESOL levels, yet there are significant differences across ESOL levels in the performance-based assessment. In this investigation, the traditional assessment appears to be a better indicator of achievement for ESOL students than the performance assessment does. These findings present a challenge to the conventional wisdom that performance-based assessment provides a less-biased picture of ELL’s mathematical knowledge than does standardized testing. One may have expected that the accommodations made by the interviewer during performance-based assessments would have mitigated the bias due to language. However, there was little evidence of that. This finding is important since the validity of an accommodation is based on its differential functioning for the target population, ELLs in this case.

There are several possible reasons for these findings. The standardized written test, FCAT, provides for ELLs additional resources, if they elect to use them. These resources include extra time and a bilingual dictionary. If the students who were tested utilized any of these aids, then this might account for the lack of differences across ESOL levels in the FCAT. In addition, the format of the assessments was different. The traditional standardized test is a multiple-choice test and students are drilled on test-taking skills. Hence, students’ performance on the FCAT could be misleading. Moreover, due to the increased pressure that school districts face, from initiatives such as NCLB, there may have been other accommodations built into the standardized tests over the years. These could include simplified language, terminology and idiomatic expressions that minority students would be better able to relate to and comprehend. The
performance-based interview, on the other hand, was scored conservatively as evidenced by lack of credit for students not knowing vocabulary.

Post-hoc analyses showed that there were significant differences with ESOL level 1&2 and all other levels on the performance-based assessment. These findings are consistent with research that has shown that less English-proficient students perform significantly lower than the other levels (Abedi & Lord, 2001; Aguirre-Munoz, 2000; Rivera & Stansfield, 2001; Lee & Luykx, 2006). This result suggests that students with low English-proficiency skills need something other than what was provided to mitigate the linguistic demand. These students are early emergent speakers of English and perhaps they could have benefited more if they had been assessed bilingually or had access to another translation facility. It is important to remember that students were pressed to speak English, even if they were seen to have difficulty communicating in English.

ELL students showed more competencies in measurement than for which they get credit. This is particularly true when they are assessed in ways that afford them opportunities to demonstrate their understanding in an environment that is not so language dependent. The ESOL students at the lowest levels performed significantly lower than their higher-level peers. Language plays a critical role when using performance assessments. Depending on the type of accommodations used and the student’s ESOL level, accommodations do provide benefit. ELLs at the lower ESOL levels seem to benefit least from accommodations that are not directly related to language, such as translations.

A possible explanation here could be that the assessment is given in English only. Students with linguistic challenges have difficulty understanding and/or expressing
themselves in English, as they would wish. This was seen in the illustrations that are included. Perhaps, if the assessment were conducted in the student’s home language, in addition to English, there might not have been any significant difference with ESOL level 1&2 and the other levels.

From the illustration provided of the student of lowest level of ESOL, it seemed that if the student were allowed to use his/her home language, then he/she might have been better able to express his/herself. On occasion, we did have students speak in their home language and we encouraged them to take their time and speak in English. In actuality, this contravenes Abedi’s (2002) recommendation to allow the student to speak in the home language if needed. This is an indication that the lower ESOL level student could have used some language assistance. In the original study, Buxton et al. (in press) opted not to employ a bilingual assessment to ensure consistency in the level of language of the assessment and to achieve a higher reliability. This coupled with conservative scoring proved to be of little aid to the student who is new to learning in English.

**Implications**

The selection of students was nested within teacher across six schools that were pilot-testing a standards-based science curriculum in grade three and later grades four and five. Replication, that selects students randomly within a larger sample of ELL students, is necessary to test the generalizability of these results. However, within this environment, the standardized test showed less bias due to language relative to a performance assessment. We are uncertain whether students who do not learn from an inquiry-based P-SELL curriculum would evidence similar results.
Despite the continued push for more authentic assessments, the argument for traditional standardized tests remains strong. Traditional tests have undergone changes since the 1980’s at the start of the performance assessment movement. In many ways, they are not so traditional anymore. Most of the research advocating against the traditional standardized test is at least ten years old. Perhaps pressure from the initiatives coming out of the education reform has led to changes in the traditional tests that may not yet have received their due recognition. These tests seem to be more accommodating than they previously were. While some may still argue that these traditional standardized tests are designed for the white mainstream middle class student population, perhaps a shift is slowly occurring and assessments are becoming more carefully constructed as to be more equitable to all students. This trend ought to be expected when one considers that the minority population, particularly in the southern United States, is rapidly growing.

Despite these trends, there still appears to be a mismatch between the objectives of the standardized test and the purpose for which it is used. One of the published objectives of the FCAT is that it is used to measure student understanding of the subject. In light of the low correlation between FCAT and a test that was designed to assess quality of student reasoning, it remains an open question as to how well standardized tests meet, or if it can meet this goal. On the other hand, while performance test does hold an advantage for the ELL, finding evidence of validity will always remain a challenge, since not all students can be treated equally in these circumstances.

Any assessment is in part an assessment of language in addition to the subject matter. Language bias cannot be removed totally. It is clear that if students are deficient in language, then a mode of assessment that significantly removes the language bias will
go a long way in improving their understanding and performance. The type of teaching and assessment on a daily basis play a major role in students’ understanding.

While some see authentic and performance assessments as a partial solution to this challenge, Darling-Hammond (2000) points out that performance assessment is not necessarily more equitable than traditional standardized tests. There is a need for policy changes to build a more equitable system. For all the promise of more authentic and performance-based forms of assessment, their value depends as much on how they are used and what supports for learning accompany them. Changing assessment forms and formats without changing the ways in which assessments are used will not change the outcomes of education. In order for assessment to support student learning, it must include teachers at all stages and be embedded in the curriculum and teaching activities. It must be aimed at supporting student-centered teaching rather than at sorting students and sanctioning schools. It must allow for different starting points of learning and must allow for diverse ways of demonstrating competence.
References


TESOL (Teachers of English to Speakers of Other Languages) 2002: *ESL standards for pre-K-12 students*. Alexandria, VA: TESOL.


Western Australia Department of Education (2000). *Student achievement in mathematics*. East Perth, WA.


Introduction

_Preamble_

Standardized tests are a common part of the education system. However, some aspects of standardized testing make these tests unfair to certain groups, including the English language learner (ELL). Since their inclusion in large-scale, high-stakes assessments, ELLs have performed below average in mathematics and measurement, in particular (NAEP, 2003, 2007). Despite the fact that they are capable of high levels of conceptual understanding, this is not evidenced by the results of standardized test scores. One possible explanation could be that traditional means of assessment may not typically reflect their understanding of content knowledge (Solano-Flores & Trumbull, 2003; Abedi, 2004; Lenski, 2006). While ELL students may have difficulty writing in English, they can speak about mathematics and science with a level of sophistication not reflected on written assessments alone (Lenski, et al., 2006; Cox-Peterson & Olson, 2007). This leads one to believe that there is a mismatch between assessment practices and student understanding.

The topic of measurement was selected for several reasons. Firstly, because it is one of the more widely used applications of mathematics and science, we could assume that most students, including ELLs, have had some exposure to it, as it is a part of daily activity for most people. Secondly, it is one of the foundational concepts that does not require much prior mathematics knowledge, as may be the case with other topics. Thirdly, it is recognized as important enough to be included as one of the content standards in NCTM’s _Principles and Standards for School Mathematics_ (NCTM, 2000).
This review examines the body of knowledge on the impact of different types of assessments of ELLs and the topic of measurement. Some questions for consideration are: 1) Are standardized, high-stakes tests a valid measure of what ELLs know and can do? 2) What impact do accommodations have on student achievement of ELLs? 3) How do students learn measurement?

*Why is it Important?*

The population of English language learners (ELLs) in the US has seen steady growth. Between 1990 and 1997, the number of immigrant US residents increased by 30% (U.S. Census Bureau, 2000). As the ELL population has grown, so has the need to include these students in large-scale testing. Improving America’s Schools Act (IASA) of 1994 and the “No Child Left Behind” Act (NCLB) of 2002 mandated the inclusion of English language learners in large-scale assessments. The exclusion of certain groups from large-scale assessments may have been well intentioned, but arguably, it has resulted in diminished educational opportunities for ELLs. Standardized tests are linguistically complex for minorities, including ELLs. This raises concerns of test bias and validity. If students are unable to master the linguistic complexities of an assessment, then results from the assessment would not likely be a true representative of what they know and can do. In addition, interpretations and uses of those results could be flawed. Because accommodations are believed to mitigate the language barriers, performance-based assessments are seen to provide a fairer picture of English-language learners’ mathematical skills and knowledge than do traditional, standardized, written assessments (Archibald & Newmann, 1988; Abedi, 2002; Johnson & Monroe, 2004).
On the topic of measurement, the Third International Mathematics and Science Study (TIMSS, 1996), an international research study on math and science achievement, reported that students in the U.S. scored very low when compared with students in other countries. Results are similar from the study conducted by the National Assessment of Education Progress (NAEP, 2000). Some states that do statewide testing, such as Florida, are also finding low scores in this area (FCAT, 2007).

Driven by the reform efforts, researchers and educators are seeking ways to determine how best to attend to the needs of ELLs, in an effort to provide equitable education for all students.

**Definitions of Terms**

In the context of school, ELLs are students whose home language is not English and who attend a school where English is the language of instruction. These students are placed in regular classrooms for the majority of their subjects, but go to a language specialist who assists in teaching English as a second language. Throughout the history of education many different terms have been used to describe or characterize children whose second language is English. Some examples are students with Limited English Proficiency (LEPs) and students for whom English is a Second Language (ESLs). Currently educators refer to these children as English Language Learners (ELLs). This change in language denotes a more accurate reflection of the process of language acquisition.

