11-1-2008

Effects of Plain Language Revision on Item Difficulty, Discrimination and DIF

Holly E. Baker
University of Denver

EFFECTS OF PLAIN LANGUAGE REVISION ON ITEM DIFFICULTY, DISCRIMINATION AND DIF

_____

A Dissertation

Presented to

The College of Education

University of Denver

_____

In Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

_____

by

Holly E. Baker

November 2008

Advisor: Dr. Ellen Steiner

Author: Holly E. Baker
Title: EFFECTS OF PLAIN LANGUAGE REVISION ON ITEM DIFFICULTY, DISCRIMINATION AND DIF
Advisor: Dr. Ellen Steiner
Degree Date: November 2008

Abstract

Colorado assesses approximately 98% of students enrolled in grades 3-10 in the area of mathematics. In 2005, 69,872 English Language Learners (ELLs) in grades 3-10 participated in the state content area assessments. In 2007, 88,060 ELLs participated in the state content area assessments. With this dramatic growth in the ELL population in the state of Colorado and elsewhere in the nation, ensuring that ELLs have access to comprehensible materials on state assessments is of paramount importance. Accommodations provide access for students based on their individual needs. Likewise, assessment items designed and/or revised using the principles of plain language and universal design increase access to the content area information in assessments. Colorado's work in plain language revision of Colorado Student Assessment Program (CSAP) items has been extensive.

Using statewide CSAP data for 2005 and 2007, this study examined how mathematics assessment items, once revised for plain language, performed compared to the original, unrevised version of the item. Access

for English Language Learners was a focal point for this study. Plain language revision was evaluated for its effects on item difficulty, item discrimination, DIF and DOK.

DOK ratings from 2005 and 2007 were compared using Chi-sqaure. One sample t-tests were used to examine the differences in DIF and item difficulty (p-value), as well as item discrimination for ELLs and Non-ELLs comparing the 2005 non-revised items with the 2007 revised items. Unrevised items were used as controls. Revised items were classified by type of linguistic change. Each type of change was examined using ANOVA to determine if a specific type of change effected differences in DIF and item difficulty.

While results indicate little if any difference in item performance at individual grade levels, overall LEP performance on revised items was higher than on non-revised items, a change in performance not present for non-ELLs. This may indicate that plain language revision continues to hold promise as a best practice for item development. As an accommodation, however, it appears that plain language revision alone may not be sufficient to ensure access to assessment items for ELLs.

ACKNOWLEDGEMENTS

Enabling Factors

Without the assistance of the Colorado Department of Education and CTB McGraw-Hill this research study would not have been possible. The Colorado Department of Education is focused on establishing itself as a “reliable source for research, data, and analysis envied by all professionals” (Jones, 2007, p. 18) and was thus supportive of this research project. Because of the researcher’s role in the Unit of Student Assessment at the Colorado Department of Education, proprietary assessment materials, not available to the public or other researchers, were available for this study. A great deal of gratitude is owed to my colleagues at the Colorado Department of Education, whose encouragement and support was deeply important during this process, particularly Elizabeth Celva, Director of Student Assessment, and Dr. Dianne Lefly, Director of Research and Evaluation.

Finally, I am forever indebted to my family and friends for their tireless support during this process. My fellow PhD candidates in the Mountain Cohort were a constant source of inspiration. My family, particularly my husband Rob and daughter Violet were patient and flexible and encouraged me every step of the way. Thanks to you all.

TABLE OF CONTENTS

LIST OF TABLES

EFFECTS OF PLAIN LANGUAGE REVISION ON ITEM DIFFICULTY, DISCRIMINATION AND DIF

CHAPTER ONE

Introduction

English Language Learners (ELLs) are individuals who are between the ages of 3 and 21 whose first language is not English, or are in an environment (household) where English is not the dominant language. The term ELL refers to students identified as having a home language other than English on the Home Language Survey in Colorado, and for who English is not their primary language. This includes students who are just beginning to learn English (NEP – non English Proficient, and LEP – Limited English Proficient) as well as students who have developed a degree of language fluency (FEP – Fluent English Proficient). While LEP is the term used in federal legislation in reference to all individuals for whom English is not the primary language, the term ELL, used throughout this study, focuses not on language limitations, but on students' accomplishments, that in addition to regular grade level content, students

are also learning a second language (La Celle-Peterson & Rivera, 1994). While ELLs may or may not have been born in the United States, they have difficulty speaking, reading, writing or understanding English to such an extent that their opportunity to participate fully in society is marginalized.