Assessment is the process of determining and documenting, usually in measurable terms, knowledge, skills, attitudes and beliefs (Broadfoot, 1996). Assessments can be classified in many different ways; formative, summative, objective, subjective, formal,
standardized, performance-based, authentic, criterion-referenced and norm-referenced. For the scope of this review, the focus will be limited to traditional, standardized, written assessments and performance-based assessments. Traditional assessments are formal, written, standardized tests that have predominantly multiple-choice items. Students are given little or no opportunities to explain their answer or to demonstrate what they know and can do. Performance-based assessment focuses on what students can do. A well-defined task is identified and students are asked to create, produce, or do something, often in settings that involve real-world application of knowledge and skills (Broadfoot, 1996; Abedi, 2002).

A good assessment has both validity and reliability. However, in practice, an assessment is rarely totally valid or totally reliable (Archibald & Newmann, 1988). For example, a ruler which is marked wrong will always give the same (wrong) measurements. In this case, it is very reliable, but not very valid. On the other hand, asking students to tell the time without looking at a clock or watch is an example of an assessment which is valid, but not reliable. In this case, the answers will vary between students, but the average answer is probably close to the actual time. In many fields, such as educational testing, there are often trade-offs between reliability and validity due to the nature, environments and constraints of the discipline. Generally, the more reliable our estimate of what we purport to measure, the less certain we are that we are actually measuring what was intended (Archibald & Newmann, 1988).

Test accommodations are adjustments to a test or the testing procedure to adjust for the special needs of students (Sireci, 2004). In the case of ELLs, it is done to mitigate linguistic complexities and biases.
Measurement has been described as “the assignment of a numerical value to an attribute of an object” (NCTM, 2000). It involves relating numerical quantity with a spatial attribute (Bishop, 1983). It might be a simple task for adults; however, it is not a trivial task for children. It requires steps that involve several mathematics concepts. Measurement consists of identifying a unit of measure, subdividing the object by that unit and placing the unit end-to-end (iterating) along the side of the object being measured (Stephan & Clements, 2003). One of the purposes of measurement is to make comparisons, directly or indirectly, of two or more objects.

*Framework*

The theoretical framework for assessing ELLs is built around the practices and beliefs described in the TESOL *Standards for P-12 ESL Teacher Education Programs*. They are drawn from research on second language acquisition, culture, instruction and assessment. Firstly, language learning implies more than simply mastering a code (TESOL, 2002). It involves the learner using language in ways that are socially and culturally appropriate for the particular setting. Knowledge about what is appropriate is embedded within the values of communities, families and schools, and the ability to use those codes appropriately ties individuals to that community (Heath, 1986; Phillips, 1983). Children learn early what is acceptable language in their homes/families first and then within their community. They learn what they experience and live everyday (Kolb, 1984). Therefore, opportunities to use the language, in assessments, are complicated by the contexts of race, class and language (McKay & Wong, 1996; Peirce, 1995).

The role of culture in language learning also informs the theoretical framework for the teaching of ELLs. Of importance is the finding that language, cultural traditions,
beliefs and educational experiences that ELLs bring to the classroom should not be considered a problem or a deficit but rather a resource that diversifies and enriches the classroom experience (Heath, 1986; Pease-Alvarez & Vasquez, 1994). Too often, schools reflect the cultural knowledge and beliefs of the majority groups and the diversity of culture and language that ELLs bring to the classroom is ignored (McGroarty, 1986). For example, white educators by virtue of their position of power present a primarily Anglo-Americanized curriculum without taking into consideration the opinions of and the contributions that ethnically diverse groups may have made (Delpit, 1998). Working class students are far more likely to fail because their cultural capital is seen as less valuable. Minority children quickly come to realize that they do not speak the same language as the educational system. It offered them very little that is culturally useful and historically, they have tended to “vote with their feet” by leaving as early as possible (Rist, 1970). Arguably, because of major reform efforts like, NCLB and IASA, this situation is changing.

According to Messick (1989), threats to validity of the interpretation of test scores originate from two sources: construct under-representation and construct-irrelevant variance. As he claims, “Tests are imperfect measures of constructs because they either leave out something that should be included… or else include something that should be left out, or both” (Messick, 1989, p. 34). Construct under-representation is where a test measures only a portion of what it was intended to measure. Whereas, construct-irrelevant variance refers to the situation where a test measures skills or abilities irrelevant to what it was intended to measure (Messick, 1989). A relevant example of construct irrelevant variance is when a mathematics test written in English is
administered to an English language learner. In this case, the student’s level of English language proficiency is the construct-irrelevant factor that can confound the scores.

Assessment

Importance of assessment

Educational assessment generally serves three purposes: 1) to show the extent to which teachers and students have met teaching and/or learning objectives, 2) to indicate what might be done to improve teaching and learning, and 3) to determine the most promising students for promotion, college or some recognition (Archibald & Newmann, 1988). It is important to distinguish between these purposes since assessment results used for one purpose may not necessarily serve another. For example, a standardized test like the SAT was designed to provide information about students most likely to succeed in college and not to indicate student achievement. To use the results from this test as an indicator of knowledge acquisition would produce a flawed outcome.

Despite this, the use of test scores to indicate educational success or failure is almost never questioned. From policy makers, to educators, to parents and students, low scores are bad news; high scores are good news. It seems as if improving the education system is synonymous with improving test scores. This heavy reliance on standardized testing as an indicator of student achievement and performance has been heavily criticized. According to Haertel (1999), one identified "past mistake" has been the reliance on standardized multiple-choice tests, and one identified “solution” has been the reliance, instead, on performance assessments. Because of this, performance and/or authentic assessments have been at the centre of state and national education reform initiatives in the 1990s.
The heavy reliance on standardized testing as an indicator of student achievement has impacted schools, teachers, students and parents. Students and teachers experience the pressure of assessments, particularly standardized written tests. Many states also require that students pass the statewide assessment in order for them to graduate (Rivera & Vincent, 1997). In Florida, students are well aware that they may not be promoted to the next grade level if they have not passed the Florida Comprehensive Assessment Test (FCAT). Schools are rated based on their performance on the FCAT. This puts pressure on the teachers to “get good grades”. Teachers feel forced into “teaching to the test” (McNeil, 2000). They may focus primarily on giving students test-taking techniques rather than quality instruction. Students may not be encouraged to explain their reasoning or show alternative ways of working a problem. They focus primarily on selecting the ‘correct’ answer” (McNeil, 2000).

Types of assessments

There are several types of assessments that teachers can employ; 1) formative vs. summative, 2) traditional vs. alternative, 3) standardized vs. classroom, and 4) authentic vs. contrived (McMillan, J. H. 2000). A valid, authentic assessment provides information about the particular tasks on which students succeed or fail. Moreover, these tasks are meaningful and significant. Assessments that provide little substantive information and lack authenticity undermine the legitimacy not only of the numerical indicators, but also of the educational system (Archibald & Newmann, 1988). Traditional, standardized assessments provide uniformity for all test takers with regard to test content, scoring and administration conditions. They also keep the teacher on task, focusing on teaching that which will be assessed. In addition, they are inexpensive to administer. However, they
pay little attention to the depth of understanding and integration of student knowledge (Wiggins, 1993). What do standardized tests really measure? Standardized test scores are commonly viewed as indicators of knowledge, abilities, or achievements and predictors of future achievement. However, there is disagreement about the kinds of achievements measured. Some assume that a student who outperforms his peers on a standardized test knows more. Others claim that no such competencies have been demonstrated. They claim that standardized tests measure no more than the ability to take multiple-choice tests (Wigdor & Garner, 1982).

With performance-based, authentic assessment, students are thought to be afforded the opportunity to demonstrate what they know and can do, and/or are thinking, in a manner that is not so heavily dependent on language skills. One potential benefit of performance assessments is that they allow for real-time use of language-based accommodations, hence providing a more complete curricularly-valued picture of what the students really know and are able to do. Therefore, they are seen as a more appropriate and fair way to measure ELLs’ progress (Lenski, et al., 2006). However, performance assessment is not without its own concerns. How does one ensure the validity and reliability of tasks? Rivera and Collum (2004) contended that, compared to English-proficient students, ELLs must direct more cognitive resources to processing the language of the test and therefore have fewer resources to attend to the content of the test. Accommodations are intended to minimize the cognitive resources that ELLs need to process the language, thereby affording more resources to be directed to the content of the test itself.
In performance assessments, accommodations are generally considered to be “support provided to students for a given testing event either through modification of the test itself or through modifications of the testing procedure to help students access the content in English and better demonstrate what they know” (Butler & Stevens, 1997, p.5). There are several ways in which the language demand of a test can be reduced. The most common accommodations are 1) extra testing time, 2) small group administration, 3) individual administration, 4) testing in a separate location, 5) breaks during testing, 6) linguistically modified version of the test, 7) use of a glossary, 8) bilingual version of the test, and 9) translation of the test into the student’s native language (Abedi, 2006). It is hoped that performance assessments that employ some of these accommodations go a long way in lessening the linguistic demands on the student; therefore, providing a better picture of the student’s understanding (Abedi, 2006).

Another type of alternative assessment is the portfolio, a collection of the student’s work over a period of time. The portfolio is assessed, generally at the end of the specified period, for content, student growth, development and improvement. Students are given feedback throughout the semester and they have the opportunity to revise work based on the feedback received from peers and teachers (Dietel et al., 1991). Traditionally, this was seen as a more appropriate way of assessing the language arts and social sciences and not the natural sciences and mathematics. Preliminary observations of classroom instruction in Kentucky and Vermont, two states with portfolio assessment, indicate that teachers spend more time training students to think critically and solve complex problems than they did previously (Improving America’s School, 1996). A potential disadvantage of continuous portfolio assessment is that it could establish false
expectations for students. Students who are taught and assessed one way during the year and then participate in high-stakes testing could encounter many challenges because formative assessment was not linked with summative assessment (Black & Wiliam, 1998).

**ELLs, Assessment and Accommodations**

Performance-based assessment, with all its concerns, is generally seen as an authentic method of assessing what students know and can do. They are assessment methods that require students to demonstrate what they know and can do by applying specific skills and competencies to a learning task. These tasks are often referred to as authentic assessments because they typically are designed to involve students in responding to a real-life task or problem. Traditional standardized tests are written, timed, usually multiple-choice assessments that typically do not afford students the opportunity to demonstrate their knowledge. They reveal only whether the student can recognize, recall, or “plug in” what was learned out of the context in which it was learned. These conventional tests typically only ask students to select or write correct responses, irrespective of reasons.

When testing ELLs in subject matter areas, construct irrelevant variances could result from the language in which the test is presented. Accommodations are seen as one way to mitigate the language bias of tests for the ELL. Test accommodations generally fall into two categories: modifications to the test and modifications to the testing procedure. An example of the former is a linguistically simplified version of the test and an example of the latter is allowing students longer test time.
Effects and issues of various accommodations

Abedi and his collaborators (Abedi & Lord, 2001; Abedi, Lord & Plummer, 1997; Abedi, Lord, Hofstetter & Baker, 2000) have found that ELLs have difficulty with linguistically complex test items and that reducing linguistic complexity of test items narrows the performance gap between ELL and non-ELL students in mathematics and science. The studies found a large performance gap between ELL and non-ELL students in reading and writing, the areas that have a substantial amount of language demand. The gap was less in science and even less in mathematics problem solving, where the test items were less linguistically challenging. For mathematics computation, the performance gap virtually disappeared. These findings suggest that, by reducing the language demand on content-based assessments, the validity of assessments can be improved and should result in more equitable assessments for all students. Language demands can be greatly reduced by the use of accommodations during assessments (Archibald & Newmann, 1988; Butler & Stevens, 1997; Abedi, 2002).

In addition, an accommodation that is appropriate for ELLs at one level of English proficiency may not be appropriate or necessary at another level. Not all accommodations are beneficial to all ELLs and some accommodations can benefit all groups.

A study by Abedi et al. (1998) showed that although ELLs scores increased slightly when given accommodations, the gains were not statistically significant. Moreover, native English speakers also had increased scores when given the same accommodations. Johnson and Monroe (2004) examined the impact of simplified language on a state mathematics assessment. A total of 1,232 seventh graders
participated. Two forms of a 20-item math test were constructed using released items from Washington state’s education website. The test consisted of sixteen multiple-choice and four short-answer items. The primary modifications were related to shortening the items, using the active voice and using high-frequency words. Results indicate that students in general education and ELLs do not benefit from this accommodation. Their findings are consistent with studies conducted in other states on the use of simplified language (Rivera & Stansfield, 2001; Tindal et al., 2000).

Even when students receive the accommodation of their choice, Castellon-Wellington (1999) found that student performance was not significantly improved. In this, students were given the choice of extra time or reading the test items and instructions aloud. Aguirre-Munoz (2000) found that ELLs with the lowest English proficiency level benefited most from the Spanish-only accommodation, while students with intermediate English proficiency benefited more from the modified English version. This seems to support the idea of different accommodations, based on level of English proficiency.

In their study, Abedi et al. (2000b) used three types of accommodations – English dictionary, bilingual dictionary and a linguistically modified test – with NAEP items at the fourth and eighth grades. The results showed varied differences across the accommodations at the two grade levels. At fourth grade there was a significant difference between the accommodated and unaccommodated conditions for English dictionary and the bilingual dictionary. However, there was no significant difference in performance between the group that received the original items and the one that received the linguistically modified version.
Differential items functioning (DIF) techniques, such as Mantel-Haenszel approach, have been used to determine if items are operating differently for subgroups of the population. A study conducted by Ockey (2007) suggested that, even though non-ELLs outperformed ELLs on the NAEP, the score difference could not be attributed to DIF against ELLs. Only one out of ten NAEP items and none out of ten linguistically simplified items were found to exhibit DIF against ELLs.

A similar result was found Shaftel et al. (2006) when they examined the impact of language characteristics in mathematics test items on student performance for ELLs. They analyzed items on the Kansas general mathematics assessment. This is a traditional multiple-choice test of approximately 200 items. Relationships were examined for test items and students at 4th, 7th and 10th grades. Language characteristics, such as vocabulary and simplified syntax, had moderate effects on item difficulty at the 4th grade, dropping to small-to-medium effects at the 10th grade. ELLs were not disproportionately affected by language and difficult mathematics vocabulary had consistent effect on performance for all students, at all grade levels.

Moschkovich (1999b), in a study looking at the needs of Latino English language learners, disputes the myth that ELLs cannot participate in mathematical discussions because they are just learning English. The study revealed that ELLs are able to participate in discussions, even though they are learning English. Therefore, instruction and assessment should not emphasize low-level language skills over opportunities to engage in discussion of mathematical ideas.

Many states require ELLs to pass content-area tests in English to earn a high school diploma. Their approach to testing ELLs may be one of several ways. The may
permit the use of several accommodations, like extra time, translation into the student’s native language and dictionary. Although these practices offer some benefit in making high school graduation more accessible to ELLs, in reality, they only accommodate the needs of a limited amount of ELLs (Rivera & Vincent, 1997).

Do accommodations promote fairness to all students?

It is possible that an accommodation does not change the construct measured, while still providing an advantage to students who receive it. For example, if extra time is given, this accommodation addresses the issue of speed and does not change the intended construct (Sireci, Li & Scarpati, 2003). To defend the validity of accommodations, an interaction hypothesis has been proposed by several researchers (Shepard, Taylor & Betebenner, 1998; Malouf, 2001; Weston, 2002). This hypothesis states that students who need accommodations will benefit from them, while others will not. In essence, when an accommodation is given to students who need them, their test scores will improve, but other students scores will not improve if they took the test with the same accommodation. An illustration of this hypothesis, adapted from Sireci (2004) is shown below.

Figure A.1

However, several additional studies (Fuchs, et al., 2000; Anderson, et al., 2000; Meloy, et al., 2000; Albus, et al., 2001) have suggested a modification of the interaction
theory. Results showed that the most common accommodation (extended time), led to an improvement in test scores for both groups. However, the gains between the standardized and accommodated tests were greater for the students with disabilities and ELLs. These findings are consistent with the concept of *differential boost* (Fuchs et al., 2000; Phillips, 1994; Thompson et al., 2002). The differential boost hypothesis states that accommodations will yield greater score improvements for students with disabilities and ELLs than for the mainstream students. An illustration of this hypothesis, adapted from Sireci (2004) is shown below.

**Figure A.2**

If test accommodations result in a demonstration of the interaction hypothesis, then there is no disadvantage to students who do not receive the accommodation. However, if the accommodations result in a demonstration of the differential boost hypothesis, then it may not be fair to limit the accommodations to students with disabilities. As the *Standards for Educational and Psychological Testing* (1999) state “While test takers should not be disadvantaged due to a disability not relevant to the construct the test in intended to assess, the resulting accommodation should not put those
taking a modified test at an undue advantage over those tested under regular conditions” (p. 105).

The mixed results from these studies suggest that other options should be explored. The concern with accommodations is that they have been implemented with good intentions but with little empirical base to demonstrate whether using them makes any significant difference for the ELL (Spicuzza, et al., 1996; Butler and Stevens, 1997a). There is limited empirical research on the uses and effects of accommodation in large-scale assessment of ELLs (Lee, 1999). Further, accommodations may fail to correct unfair disadvantage or in some cases, may overcompensate to give unfair advantage. In either case, accommodations can reduce the validity of assessment results (Lee, 1999). Although some types of accommodations show promise for some groups of ELLs, the use of the same test accommodations with all ELLs is not likely to be an effective solution. This is due to the fact that different students, at different levels of English proficiency, have different needs (Butler and Stevens, 2001).

Validity of assessments

*Types of validity*

These pressures that students and teachers face coupled with the cultural and linguistic concerns that ELLs face raise the question of validity. A concern with all assessments, but in particular, performance assessments is the question of validity. According to *Standards for Educational and Psychological Testing* (1999), “validity is unitary concept which refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed used of tests (p. 9)”. A fundamental notion in test validity is that low test scores should not occur because of factors that are
irrelevant to the content (Messick, 1989, 1995a). As noted over seventy-five years ago, any test that employs language is to some degree a measure of language skills (Sanchez, 1934; LaCelle-Peterson & Rivera, 1994; Abedi, 2006). As Abedi (2002) asked: How valid are inferences about students’ knowledge when they are based on a test administered in a language that students may not understand? Recognizing the importance of this question, a 1999 edition of *Standards for Educational and Psychological Testing* devoted an entire section to testing English language learners, cautioning that test results should be weighed carefully because of the potential for confounding influences due to language deficiencies.