Colorado is considered a destination state in that, since 1994, it has seen an over 200% increase in the number of ELLs. While the total K-12 enrollment rate over the past 12 years has increased 13.37%, the enrollment for ELLs over the same time period has increased 352.68% (CDE & ELAU, 2007).

In Colorado, there is a significant persistent gap between the performance of ELLs and native English speakers as measured on their performance on Colorado's standards based achievement test. This achievement gap is the discrepancy between different subgroup populations achievement with the highest performing subgroup. The Colorado Student Assessment Program (CSAP) is Colorado's annual assessment that measures student performance relative to the Colorado Model Content Standards in reading, writing, math and science. Every student enrolled in a public school in the state of Colorado in grades 3-10 is required to take the assessment (CRS 22-7-409). This requirement is in line with the No Child Left Behind Act (NCLB) requirements that all

students be assessed, including students for whom English is not their primary language. NCLB requires annual assessment of students in reading, math and science. The results of these assessments are tied to accountability measures such as Adequate Yearly Progress (AYP) that evaluate schools based on student assessment results.

NCLB also requires that each state's assessment system be valid and reliable. While the Colorado Assessment Program received full approval with recommendations from the United States Department of Education (USDoE) in 2006 (Johnson), there is an ongoing need to examine the validity of the scores for ELL populations (Kopriva, 2000; Rivera & Stansfield, 2001).

Statement of the Problem

While the United States Department of Education (USDoE) has determined the Colorado Assessment System to be reliable and valid per No Child Left Behind (NCLB) requirements, a continuing concern lies in the issue of score comparability for different subgroups of students, particularly for English Language Learners (ELLs). This is a concern of Colorado, as well as every state on a national level. Conversations are centered on issues of validity and comparability of scores earned on those assessments. The CSAP assessments are created with both multiple choice and constructed response questions as outlined in the standards

and legislative requirements (CRS 22-7-409(1.5)(II)). The majority of the math items are constructed as word problems. For an ELL student, any problem involving words also becomes a test of language ability. This inherently introduces construct irrelevant variance for ELLs taking the assessments. Messick (1993) describes construct irrelevant variance as one of two primary threats to construct validity and as a "contaminant with respect to score interpretation" (p. 34). Construct irrelevant difficulty includes "aspects of the task that are extraneous to the focal construct make the test irrelevantly more difficult for some individuals or groups" and "leads to construct scores that are invalidly low for those individuals adversely affected" (Messick, 1993, p. 34).

While ensuring the validity and comparability of CSAP scores for ELLs has been a focus of the Colorado Department of Education's Unit of Student Assessment since the early stages of CSAP development, until recently there has not been consistent practice in development to facilitate this process. Kiplinger, Haug and Abedi (2000) examined impact of linguistic complexity on the new 5th grade mathematics assessment. The researchers used three test forms for the study, an original English version, a plain language (simplified English) version, and a glossary version where definitions were given for the non-mathematics terminology. Each test booklet contained the same test items, and only the plain

language version varied in linguistic structure. A sample of classrooms from thirteen Colorado school districts was used, yet the researchers had to oversample some schools to obtain sufficient numbers of Spanish speaking students to mirror the population of the state. While that study found that reducing the linguistic complexity of mathematics items benefited all students and did not compromise the mathematical integrity of the items, this finding was not permanently implemented into item development processes. Without this guidance, over time, linguistic complexity and superfluous language found its way into the mathematics items on the CSAP. In 2004, the Unit of Student Assessment began the work to ensure the CSAP items reflected best practice in terms of universal design and plain language. Universal design and plain language focus on ensuring access to material by presenting it in simple, clear formats, and reducing linguistic complexity and unnecessary verbiage which may impede comprehension and access. Plain language revisions applied to assessments in which content other than language is being measured, such as mathematics and science, reduce the construct irrelevant variance measured. Preliminary revisions were conservative, with most of the major linguistic revisions taking place for the 2006, and to a greater extent, the 2007 assessments in all content areas. As they appeared to be more conducive to linguistic revisions, the mathematics

assessments received the greatest focus in terms of plain language revisions.