There are several types of validity that one can consider here; predictive, content, face and construct validity. Predictive validity is the degree to which a test is able to predict how well a student will do in the future, while content validity exams the degree to which a test measures the content area (Gay, Mills & Airasian, 2006). Face validity, the degree to which a test appears to measure what it claims to measure is perhaps the least important in the assessment of ELLs. Construct validity is the most important form of validity to be considered with regard to the English language learner because it asks the fundamental validity question: What is the test really measuring? Construct validity is the degree to which a test measures the intended construct of student’s content knowledge and skills in measurement and its applications (Kane, 2001; Gay, Mills & Airasian, 2006). Evidence of construct validity can be obtained from an analysis of the relationship between a test’s content and the construct it is intended to measure. (Kane, 2001).
Sources of validity

According to the *Standards* (1999), to evaluate test validity, one should begin with an explicit statement of the proposed interpretation of the test scores. This should include a clear description of the constructs or concepts that the test is intended to measure. Only then can one begin to discuss how valid a test might be, by measuring it against its intended purpose. The validation process is the development of a scientifically sound set of arguments to support the intended interpretation of the scores. While many may view this as the sole responsibility of the test developer, validation is the joint responsibility of both the test developer and the user. While the test developer is responsible for the development of the test, the user is indirectly responsible for providing evidence in support of its use. Student scores can be analyzed to give support to the argument for validity of the test.

Sources of evidence of validity

While there are several sources of evidence that can be used to support the argument for test validity (*Standards*, 1999), the ones highlighted here are those directly related to the assessments of ELLs.

*Content.* Evidence that is based on test content refers to the wording, themes, questions and format of the test, as well as guidelines for administration and scoring. This source of evidence can be evaluated by “experts” in the field who may supply tests items or vet them. Evidence about content can be used to start to address questions about differences in scores across different subgroups. Upon examination of the test, experts may highlight sources of irrelevant difficulty for certain subgroups. This looks at how broadly the content of a test is constructed, i.e. if it includes real-world, real-time
applications, or if it is limited to what is taught. For example, measurement in the traditional, standardized test may be restricted to knowing how to read the measurement from a chart; whereas the performance-based assessment affords students the opportunity to actually select the appropriate tool and do the measurements. In addition, the performance-based assessment can highlight how deeply the content was covered because students are afforded the opportunity to talk about what they are doing and how it relates to what they do at home. When students are asked low-level questions, as are most FCAT measurement questions, the answers will not reveal high-level thinking.

*Match to the intended construct.* Another source of evidence is evidence based on the internal structure of the test. This looks at the degree to which the relationships among test items and test components conform to the intended construct. Standardized tests like the SAT and the GRE make no claim about alignment to instruction, but FCAT does. However, results from Porter’s (2002) study reveal something different on the national level. He conducted content analyses of six states eighth-grade assessments to investigate the degree of alignment to instruction. The average within-state alignment was .22 and the alignment of instruction to NAEP was slightly higher at .39. The table, adapted from Porter (2002) to highlight the results, is given below in Table 2.4.

### Table 2.4

**Alignment of Instruction With Assessments: Eighth-Grade Math—SCASS Study**

<table>
<thead>
<tr>
<th>Instruction</th>
<th>H</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>E</th>
<th>O</th>
<th>NAEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>State  H</td>
<td>.35</td>
<td>.22</td>
<td>.19</td>
<td>.28</td>
<td>.21</td>
<td>.04</td>
<td>.38</td>
</tr>
<tr>
<td>J</td>
<td>.34</td>
<td>.21</td>
<td>.18</td>
<td>.25</td>
<td>.20</td>
<td>.05</td>
<td>.38</td>
</tr>
<tr>
<td>K</td>
<td>.42</td>
<td>.28</td>
<td>.21</td>
<td>.29</td>
<td>.25</td>
<td>.05</td>
<td>.39</td>
</tr>
<tr>
<td>L</td>
<td>.36</td>
<td>.24</td>
<td>.19</td>
<td>.29</td>
<td>.22</td>
<td>.05</td>
<td>.40</td>
</tr>
<tr>
<td>O</td>
<td>.35</td>
<td>.21</td>
<td>.16</td>
<td>.26</td>
<td>.20</td>
<td>.05</td>
<td>.38</td>
</tr>
</tbody>
</table>

*Note. Average within-state alignment = .22; average between-state alignment = .23; average state-test-to-NAEP alignment = .39.*
Authentic assessment, on the other hand, claims to be aligned with instruction (Stiggins, 1994; Valencia, 1990; Wiggins, 1989).

The concern is whether test items are tapping into the same knowledge of measurement. Test items may look like they do. However, Test A and Test B may look similar, but may not give similar results. If they do not give similar results, then either they are tapping into different constructs or they are tapping into different dimension of the same construct.

The student’s response process can also provide evidence concerning the fit between the construct and the response enacted by the students (Standards, 1999). Aspects such as eye movement or response times may be relevant to some constructs, such as those that assess mental or emotional abilities. Capabilities irrelevant or ancillary to the construct may influence student performance. In addition, consequences of the test can also be a source of evidence of test validity. Students and teachers feel the pressure of performing well, particularly if their promotion or salary is tied to the results. While this may not be a direct source of evidence, it could detract from the validity of the intended interpretation.

Use of tests. A sound argument for the valid use(s) for any test should include and integrate several of the aforementioned sources of evidence. The degree of validity for an intended interpretation of a test score should rely on all available evidence (Standards, 1999). This includes evidence of test construction, appropriate test administration and scoring, and accurate score scaling. The Standards document outlines twenty-four standards that ought to guide test development, administration and scoring. In the interest
of brevity and relevance, a few are highlighted here that might be more applicable to the assessment of the ELL.

A rationale should be clearly outlined for each interpretation and use of the test scores. The population(s) for which the test is appropriate should be clearly described. If a test is used in a different way, it is incumbent on the user to justify the new case. Since ELLs are typically omitted from the development of a test, its use of these standards with regard to ELLs needs to be justified. While one of the objectives of FCAT is to assess students’ knowledge of content, the test is also a test of language. This questions the validity of the test, particularly with regard to ELLs. In an environment where the home language is the language of instruction and the language of assessment, this concern is lessened. When validation rests on “experts”, procedures for selecting experts should be clearly described. When validation is determined by appropriateness of test content, the procedures followed for item development should be described and justified. Since the demographic makeup of the student country’s population is always changing; the environment under which the tests are taken, and the use of scores are constantly evolving, test validation should go through several iterations and should be continuous process.

Issues of Bias

Bias is the systematic tendency for one group to perform better or worse than another group for reasons that are not intended by what is being measured. It is the degree to which a test systematically over-estimates or under-estimates true performance on a test for reasons that are unrelated to one’s knowledge of the test content. A test item is possibly biased when the probability of a correct answer is associated with group
membership (Camilli, 1993). There are many possible sources of bias. For example, ELL students may be administered a mathematics test that is written beyond their ability to read. In this case, the students who performed poorly may do so not because they did not know mathematics, but because they could not read the test. In this case, the test contains a language-based bias against less proficient readers or ELLs. Another example is the cultural references or background information that ELLs need and might be lacking in order to correctly answer some items. Source of bias could also arise if the person who administers the test and/or scores the test knows and/or likes the student. Test performance could be over-estimated because of that person’s subjectivity.

**Identifying and Removing Bias**

There are two general ways for identifying bias: judgmentally and empirically (Shepard, 1981). Judgmental reviews are conducted throughout the development of the test and concerned with the individuals (judges) representing the relevant subgroups, such as ELLs, gender, students with disabilities, etc. Judges are asked to identify such things as stereotypes, unfamiliar content, cultural barriers and complexities of language. The other method of detecting bias is an empirical review to statistically determine whether individuals perform differently, using techniques such as differential item functioning (DIF) analysis. DIF analysis has been used to determine if items are operating differently for subgroups of the population. DIF is present when persons in two separate groups have the same ability or total test score (Howard, 1993), but have a different probability of correctly responding to a particular test question. Once DIF has been identified, representatives of the group should review the items exhibiting DIF to revise, remove or approve the item in question.
McNeil (2000) described the effects of differential education for minority and ELLs in mathematics. Their education emphasizes test-taking strategies and low levels of knowledge. The study of mathematics becomes “having students choose among four or five possible answers. They are not asked to explain their answers, so if students have alternative ways of working a problem, their reasoning is not made visible on the test, nor are their reasons for selecting the ‘correct’ answer” (McNeil, 2000, p.241).

Since most standardized are administered in English, and are designed for the native-English speaker, they may inadvertently function as language proficiency tests rather than content-based test. The ELL may be unfamiliar with the linguistic complexities of the language. Therefore, language factors are likely to reduce the validity and reliability of these tests and the inferences drawn about student knowledge (Standards for Educational and Psychological Testing, AERA, APA, & NCME, 1999).

Supporters of the standardized tests argue that externally developed and scored tests are more objective and hence safer for making decisions because local judgment is avoided (Wiggins, 1993). This type of test may remove one source of bias (e.g. prejudice), but may open itself to other sources of bias. Sources of bias in the assessment of ELLs include language, culture, background knowledge and SES factors.