NCLB requires that students be given appropriate accommodations to access the content of any assessment (Section 1111 (4) (a)), and a growing body of evidence for accommodations that potentially benefit ELLs specifically is becoming available. Accommodations include two types of changes, modifications of the test itself, and modifications of the test procedures. All accommodations are intended to "level the playing field" in order to obtain a more accurate picture of what students know and are able to do on content based assessments. Accommodations should not change, or simplify the content of the assessment, and thus should not provide an unfair advantage to students who receive them over students that do not receive the accommodation.

One of the most effective accommodations in providing access to assessment items, and decreasing the performance gap, as measured by standardized assessments is plain language revision (Abedi, Hofstetter, & Baker, 2001; Abedi & Lord, 2001; Brown, 1999; Kopriva, 2000; Rivera & Stansfield, 2001). This direct linguistic accommodation focuses on rewording items without changing the content and concept being measured. Plain language revision is referred to synonymously in the literature as linguistic modification, plain English, or simplified English. The

term used in this research project will be plain language revision. This term refers to the revision of complex linguistic structures, particularly in assessment items, to provide increased access to the content being measured.

Plain language falls under the larger umbrella of universal design, and has become a standard for government documents. Abedi, Lord and Plummer (1997) introduced the concept of plain language revision for assessment and identified features of text in assessments that may introduce unnecessary complexity and create construct irrelevant variance.

Plain language revision focuses on the following aspects in terms of item development and revision: reducing wordiness, using common words rather than unusual ones, reducing or eliminating unnecessary or low frequency words, reducing linguistic complexity, using shorter, simpler sentences, using words with only one meaning and avoiding irregular spellings of words (Hanson, Hayes, Schriver, et al., 1998).

Using the research on plain language, Colorado began to revise existing CSAP items for plain language and universal design and develop new assessment items using these principles. In 2007, major revisions were incorporated into both new item development as well as revisions to previously published items, with specific attention paid to mathematics,

science and writing assessments. Initial analysis (Lefly & Karkee, 2007) indicated that these revisions may have indeed contributed to some closing of the performance gap for ELL students in math in 2007 as evidenced by a gain in mean scores for Hispanic (linguistic minority) students between the 2004 and 2006 administrations of CSAP, a gain not shared with other subgroups, including non-linguistic minority students.

Some states, Delaware for instance, utilize an alternate form of their content area assessments designed using plain language (Brown, 1999). This is not a form available for all students.

This study is important as further analyses, such as item-by-item examination of the specific type of change and effect on performance and item bias for particular subgroups, are needed to determine the effectiveness and validity of these revisions. Since CSAP scores are used in part to make educational programming decisions for Colorado ELL students, closely examining the effects of the revisions is essential. An accurate and adequate measure of student knowledge relative to the Colorado Model Content Standards is essential, therefore, construct irrelevant variance must be minimized and access increased.

Text comprehension processes impact dramatically a student's ability to respond to mathematics word problems. Namely, the student must be able to map and decipher linguistic characteristics into their

knowledge about mathematics. This is problematic on many levels, but particularly troublesome in terms of construct validity, as developing linguistic skills can impede or hinder a students' ability to employ known problem solving strategies. Studies confirm that children find mathematical word problems difficult in part because they have difficulty interpreting key words and phrases, not because of a lack of conceptual knowledge required to solve a certain mathematics problem. For students with conceptual schemata necessary to solve a given mathematics problem, the textual structures can greatly determine student success, as students with more robust linguistic backgrounds are less likely to be hindered by linguistic structures, while students with weaker linguistic backgrounds are more likely to be confounded by the text itself, regardless of conceptual knowledge of the related content (Cummins, Kintsch, Reusser, & Weimer, 1988). As studies have indicated, reducing the linguistic complexity of mathematics items seems to benefit all students, not just linguistic minority students (Abedi et al., 1997; Kiplinger, Haug, & Abedi, 2000; Lefly & Karkee, 2007). By focusing on maintaining the integrity of the construct while minimizing unnecessary non-content related language, it seems that item development and revision using plain language principles may produce more clearly and concisely written items, potentially adding clarity

to the purpose of each assessment task, and paying particular attention to the potential presence of cultural bias.

Little research to date has examined how the different types of linguistic changes impact a student's performance on a particular item; however, because the types of changes are so diverse, and combinations of changes are often made, there may be varying levels of impact by type of linguistic change. Sato (2007) suggests that plain language revision of items can be examined according to changes in item context, item graphics, item vocabulary and wording, item sentence structure and item format. Additionally, the interaction of these different types of changes when combined in revisions is important to examine.

Purpose of the Study

The purpose of the study was to examine how Colorado Student Assessment Program mathematics assessment items, once revised for plain language, compare to the original version of the items and determine if plain language revision of items allows better access for English Language Learners. Plain language revision was evaluated for its effects on item difficulty, item discrimination, and differential item functioning. Analysis of item performance must also examine the results of statistical analysis which identify differential item functioning and bias for different subgroups of the population (Kopriva, 2000). This study examined the

effects of plain language revision of CSAP mathematics items for English Language learners (ELLs) in varying levels of language proficiency [non-English proficient (NEP), limited English proficient (LEP), and fluent English proficient (FEP)] on differential item functioning, item difficulty and item discrimination between two years of test administration, 2005 and 2007. Using these two years of test forms was essential. The 2005 assessment includes original, unrevised forms of mathematics items that in 2007 were revised for plain language. Using these two years of test forms provided the opportunity to compare item performance in original and plain language revision formats.

Differential item functioning (DIF) is used to help identify whether an individual test item yields differences in subgroup performance (for example performance between Caucasian students as compared to Asian students). It is an assumption that the difference is related to group membership, not due to the content of the construct (there must be an assumption of opportunity to learn). DIF may help identify the presence of construct irrelevant variance. The higher the DIF, the more unequally the item functions for different subgroups. Once issues with individual items are identified by DIF, further investigated, then addressed, differences between the observed and expected performance of the subgroups should be lessened.

For CSAP, item difficulty for multiple choice items is the percent of students that answered that particular item correctly. For constructed-response items, item difficulty is the mean percent of the maximum possible score for each item. Item difficulty for both multiple choice and constructed response items are communicated by p-value (CTB McGraw-Hill, 2007). Examining the item difficulty of items provides insight as to whether or not students of a particular subgroup (in this case ELLs) perform better on one form of the item versus another. A 1984 court ruling (Golden Rule Insurance Company and the Illinois Department of Education/ETS) specified that a difference of 0.15 or more in p-value between majority and minority subgroups was evidence the item was biased. For this examination, the item difficulties were examined from 2005 and 2007. Anchor items and unrevised items for both forms of the test were used as controls to determine if the change could be attributed to an overall change in performance between the two years, versus a change in performance due to plain language revision of the items themselves.

In an earlier study (Lefly & Karkee, 2007), CSAP items were examined for significant differences in item difficulty according to demographic variables. Items were classified as having a major or minor change. Minor change items included only those with formatting changes

such as adding bullets, adding white space, etc. consistent with universal design principles. All major change items constituted some type of linguistic change (possibly with additional formatting revisions). Neither the nature nor the degree of the linguistic change was analyzed as a part of that study. It is necessary to further examine the items and identify differences in item performance between the 2005 administration and the 2007 administration with revisions for plain language. This step helps determine the impact of the type of linguistic change as well as the interaction of students' English proficiency level with the plain language revised items in determining whether or not these revisions increase access for ELLs on the mathematics assessments.

For this study, the anchor and core item selections from 2005 and 2007 were examined. Anchor items are multiple choice items that are consistent across forms of the assessments. They appear in approximately the same location in both test forms, and are placed on a grade-specific scale through a common equating design. The anchor items were used to link the tests across years (CTB McGraw-Hill, 2007). The core items are those that also appeared in both assessments, but may have been revised for universal design and/or plain language. Computation items from both the core selections and anchor items sets

that appeared in both the 2005 and 2007 assessments were also included in this study.