Perhaps the most significant source of test bias affecting the ELL is language. When students are expected to perform on a timed test and answer questions in a language in which they are not proficient, this creates an automatic, systematic disadvantage. ELLs struggle with the meaning and concepts of individual words; they have trouble putting the words together in order to extract a sentence’s complete meaning (TESOL, 2002). For example, a student may know the meaning of the word “off”, as in
“Is the clock on or off?” However, when “off” is used in the phrase “the alarm clock went off”, an ELL student may struggle with the meaning that is clear for a native English speaker, who is familiar with the American culture. Another example is a student may be able to measure in inches but not in cubits (an old system of measurement).

**Validity and Bias with ELLs**

With regard to the ELL, the most significant area of bias is language. When students are expected to perform on a timed test and answer questions in a language in which they are not proficient, this creates an automatic bias (Abedi, 2002). Not only do they struggle with the meaning and concepts of individual words, they have trouble putting the words together for a complete meaning. For example, a student may know the meaning of the word “off”. However, when this is used in the phrase “the alarm clock went off”, they may struggle with the meaning. Did the alarm turn off or did it turn on? For the native English speaker, familiar with the American culture, the meaning is clear.

Critics have long argued that standardized tests are developed for and standardized against the performance of the white, middle-to-upper class groups and not minority groups (Darling-Hammond, 2000; Ladson-Billings, 1995). A number of researchers have pointed to the inherent flaws underlying the procedures and instruments used in the standardized testing of minority students and for their placement in special education (e.g., Jitendra & Rohena-Diaz, 1996; Ortiz & Garcia, 1995; Oller & Eilers, 2002; Harry & Klingner, 2006). According to Gonzalez, Brusca-Vega and Yawkey (1997), some of the problems include “a lack of appropriate norms, use of verbal standardized tests that are not valid and reliable for minority students” (p. 66).
Whenever a test is used as part of the data collection process, the concerns of validity, reliability and bias of that test are important. Just as we would not use a math test to assess verbal skills, we would not want to use a verbal skills test to assess mathematics. We would not want to use a measuring device that was not truly measuring what we purport it to measure.

Measurement

Importance & uses of measurement

Measurement tasks are frequently encountered in real-world situations. We are constantly measuring or estimating things on the job, in our homes, at the market, on the golf course, as we drive, etc. In addition, measurement estimation is important in its own right, because it is often the only method available for solving problems (Levin, 1981). Usiskin (1986) commented that although estimation may be considered the “weaker sister” to computation, in real life it is often the “stronger sister”, because tools may not be readily available and estimation may be the only option. In addition, successful estimation skills make it unlikely that students will mindlessly apply memorized procedures, but more likely that they will develop a conceptual understanding of number sense. Hence, it is a foundational topic that students need to master early. It is sufficiently important to be recognized by the mathematics education community and to be included as one of the content standards in NCTM’s Principles and Standards for School Mathematics (NCTM, 2000).

There are two kinds of reasoning that students must use to measure objects: transitive reasoning and unit iteration (Piaget et al. 1960). These reasoning abilities on how to measure apply to all objects to be measured: length, weight, volume, time and
temperature. The measurement of volume requires more cognitive development than the measurement of length because volume involves thinking and operating in three dimensions, whereas length is a single dimension. Length is a simpler idea for students to process, hence students are developmentally ready to learn the measurement of length before the measurement of volume (Reece, et al., 2001).

**Linear Measurement**

The process of measuring length involves the process of moving along an object, segmenting it and counting the segments (Piaget et al., 1960, Sophian, 2003). Measurement consists of identifying a unit of measure, subdividing the object by that unit and placing the unit end to end (iterating) along the side of the object being measured (Stephan & Clements, 2003). Concepts needed to complete a task include a) establishing a linear unit, b) partitioning, c) unit iteration and maintenance, d) accumulation of length, e) transitivity, f) conservation, and g) relation of unit iteration to number (Stephan & Clements, 2003). Partitioning is the activity of slicing up the length into equal units. Transitivity is the understanding that the relation follows through logically, where one is able to deduce a third relationship from two other relationships, as in equality.

Conservation is the idea that as an object is moved, its length does not change. The relationship between number and measurement is the understanding that measuring is more advanced than simply counting.

Piaget and his colleagues conducted activities with students that demonstrate these concepts (Piaget, Inhelder & Szeminska, 1960; Inhelder, Sinclair, & Bovet, 1974). They found that young children have limited reasoning skills in both transitivity and conservation. For children who cannot reason transitively, rulers are useless for
comparing two lengths that are not placed next to each other. Students need to see and experience this idea in the physical sense before they are able to view and internalize it abstractly.

In another activity conducted by Barrett et al. (2003), second grade students incorrectly measured the length of a side of a rectangle when given rulers that had the first two inches broken off. They did not take into consideration that the “zero point” was not zero but two. Stephan, Cobb and Gravemeijer (2003) describe a similar misconception that students have when they measure by covering distances. When students were measuring the length of a rug with their feet, most did not count the first position of their foot as “one”, but they started counting only after they moved to the second step. They did not see their initial pace as covering distance. They considered that they had only covered a distance when they moved. With the teacher’s help, students were able to describe their methods and their interpretations of what measuring length is.

In addition, Barrett’s (2003) study shows that children’s thinking about length fall into one of three profiles. In the first profile students are able to compare objects, but do not use units. They may recognize that one pencil is longer than another is, but cannot determine by how much. Other students, in the second profile, use addition and some notion of units. They are able measure by adding the unit measures. However, they encounter difficulty if with amounts left over. Students in the third profile use multiplication and units in their measurement tasks. These students are able to measure by multiplying the number of unit measures.

Piaget, Inhelder and Szeminska (1960) described the establishment of continuous linear units as an advance over the discrete units of number. They argued that “the
coordination of numerical (counting and ordering) and spatial (subdividing) schemes needed to form an abstract length unit often serves as a “coping stone” for operational thinking about measurement” (pp. 148-49). They base this on experiments they conducted with children completing measurement tasks involving simple, straight paths. One might question what accommodations, if any, are needed if the child were asked to measure non-straight paths such as perimeters. Studies conducted by Bailey (1973) and Hiebert (1981) suggest that students in grade one to three struggle to compare and order bent path lengths. Barret et al., (2006) conducted a study with students in grades two through ten, to attempt to answer this question. Students were asked to find all cases of a rectangle or triangle using a flexible straw with length of 24. As expected some students used their knowledge of algebra and perimeter to justify their claims. They found that students in the lower grades (2nd & 3rd) were focused on the issue of units and iterating units. Students in the middle grades were focused on generalizations, however they still maintained a reference to the physical representations of unit. As expected, the students in the upper grades immediately focused on generalizations without being concerned about the physical representations. Through their reasoning ability, they were able to apply their knowledge of algebra and perimeter in justifying their claims.

In the attempt to balance instruction between procedural and conceptual, educators must address the diverse understanding among their students at the different grade levels. Through their research on measurement, Barrett et al. (2006) were able to show that students were in some ways forced to draw on their reasoning skills to examine the relationship between shapes, parts of shapes, their measure and relation to one another.
Area and Volume Measurement

The concepts of area and volume are similar to those of linear measurement, with an obvious difference being the consideration of two or three dimensions. It is easy to visualize this idea as an array or little squares. Construction of the array is paramount to students’ understanding about area as a multiplication of lengths. In addition, construction of an array enables students to understand the unit of measure of area. In my experience, students tend to use the unit of length, despite the fact that they computed the area of a shape. Even when they write the square units, many do not understand the concept behind it. Construction of an array also enables students to comprehend the commutative property of multiplication, using rows and columns (Battista, 2003). Because of the prolific use of calculators in the classroom, many students do not learn the multiplication tables and miss the opportunity to explore its commutative property. Additionally, many children learn to “calculate” areas of shapes by a formula without understanding this concept or applying any reasoning behind it (Battista et al., 1998).

Although there are several points of concern with the measurement of area, a challenge for younger children is the idea of reallocation or rearrangement. They encounter challenges in arranging and rearranging the pieces of an object to verify that the areas are the same. When children examine rectangles with the same area from different perspectives, they typically focus on one dimension. They use terms like the “skinny one” or the “fat one” (Lehrer, Jaslow & Curtis, 2003). For example, they have difficulty seeing that the two shapes in the figure below have the same area.
Measurement becomes a way to resolve these conceptions in a concrete manner, by giving students the opportunity to decompose and compose an object to establish equivalent areas. This process of decomposing and composing also facilitates their calculations of areas of irregularly shaped polygons.

At the elementary school level, the measurement of volume can be viewed from the perspective of a solid or a liquid. Traditionally, in mathematics education, the focus is on solids and the volume of a liquid was left in the domain of science instruction. This may be the cause of confusion for some students. Students may develop two concepts of volume, rather than one concept being applied to the different states of matter.

Battista (2003) argues that the foundation for developing competence in area and volume measurement is to determine the number of squares or cubes in the object that we are measuring. He posits that there are four mental processes essential for meaningful measuring of area (squares) and volumes (cubes). These mental processes are 1) forming and using mental models, 2) spatial structuring, 3) units-locating, and 4) organizing-by-
composites. Students need to create or recall a mental representation about a situation. They need to see the situation in their “mind’s eye”. Then students need to form a spatial structure of the object’s components. For example, they may see an object as an array of cubes. This is only possible if they have previously encountered that idea and have it as a mental model. Thirdly, they must be able to locate the squares and cubes by describing their position in the dimensions of the array.