Plain language revised items were examined and classified by type of linguistic change. The kinds of linguistic changes and revisions varied greatly between items and grades. Since the extent to which different types of linguistic changes impact performance had not been examined previously, it was important to explore whether different types of linguistic revisions yielded different effects in terms of performance leading to greater access for ELLs in the mathematics assessments. DIF measurement as well as item difficulty and item discrimination were examined for each of the different ELL subgroups (as well as for non-ELL students). Unmodified anchor items and computation items (not revised for plain language) were used as controls to examine if any changes in directionality for all students, as evidenced by Lefly and Karkee (2007), were changes that can be attributed by overall differences in group performance.

Research Questions

1. Does the cognitive complexity (DOK) of items change after plain language revisions?
2. Is item difficulty impacted by plain language revision?
3. Is item discrimination impacted by plain language revision?

4. Do CSAP mathematics items revised for plain language provide increased access to assessment items by reducing differences on individual item performance as measured by Differential Item Functioning (DIF) for all groups of ELLs (NEP, LEP and FEP)?
5. Are there certain plain language revisions (context, graphics, vocabulary/wording, sentence structure, and format/style) of mathematics assessment items on CSAP that reduce DIF and effect difference in item difficulty?

Null Hypotheses

NH_1 – Depth of Knowledge changes when grade 3-10 CSAP mathematics items are revised for plain language.

NH_2 – Item difficulty does not change on grades 3-10 CSAP mathematics items revised for plain language for any language proficiency levels.

NH_3 – Item discrimination does not change on grades 3-10 CSAP mathematics items revised for plain language for any language proficiency level.

NH_4 - DIF does not change for grades 3-10 CSAP mathematics after items are revised for plain language for any language proficiency level.

NH_5 - No specific type of plain language revision reduces DIF or effects item difficulty differences for students of any language proficiency level on grade 3-10 CSAP mathematics items.

Summary

With the increased focus on English language learners relative to inclusion rather than exclusion on assessments as well as heightened attention to the validity of assessment results of ELLs, ensuring that ELLs have access to comprehensible materials on state assessments is of paramount importance. While accommodations do provide access to students, assessments can also be designed and/or revised to provide increased access to content area information. Colorado's work in plain language revision of CSAP items to provide greater access to comprehensible content has been extensive, yet the merit and validity of this work must be examined in order to determine if the goal of increasing access for ELLs to the mathematics content of CSAP has been achieved. This study aims to provide important insight into the effectiveness and validity of items revised for plain language.

CHAPTER TWO

REVIEW OF THE LITERATURE

Introduction

State achievement assessment may be one of the most controversial aspects of the No Child Left Behind Act of 2001. Every student enrolled in a public school, regardless of language background, is required to take an annual assessment. In Colorado, statute dictates that all students must be assessed on the Colorado Student Assessment Program (CSAP) in reading, writing, and math in grades three through ten, and in science at grades five, eight and ten. These annual assessments are given at some point during a state-established five-week window of time each spring (C.R.S. 22-7-409). Results from these assessments are linked to various accountability measures such as Adequate Yearly Progress (AYP), the Colorado School Accountability Report (SAR) and Annual Measurable Achievement Objectives (AMAOs) as well as state accreditation. Because of the greater level of accountability for schools and districts relative to all students, and increasingly for ELLs, ensuring

the validity of the assessments is critical. States have numerous methodologies to build the validity argument for the assessments, from validity and alignment studies to post hoc analysis. During administration of the assessment, however, ensuring that students have access to comprehensible content leads to solid, accurate data on which states can begin to build a validity argument.

Universal Design

The concept of Universal Design began with Ron Mace, an architect who was also a wheelchair user. Mace began the focus on creating structures that worked for most people including everyone from young children to those for whom disabilities may have prevented access to structures without these design elements (Connell et al., 1997). These general ideas have been adapted for education generally and increasingly for assessment in order to ensure that students have access to comprehensible information in the assessment (Johnstone, 2003). When assessments are created with universal design as part of the development considerations, they “allow participation of the widest possible range of students” and thus “result in valid inferences about performance for all students who participate in the assessment” (Thompson, Johnstone, & Thurlow, 2002, p. 5). As it applies to large scale assessment, universal design focuses on aspects such as using plain simple fonts, use of

headings to organize information and help guide the reader, or using white space around items. Ultimately universal design should also make the assessment more amenable to accommodations a student may need in order to access the content of the items in the assessment.

Equity in Education

Achievement differences as measured on the CSAP between ELLs and native English speakers are pervasive, with few schools and districts successfully closing the gap between performance levels. In 2007, the performance of ELLs on CSAP fell far below the state average in reading, writing, math and science, with the greatest discrepancy in grades six through twelve (CDE & ELAU, 2007).

Access for ELLs

The No Child Left Behind Act of 2001 ensured that schools and districts were no longer able to exclude English Language Learners from assessment requirements, thus ensuring equitable access to not only assessment, but instruction as well. NCLB states that ELLs “must be provided reasonable accommodations, including “to the extent practicable,” in the language and form most likely to yield accurate and reliable information on what they know and can do in content areas” (Section 1111 (4) (a)). When a content area assessment is given in

English, regardless of one's content knowledge, for an ELL student, this assessment becomes a test of English language development.

> For all test takers, any test that employs language is, in part, a measure of their language skills. This is of particular concern for test takers whose first language is not the language of the test. Test use with individuals who have not sufficiently acquired the language of the test may introduce construct-irrelevant components to the testing process. In such instances, test results may not reflect accurately the qualities and competencies intended to be measured (American Educational Research Association, American Psychological Association, National Council on Measurement in Education1999, p. 91).

Language ability as well as background factors may confound an ELL student's ability to demonstrate content knowledge and reduce the validity and reliability of any inferences drawn from results of content standards-based assessments. Garcia (1991) contends that while ELLs may have a familiarity with vocabulary that is similar in both English and their native language, they may not have a complete understanding of the nuances that make them distinctly different in English. When assessments have been developed for the general population, and have not taken into account the unique needs of ELLs, the language complexity can provide a further challenge to this groups of students relative to what they can accomplish (Abedi & Gandara, 2006; Stevens, Butler, & Castellon-Wellington, 2000). For ELLs, linguistic complexity of assessment items is a major source of measurement error, and reducing the language load in

assessment items, particularly for math and science, may improve the validity and reliability of the assessments (Abedi, 2007).

Assessments of ELLs without appropriate accommodations could yield invalid results (La Celle-Peterson & Rivera, 1994). Since NCLB requires that states improve the validity and equitability of the inferences made from the state's standardized assessments, and requires an assurance of comparability between assessments for all learners ("No Child Left Behind Act of 2001," 2002) greater attention is now being given to using accommodations to increase the validity of assessment results. Students learning English as their second language need targeted accommodations necessitating a different approach to accommodations, as being an ELL is not a disability (La Celle-Peterson & Rivera, 1994).

Academic Language Development. The development of language is a process over time, and it may take several years for a student to attain the academic English language skills necessary to completely access required grade level content area knowledge. The term academic language refers to language and vocabulary commonly used in the classroom or other academic context where one is gaining knowledge (Stevens, Butler, & Castellon-Wellington, 2000), and includes knowledge of less frequent content vocabulary. Students who are proficient with academic language can interpret, understand, and communicate in more

complex linguistic structures. Students with a high level of proficiency in academic language should be able to understand complex content area information (Abedi, 2007).

Sato (2007) presents a distinction between language skill and academic language as they relate to necessary language in assessment items. There is language that is important for the construct being measured, and then there is language that is not important to the construct, such as language found in general test directions. A language demand that is essential to the development and use of language Sato considers a linguistic skill. Linguistic skills include demonstrating knowledge of phonemes, syllables, morphemes, vocabulary words, phrases and sentences, sound-symbol correspondence and written English conventions. These are all related to the linguistic domains of listening, speaking, reading and writing. If the language is applied contextually, however, it is considered academic language. This distinction is important in determining language skills from content area specific skills while revising assessments for plain language (2007).

Accommodations for ELLs

When a student with limited English proficiency is not provided appropriate accommodations, and thus cannot access the test materials to demonstrate their content knowledge, it is difficult to accurately measure students' content knowledge on an achievement test.