In a study investigating transitivity and unit iteration with volume, Reece et al. (2001) found that the minority of the students in 2\textsuperscript{nd} grade (33\%) successfully demonstrated transitivity. This result rose to 51\% in the 3\textsuperscript{rd} grade and 66\% in the 4\textsuperscript{th} grade. The percentage of students who demonstrated unit iteration also increased from 3\textsuperscript{rd} grade through to 4\textsuperscript{th} and 5\textsuperscript{th} grades. In addition, she found that all students who demonstrated unit iteration also demonstrated transitivity. This leads one to believe that students must learn transitivity before unit iteration. The order in which students develop measurement skills appears to be length, volume and then time. Although Reece (2001) criticized NCTM for expecting students to “understand how to measure using standard and non-standard units” in pre-K – 2, I do not see a conflict. It seems obvious that students will not learn all there is to learn about measurement in the early grades.

Students can be introduced to linear measure first and then volume and time. As students progress developmentally, they are able to think about a unit as one of many equal parts making up the whole. Therefore, teaching measurement precision is ineffective when students cannot make the mental connection between a cupful and the total amount.

Finally, students must combine basic squares and cubes into more complicated, composite units that can be repeated to make up the entire object. Instruction should
focus on supporting students’ movement through these processes with a view to develop the conceptual knowledge.

**Time Measurement**

At first glance, some persons might view the measurement of time as learning how to tell time. However, reading a clock is different from measuring time. The former is the record of time at a given point and the latter is the measure of the time elapsed. The measurement of time is a question of how does one measure something that one cannot see, feel or touch. Educators can teach the measurement of time by making an abstract concept more concrete.

Kamii and Long (2003), in a study with 120 students from California in grades K – 6, demonstrated transitivity by having students determine which piece of recorded music played the longest. The materials provided were a funnel, a bottle with water, an empty bottle and a marker. The same selection of music was played for 30 seconds and 40 seconds. The older children, with some prompting, were able to pour water through the funnel as the music played and then measure the amount of water in the bottle. The bottle with the larger quantity of water indicated that music played the longest. Their results indicated that 53% of the second graders and 90% of fourth graders successfully demonstrated transitivity. This is an illustration of going from abstract to concrete. Students were able to reason that if the length of time the music was playing transfers to the amount of water in the bottle, then the comparison can be made with the quantity with which they are more familiar and which is more concrete.

The same group of students was given a second task that would demonstrate unit iteration. Unit iteration applies in the measurement of time as it does in length, area and
volume. Students determined what a unit of time is and then performed several iterations to see how much time had elapsed. Students were asked to demonstrate that the second song was longer by using the materials provided. They were given two small vials, each with a hole in the bottom. They filled the vial with water and plugged up the hole with their finger before the music started. As the music started they removed their finger and allowed the water to flow out the vial. Since the song played longer than the time it took the vial to empty itself, the students had to compare a part to the whole. In this case the vial is the unit of measure which is used to determine the length of time elapsed. The results from this task indicated that not until 5th grade were most students (70%) successful at demonstrating unit iteration. Of the 4th graders, approximately 36% were successful at demonstrating unit iteration.

The younger children did not have clear understanding of transitivity nor unit iteration. However, there was a marked and consistent progression in abilities, as the children got older. If educators understand the developmental stages at which students are able to do certain tasks then they will be able to help them through the process.

Estimation and Measurement

Most adults estimate measurement everyday. Estimation is important in its own right, because it is often the only method available for solving problems (Levin, 1981). Usiskin (1980) comments that although estimation may be considered “weaker” because it does not utilize precise computations, in real life it is often the only option since tools may not be readily available in some environments. Some researchers in the area of estimation believe that measurement should not be taught with estimation because students may get the wrong idea about estimation. They may view it as a poor alternative
to precision measurement and hence undervalue its purpose. In reality, there are advantages and disadvantages to estimation. Estimate can save time, it can help verify the validity of the methods or it may be the only option available. The argument against measurement is that sometimes precise measurement is not only desired, it is essential, as there may be risks if it is not done (Adams & Harrell, 2003). An example of this might be that is not wise to estimate a person’s blood pressure or his cholesterol level. Certain disciplines require precise measurement.

Children begin to talk in estimation terms early. They often use phrases like “a little more”, “a lot taller”, etc. The ability to estimate improves as the child develops and it is related to the child’s perception of the realities around him. Children are often misinterpreted to be exaggerating when, in reality, they may not developmentally able to handle complex estimations. A child who sees ten dollars as “a whole lot of money” is perhaps a child whose reality is limited to dealing with small amounts of money.

Estimation has been identified as a critical area for development in mathematics, yet little is known about how students make their judgments (Joram, Subrahmanyam & Gelman, 1998). They highlight the importance for teachers to understand how students move back and forth between the written numbers and their meaning and positioning on a mental number line. For example, students need to recognize that 42 is closer to 40 than it is to 50. Therefore, when a student says “a little more than forty”, he/she will likely not be referring to fifty. When students are able to “see” the relative position of quantities, they will be able to make accurate and meaningful estimates. This is enhanced when students are given opportunities to compare quantities. Teachers should give students hands-on experience in comparing lengths, time, size, etc. In addition, successful
estimation skills make it unlikely that students will mindlessly apply memorized
procedures, but more likely that they will develop a conceptual understanding of number
sense. Students will be able to determine if they answers to problems make sense
(NCTM, 1989). This will promote conceptual understanding of measurement rather than
only a procedural one.

ELL performance on assessments and measurement

In the measurement strand, the NAEP results showed that the achievement gap
increases sharply from fourth to eight grades. The figure below, adapted from Lubienski,
Camburn and Shelley (2004), summarizes these differences. Overall, we see that
Hispanic-White gaps are large, but not as large as the Black-White gaps.

Table A.1 Achievement Differences by Ethnicity

<table>
<thead>
<tr>
<th></th>
<th>4th Grade</th>
<th>8th Grade</th>
<th>12th Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispanic-White</td>
<td>24</td>
<td>33</td>
<td>25</td>
</tr>
<tr>
<td>Black-White</td>
<td>31</td>
<td>39</td>
<td>34</td>
</tr>
</tbody>
</table>

The research reveals similar results from Western Australia. (Western Australia
Department of Education, 2000). Aborigines and ELLs also perform lower than their
white counterparts. The results are summarized below.

Table A.2 Western Australia measurement scores by strand

2000 Mean Scores on the Measurement Strand by Sub-group

<table>
<thead>
<tr>
<th></th>
<th># of Students</th>
<th>Mean Score</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>1634</td>
<td>481</td>
<td>105</td>
</tr>
<tr>
<td>Girls</td>
<td>838</td>
<td>486</td>
<td>104</td>
</tr>
<tr>
<td>Boys</td>
<td>792</td>
<td>477</td>
<td>105</td>
</tr>
<tr>
<td>ATSI</td>
<td>218</td>
<td>385</td>
<td>101</td>
</tr>
<tr>
<td>LBOTE</td>
<td>226</td>
<td>449</td>
<td>98</td>
</tr>
</tbody>
</table>
Table A.3 2002 Mean Scores on the Measurement Strand by Sub-group

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>1074</td>
<td>528</td>
<td>99</td>
</tr>
<tr>
<td>Girls</td>
<td>528</td>
<td>525</td>
<td>97</td>
</tr>
<tr>
<td>Boys</td>
<td>543</td>
<td>531</td>
<td>100</td>
</tr>
<tr>
<td>ATSI</td>
<td>205</td>
<td>457</td>
<td>87</td>
</tr>
<tr>
<td>LBOTE</td>
<td>136</td>
<td>527</td>
<td>103</td>
</tr>
</tbody>
</table>

ATSI = Aborigine
LBOTE = Language Background Other Than English

One can conclude that minorities are not receiving equitable education in mathematics and specifically measurement. There are many factors that could contribute to the differences in achievement. Some of these could include, but are not limited to, opportunities for learning, instructional practices, school culture, student learning styles, parental involvement, family beliefs, available resources and teacher knowledge about the content and their students.

Conclusion

This review highlighted the condition of the English language learners with regard to assessment and measurement. Assessments should show the extent to which students have met the learning objectives. Through demonstrating and documenting what students know, they should inform the teaching and learning of students. If the objectives of assessment are to be realized, continued efforts are needed to address students with special needs and from diverse backgrounds. While some may advocate for equal
treatment of students, not all students are created equal. Students enter the education system with intellectual, physical, emotional, cultural and language differences. Some students need more than others to bring them to a standard, acceptable level. In this vain, the objective for researchers and educators should be to find ways to provide equity for all students, providing the resources that addresses their special needs and circumstances.

As Thompson et al. (2002) so aptly puts it, “further research should…explore the effects of assessment to see whether more flexible testing conditions reduce the need for accommodations” (p. 17). It is possible that the need for accommodations will decrease and the measurement of what students know and can do will improve for all students.

References


TESOL (Teachers of English to Speakers of Other Languages) 2002: ESL standards for pre-K-12 students. Alexandria, VA: TESOL.


Western Australia Department of Education (2000). Student achievement in mathematics East Perth, WA.


Appendix B

Promoting Science among English Language Learners (P-SELL)

University of Miami – School of Education

Measurement Reasoning Task
Student Protocol

Forms
1. Interview protocol (bring with you)
2. Ranked list of students to interview (bring with you)
3. Parental consent and student assent forms (get from teacher)
   * UM pencil (thanking the student)

Supplies
1. 2 (or 3) plastic cups
2. Water
3. 1 container to hold water (e.g., 1-gallon container or 2-liter bottle)
4. 1 ruler (both centimeter and inch)
5. 1 kitchen scale (both grams and ounces)
6. 1 measuring cup (both cups and milliliters)
7. 1 graduated cylinder
8. 1 thermometer (both Fahrenheit and Celsius)

Equipment
1. Video camera (with power cord & charged battery)
2. Tripod for camera
3. Blank videotape
4. Microphone for the student

Contact your teacher in advance to arrange for three things:

1. Ask the teacher to arrange a room (a quiet place) to videotape the session. Set up a date and time that is good for her and her student as well as yourself (of course).