"Access" deals specifically with removing barriers for students so that construct irrelevant variance is eliminated and ensuring that the construct that is being measured does not change. Construct irrelevant variance can occur when access is hindered and something other than the intended construct is actually measured, or measured in addition to the intended construct. When a student is presented barriers, for ELLs these would be linguistic barriers, that impede access to assessment items, the student's skills and abilities may not be accurately represented and construct irrelevant variance may be introduced. Lack of access introduces a threat to validity (Abedi, Courtney, & Leon, 2003; Messick, 1993; Sato, 2007).

> Access is improved by systematically minimizing or removing sources of construct irrelevant variance – without significantly altering the assessed construct – in order to facilitate students' ability to demonstrate their construct-relevant knowledge and skills (Sato, 2007, p. 8).

Accommodations are particularly important when student performance may be impacted by level of English language proficiency or other background variables (Abedi, Courtney, & Leon, 2003). Regardless of level of English academic language proficiency, students are required to take Colorado's standards-based content area assessments. Students who require accommodations to access content area material are assumed to have such opportunity regularly provided in instruction and assessment. Students must not only be provided an equitable opportunity to learn in the classroom, but also an equitable opportunity to demonstrate their content knowledge on assessments.

Linguistic Accommodations

Items that have complex linguistic structures and a large amount of text will impact ELLs or students with disabilities who have difficulty with English text on tests (Kopriva, 2000). When assessments have a high language load, and that language is complex, ELL students' performance may be systematically underestimated and inaccurately represented. Assessments that are well designed are better measures of student knowledge and skills for all students, including ELLs. All students benefit from assessment items that use clear and concise language (Abedi, 2007).

Abedi (2007) examined the performance gap between ELL and non-ELL students in reading and mathematics using data from seven locations across the nation. In addition to reading, the assessed mathematics domains included math problem solving, math computation, math concepts, and estimation. The content areas with the largest gaps in student performance were those areas with larger language demands. The largest gap was identified in reading, and the smallest performance gap was in math computation - the area with the least language demand. While language factors have a greater impact on an ELL students' performance than any other factor, the pervasiveness of the performance gap between ELLs and native English speakers is further compounded by other factors such as poverty and the level of the parent's education (Abedi & Gandara, 2006).

Effectiveness, Validity, and Feasibility

Abedi, Courtney and Leon (2003) explained that certain conditions need to exist for an accommodation targeted specifically for ELLs to function in a way that provides access for students. Each accommodation considered must be effective, valid, and feasible. For an accommodation to be considered effective it must significantly improve the performance of those ELLs who receive the accommodation over those ELLs that do not receive the accommodation. There is continual concern that

accommodations may invalidate the construct being measured, so it is with careful consideration that accommodations must be selected. For ELL accommodations, when both ELL and non-ELL students receive the same accommodations only ELL students should show improvement in performance.[1] Finally, for an accommodation to be feasible, it must practical to administer, in that it does not require exorbitant resources. An accommodation must help increase the level of inclusion of ELLs into the assessment system (Butler & Stevens, 1997).

Direct Linguistic Accommodations

Direct linguistic accommodations, such as dual language versions, oral presentation in English or in a students native language as well as plain language revision provide adjustments to the text of the assessment. In studies examining accommodations appropriate for ELLs, it was determined that if an accommodation was an effective accommodation, it removed linguistic barriers for students (Abedi, 2007; Abedi et al., 2003, 2003; Abedi, Courtney, Mirocha, Leon, & Goldberg, 2005; Abedi et al., 1997). Effective accommodations proved to be a valid accommodations as well, for the accommodation could be provided to ELL and non-ELL learners, with only the ELL students benefiting (Abedi et al., 2003). Abedi, Leon and Mirocha (2001) found that as the level of language demand in assessments items decreases, so do the language demands on the

student, thus the performance gap between ELLs and non-ELLs also decreases. Higher language demands in test items increased the performance gap between ELLs and non-ELLs on those items, while on math computation problems there was zero performance difference. Appropriate use of linguistic accommodations can decrease the performance gap between ELLs and non-ELL students. When needed accommodations are provided to a student, access to the content is increased as the validity of the results. Accommodations are not, however, the only way to ensure all students have access to comprehensible content.