2. Ask the teacher to make a list of the names of the students who returned signed parental consent and student assent forms. On the day of the interview, you will start with the student who is listed first. If that student is unavailable for any reason, move on to the second student on the list, etc.

3. The teacher will be asked to watch the student interview prior to the follow-up teacher interview. Ask the teacher if she would prefer the interview on VHS tape or on DVD.
**On the Day of the Interview:**

1. Sign in at the office as a visitor and inform the receptionist of what you are doing.
2. Go to the classroom. Please ask the teacher to introduce you to the student to be interviewed.
3. Set up the videotaping equipment in the designated place (likely to be a library or a media lab).
4. Each interview session will take about 30-45 minutes.
5. After completing the session, take the student back to the classroom.

**Videotaping**

1. Videotape the session.
2. Use a new tape for each student interview.
3. Make sure to label the tape with the student’s name, school, teacher, grade, and date.
4. Set up the video camera to focus on the student and the materials on the table. Most important is to be able to see how the student manipulates the materials. Remove anything (e.g., a gallon container) that may block the view for the video camera.
5. Attach the clip-on microphone to the student and be sure it is turned on. If the student talks quietly, you may need to prompt the student to speak more loudly. If possible, check the audio level before and during the interview.
Interview Guideline

**Objectives:** The interview examines students’ reasoning that takes place during a measurement task. An environment of acceptance and support is essential in promoting students’ performance. **This is not a test.** It is an opportunity to learn from the student. Some students may be able to give detailed explanations of the tasks, others may be able to describe what they are doing but not explain why, and still others may not even be able to describe what needs to be done. Whatever the case, documenting the student’s actions and thinking behind those actions will allow the teacher to consider the role of student reasoning in providing relevant and effective science instruction and language development activities.

**Procedures:** To ensure the reliability of the data collection process, it is important to follow the procedures as they are presented in this document.

1. Make sure to record the student’s first and last name, as well as the date, on the paper and on the videotape.

2. The elicitation process involves a student and an interviewer – no one else should be involved in the activity.

3. Have the student DO the activity – Pour water in the plastic cup, take the actual measurements etc. – NOT just describe what they WOULD do.

4. Bring enough water in a gallon container or 2-liter bottle.

5. After each question, make sure that the student is allowed to give a full response. Do not begin the next question until you are sure that the student has finished thinking about the last one. When in doubt, ask “did you have something else to add?” or “are you still thinking about that?”

6. Encourage the student to initiate his/her response. If the student does not respond, use probing questions. Do not omit probing questions, since we want to be sure that all students are asked the same things.

7. Guidelines for interviewer language use:
   - DO use non-evaluative affirmation (e.g., ok, or anything else?)
   - DO NOT use evaluative affirmation (e.g., good, excellent, a great answer)
   - DO NOT use leading questions (e.g., What do you think would happen if you tried it this way?)
   - DO use the probing questions included in the guidelines
   - DO NOT make up your own prompts or probing questions

8. The interviewer should be careful when asking the student for clarification or making summary statements of student responses. These may become leading questions or put words into the student’s mouth. The interviewer can repeat the student’s answer
back to the student, but should not paraphrase it or restate it in the interviewer’s own words.
Grade 3 Student Reasoning Task #1 – Measurement

Opening: Tell the student, “We are going to do a science activity today about measuring things. I’m going to ask you some questions to see what you know about measurement. This isn’t a test and I don’t expect you to know all the answers. I just want to know how a smart kid like you thinks about these things. Ok? You can ask me any questions about what we are doing whenever you want to. Ok?”

Section 1 – Home Context
The purpose of the first section of the interview is to get the student to reflect on ways that he/she has engaged in measurement tasks (or observed others engage in measurement tasks) in the home context.

Initial prompt: We are going to start by talking about times when you measure things at home.

1. Have you ever seen anyone in your family measure how long or how tall something is? What were they measuring? How did they do it? What did they do it with?

   PROBE:

   How tall are you? How do you know?

2. Have you ever seen anyone in your family weigh something to see how heavy it is?

   What were they weighing? How did they do it? What did they do it with?

   PROBE:

   Who does most of the cooking in your family? Have you ever seen them weigh anything when they are cooking? What did they weigh? How did they weigh it?
Do you ever go to the food store (market)? Have you ever seen anyone weigh anything there? What did they weigh? How did they weigh it?

3. Have you ever seen anyone in your family find the volume or capacity of something to see how much space it takes up? What were they finding the volume of? How did they do it? What did they do it with?

PROBE:

Have you ever needed to take medicine when you were sick? Who gave it to you? How did they know how much to give you? How did they do this?

Have you ever helped do the laundry? How do you know how much detergent (soap) to put into the washing machine?

4. Have you ever seen anyone in your family take the temperature of anything? What were they taking the temperature of? How did they do it? What did they do it with?

PROBE:

Have you ever had a fever? Did someone find out what your temperature was? How did they do this?

Have you ever heard people ask about the temperature outside? How do they find out what it is?

Section 2 - School Context: Tool Use

Now I’m going to ask you to measure some things for me using some tools. This may be like the kind of activity you are doing in science time here in school.

First, I’d like you to take this plastic cup and fill it about half way up with water from this container.

Is that about half way full?
1. Now can you tell me the **height** of the water in the cup? Which tool do you want to use to find out how high the water is in the cup? Ok, go ahead and do it. Can you write your answer in the box for length? (point to place on student sheet)

   (NOTE: for all of the measurements, the tools have both metric and traditional systems. Let the student measure in whichever system she wishes. If the student asks you which system to use, tell her to use the one she prefers.)

2. Now I’d like you to **weigh** of the water in the cup. Which tool do you want to use to find out how heavy the water is in the cup? Ok, go ahead and do it. Can you write your answer in the box for weight? (point to place on student sheet)

   (NOTE: If student struggles with the units on the scale prompt with “Do you know how many grams are in a kilogram or how many ounces are in a pound? Can you find 1 kilogram or 1 pound on the scale? What do you think the 200, 400, 600, that go up to 1 kg or the 4, 8, 12 that go up to 1 lb stand for? Now can you tell me how heavy the water in the cup is?”)

3. Now can you tell me the **volume or capacity** of the water in the cup? (Prompt: Remember that means how much space it takes up) Which tool do you want to use to find out how much space the water in the cup takes up? Ok, go ahead and do it. Can you write your answer in the box for volume? (point to place on student sheet)

4. Last, I’d like you to tell me the **temperature** of the water in the cup. Which tool do you want to use to find out how hot or cold the water in the cup is? Ok, go ahead and do it. Can you write your answer in the box for temperature? (point to place on student sheet)
**Section 3 – School Context: Comparing and Estimating**

Now take this second plastic cup just like the first one. This time, I’d like you to fill it completely with water, all the way to the top.

Now I want you to compare the water in the half-filled cup with the water in the full cup.

I have some questions for you to answer. These are sentences that have multiple-choice answers and I want you to tell me which is the best answer. Can you read the first sentence to me?

1. The **height** of water in the full cup is higher than, lower than, or the same as the height of water in the half-filled cup.
   
   [ ] higher  [ ] lower  [ ] same

Remember when you measured the height of the water in the half cup you said it was [whatever student wrote down]. Now without measuring, what do you think the height of the water in the full cup is?

Now can you read me number 2? It’s another multiple-choice sentence.
2. The weight of the water in the full cup is heavier than, lighter than, or the same as the weight of water in the half-filled cup.

- [ ] heavier
- [ ] lighter
- [ ] same

Remember when you weighed the water in the half cup you said it was [whatever student wrote down]. Now without measuring, what do you think the weight of the water in the full cup is?

_________

Now can you read me number 3?

3. The volume of water in the full-cup is larger than, smaller than, or the same as the volume of water in the half-filled cup.

- [ ] larger
- [ ] smaller
- [ ] same

Remember when you measured the volume of the water in the half cup you said it was [whatever student wrote down]. Now without measuring, what do you think the volume of the water in the full cup is?

_________

Now can you read me number 4?

4. The temperature of water in the full-cup is warmer than, colder than, or the same as the temperature of water in the half-filled cup.

- [ ] warmer
- [ ] colder
- [ ] same

Remember when you took the temperature of the water in the half cup you said it was [whatever student wrote down]. Now without measuring, what do you think the temperature of the water in the full cup is?

_________

Section 4 – Connection to Play

So, we talked about measuring and estimating today. We practiced using tools like the ruler and the thermometer. Now the last thing I want to ask you about is if you
ever use measurement when you are playing with your friends or with your brothers or sisters.

1. Can you think of a time playing with your friends that you might need to measure something? Tell me about it.

   PROBE: Have you ever wanted to figure out how far something goes? How heavy something is?

2. In the example you were just talking about of [give a student example], do you think it would make a difference if you measured using a tool like the ruler or the scale, or would estimating be good enough? Why do you think so?

4. The tools we used today had two different systems, the metric system (centimeters, grams, milliliters, and Celsius degrees) and the traditional system (inches, ounces, cups, and Fahrenheit degrees). Which of these two systems is easier for you to use? Why do you think so?

   Ok, that’s all the questions I have for you. Do you have any questions you want to ask me about anything we talked about today? I want to give you this little present to thank you for helping me before we go back to class.
Analytical Framework for Grade 3 Measurement Student Reasoning

A four step analytical and coding scheme was developed to highlight the reasoning complexity of the English language learners in the study. This scheme adapted the analysis procedures developed by Hogan, Nastasi and Pressley (2000), Novak (1990), and Resnick, Salmon, Zeitz, Wathen and Holowchak (1993). Unlike Hogan et al. (2000) who looked at cognition as primarily situated in interpersonal interactions (but not in the broader social, cultural, and institutional settings), we look at cognition as situated in cultural settings of home and play as well as personal knowledge construction around tool use for measurement.

In the first step of the analysis, a semantic map of verbal and procedural responses and actions was constructed for each of the 27 student interviews. This step created an overall verbal and diagrammatic picture of each interview. A word count for each interview was completed. Construction of the semantic maps also served to identify three macrocodes (major types of statement categories): (1) unprompted responses, (2) prompted responses, and (3) clarifying questions. Figure 1 shows an example of a semantic response map.
Step two of the analysis involved the development of various conceptual proposition maps for each interview. Interviews were conceptually divided in two ways. First, conceptual propositions maps for each interview were developed for reasoning about each of the four types of measurement (length, weight, volume, and temperature) covered in the interview protocol. Figure 2 shows an example of a typical conceptual proposition map of a student’s reasoning about length. This analysis highlights the situated reasoning complexity about each measurement topic for each student. In this process, the initial semantic response maps were restructured to highlight the conceptual flow of each measurement topic, allowing for analysis of reasoning complexity within each topic. Second, the original proposition maps were divided along non-school-based
themes (home and play connections) and school-based (measurement with tools and estimation) themes, allowing for analysis of reasoning complexity across these two settings.

Figure C.2. Conceptual Proposition map of length

Step three of the analysis involved taking each of the conceptual proposition maps developed in step two and scoring them using a reasoning complexity rubric (see Table 1). This rubric was adopted from the one developed by Hogan et al. (2000), and was based on both emergent and theoretical descriptors to analyze the sophistication of
student’s reasoning. Reasoning complexity was scored along five criteria: (1) Generativity (amount and types of topics and assertions brought forth by students); (2) Elaboration (details added to ideas within a topic); (3) Justifications (how well ideas are supported with evidence and inference); (4) Explanations (how mechanisms are used to explain assertions); and (5) Logical Coherence (overall quality of connections within reasoning). The first two criteria account for the variety and richness of ideas raised by students, the third and fourth criteria assess the structure of student reasoning, and the final criterion assesses the overall quality of student thinking.

Table C.1
Student Reasoning Complexity Rubric

<table>
<thead>
<tr>
<th>Criteria</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generativity</td>
<td>No observations</td>
<td>One or two observations or simple</td>
<td>Three or more observations or simple</td>
<td>One or two assertions or conjectures</td>
<td>Three or more assertions or conjectures</td>
</tr>
<tr>
<td></td>
<td>observations</td>
<td>statements</td>
<td>statements</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elaboration</td>
<td>No elaboration</td>
<td>One or two elaborations of one idea</td>
<td>One or two elaborations of more than</td>
<td>Three or more elaborations of one idea</td>
<td>Three or more elaborations of more than</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>one idea</td>
<td></td>
<td>one idea</td>
</tr>
<tr>
<td>Justifications</td>
<td>No justifications</td>
<td>One or two justifications of one</td>
<td>One or two justifications of more than</td>
<td>Three or more justifications of one</td>
<td>Three or more justifications of more</td>
</tr>
<tr>
<td></td>
<td></td>
<td>assertion</td>
<td>one assertion</td>
<td>assertion</td>
<td>than one assertion</td>
</tr>
<tr>
<td>Explanations</td>
<td>No explanations</td>
<td>Single mechanism explaining one</td>
<td>Single mechanisms explaining more than</td>
<td>Multiple or chained mechanisms</td>
<td>Multiple or chained mechanisms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>assertion</td>
<td>one assertion</td>
<td>explaining one assertion</td>
<td>explaining more than one assertion</td>
</tr>
<tr>
<td>Logical Coherence</td>
<td>No logical connections</td>
<td>Vague connections</td>
<td>Clear and reasonable</td>
<td>Clear and reasonable</td>
<td>Clear and reasonable</td>
</tr>
</tbody>
</table>
When taken together, these five criteria present a balanced judgment of a student’s reasoning complexity on a given topic. It is important to note that this rubric does not assess the canonical quality of correctness of student answers regarding measurement; instead, it focuses on the quality of reasoning within the bounds of what the student does. Thus, a logical and well-connected explanation with many supporting details could receive a high reasoning score even if it is based on canonically incorrect factual information. We believe this approach is important in that students’ thinking in the context of science education is rarely considered in this light. While we acknowledge the value of studying students’ canonically correct conceptual understanding, we believe that studying students’ reasoning complexity is worthwhile in its own right.

In the fourth and final step in the analysis, a table was constructed to look at cross-case analysis of patterns in the results of the reasoning complexity rubrics. We looked for patterns in each of the four measurement sub-topics (length, weight, volume, and temperature), the two measurement contexts (school-based and non-school-based), the five reasoning complexity criteria, and across the student demographic variables of ESOL level, home language, gender, and school.
**APPENDIX D:** Table D.1

**3rd GRADE MEASUREMENT CONTENT SCORING RUBRIC**

<table>
<thead>
<tr>
<th>Topic</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Length:</strong> Did the student give an appropriate example for measuring length?</td>
<td>No response, inappropriate response, or solely responds to prompt</td>
<td>Example is appropriate and extends beyond responding to the probe</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Did the student name an appropriate tool?</td>
<td>No response or inappropriate response</td>
<td>Tool is appropriate and correctly named</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Weight:</strong> Did the student give an appropriate example for measuring weight?</td>
<td>No response, inappropriate response, or solely responds to prompt</td>
<td>Example is appropriate and extends beyond responding to the probe</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Did the student name an appropriate tool?</td>
<td>No response or inappropriate response</td>
<td>Tool is appropriate and correctly named</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Volume:</strong> Did the student give an appropriate example for measuring volume?</td>
<td>No response, inappropriate response, or solely responds to prompt</td>
<td>Example is appropriate and extends beyond responding to the probe</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Did the student name an appropriate tool?</td>
<td>No response or inappropriate response</td>
<td>Tool is appropriate and correctly named</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: To get point for giving an appropriate example within the prompt, it has to describe the process of tool use.*
**Temperature**: Did the student give an appropriate example for measuring temperature?
- No response, inappropriate response, or solely responds to prompt
- Example is appropriate and extends beyond responding to the probe

Did the student name an appropriate tool?
- No response or inappropriate response
- Tool is appropriate and correctly named

### School Context: Tool Use

<table>
<thead>
<tr>
<th>Topic</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Length</strong>: Did the student use the appropriate tool correctly?</td>
<td>No</td>
<td>Correct Tool</td>
<td>Correct tool AND number or unit</td>
<td>Correct tool, number, and unit</td>
</tr>
<tr>
<td><strong>Weight</strong>: Did the student use the appropriate tool correctly?</td>
<td>No</td>
<td>Correct Tool</td>
<td>Correct tool AND number or unit</td>
<td>Correct tool, number, and unit</td>
</tr>
<tr>
<td><strong>Volume</strong>: Did the student use the appropriate tool correctly?</td>
<td>No</td>
<td>Correct Tool</td>
<td>Correct tool AND number or unit</td>
<td>Correct tool, number, and unit</td>
</tr>
<tr>
<td><strong>Temperature</strong>: Did the student use the appropriate tool correctly?</td>
<td>No</td>
<td>Correct Tool</td>
<td>Correct tool AND number or unit</td>
<td>Correct tool, number, and unit</td>
</tr>
</tbody>
</table>

### School Context: Comparing and Estimating

<table>
<thead>
<tr>
<th>Topic</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Length</strong> correct answer: Higher</td>
<td>Neither correct comparison NOR reasonable estimate</td>
<td>Correct comparison OR reasonable estimate</td>
<td>Correct comparison AND reasonable estimate</td>
<td></td>
</tr>
<tr>
<td><strong>Weight</strong> correct answer: Heavier</td>
<td>Neither correct comparison NOR reasonable estimate</td>
<td>Correct comparison OR reasonable estimate</td>
<td>Correct comparison AND reasonable estimate</td>
<td></td>
</tr>
<tr>
<td><strong>Volume</strong> correct answer: Larger</td>
<td>Neither correct comparison NOR reasonable estimate</td>
<td>Correct comparison OR reasonable estimate</td>
<td>Correct comparison AND reasonable estimate</td>
<td></td>
</tr>
<tr>
<td>Topic</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>------------------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td><strong>Temperature</strong></td>
<td>correct answer: Same</td>
<td>Neither correct comparison NOR reasonable estimate</td>
<td>Correct comparison OR reasonable estimate</td>
<td>Correct comparison AND reasonable estimate</td>
</tr>
<tr>
<td><strong>Systems of Measurement</strong></td>
<td>No response or inappropriate response</td>
<td>Either measuring or estimating is chosen AND some correct reason is mentioned.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Measurement vs. Estimation</strong></td>
<td>No response or inappropriate response</td>
<td>Either traditional or metric is chosen AND some correct reason is mentioned.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>