Design for Access

A requirement for universal design is now found in the reauthorized Individuals with Disabilities Education Act of 2004 (IDEA). Referring particularly to assessments given by a state or district that "to the extent feasible," universal design principles should be used in developing or administering any required assessments. This focus ensures that "barriers" to students are removed, thus increasing the likelihood that content knowledge can be accurately measured. Barriers students experience in demonstrating their content knowledge can often be mediated with an accommodation, but accommodations that have been traditionally available on standardized assessments have been focused on

providing access for students with disabilities. Universal design may actually reduce the need for accommodations by eliminating barriers within the test design itself, and may benefit all students in ensuring universal access to test item content (Thompson, Johnstone, Anderson, & Miller, 2005).

Assessments developed using universal design principles, and continually refined based on those principles, will provide the greatest likelihood that all students across all populations have the opportunity to participate in assessment, and in turn increase the validity of inferences of scores (Thompson et al., 2005). Continual reviews must also examine the results of statistical analysis which identify differential item functioning (DIF) and bias for different subgroups of the population (Kopriva, 2000). Modifying items for universal design alone is not sufficient to ensure access for ELLs (Sato, 2007). The provision of appropriate accommodations, in addition to assessments designed for access, plays an important role in ensuring students have access to comprehensible content.

Plain language revision

Plain language revision (plain language, linguistic modification, plain English, linguistic simplification are all used synonymously in terms of test development) is a subset of universal design principles. In terms of

plain language revision, several studies have indicated this approach as a linguistic accommodation to be not only valid, but also effective and feasible. Plain language revision is a

> "linguistically-based, systematic means for purposefully targeting and reducing the irrelevant variance in test performance that is attributable to individual differences in English proficiency so that ELLs are able to demonstrate fully their knowledge and skills related to the tested content" (Sato, 2007, p. 5).

While early studies showed mixed results in terms of student performance (Abedi, 1995), subsequent studies have confirmed that plain language revision provides greater access to assessment items by reducing the linguistic complexity and language barrier for ELL students. Abedi et al. (1998) found clarifying and simplifying language on math items increased all students' performances, but the LEP students benefited more than the fluent English speakers on 34% of the items on the plain language revised assessment.

Many studies have examined the impact of plain language revision of assessment items relative to ELL performance, and found that by reducing the complexity of the language, ELL students performed better on these revised items. Additionally, these studies have indicated that in the process of simplifying the language in the items did not alter the construct being measured, thus the validity of the assessments were not negatively impacted (Abedi et al., 2003; Abedi et al., 2005; Abedi & Lord,

2001; Abedi et al., 1997; Rivera & Stansfield, 2001). Abedi and Gandara (2006) asserted that linguistic complexity of assessment items may effect the performance of ELLs on those assessments and thus, increase the performance gap between ELLs and native English speakers, while plain language revision of assessment items may decrease that performance gap (Abedi et al., 1998).

A 1997 CRESST (Center for Research on Evaluation Standards and Student Testing) study examined linguistic simplification as an accommodation using NAEP mathematics items. Three different test forms were used, a Spanish version, a plain language version, and the original English version. Attention was paid to ensure that the construct and difficulty of every item was maintained and comparable to the original version. In this experimental study including 1400 California 8th grade students who were either language minority or national-origin minority students, only Hispanic students received the Spanish version. Both LEP and non-LEP students performed better on the plain language version of the math assessment. There were significant differences in p-values in 34% of the simplified items. From this study it was suggested that plain language items may be beneficial for all students (Abedi et al.).

Rivera and Stansfield (2001) examined the effects of plain language revision of science items on the Delaware state assessment (DSTP). Their study examined plain language revised items that were part of the embedded field test items in the regular test administration. The goal of the linguistic simplification was to "further clarify for LEP examinees the task or the context of each item and to reduce its reading difficulty level" (p. 8). Each form was assigned randomly, with two of six test forms including linguistically simplified items. Each form was administered to almost 1500 students in grades 4 and 6. Only a small number of ELL students participated in the 2000 DSTP, therefore, results were only used from the non-ELL samples. The ELL samples were, however, too small to provide results that could be generalized to the larger population. In addition to statistical analysis to examine differences in item-by-item difficulty, for items that showed a significant difference in item difficulty (p-value) the changes in wording were further examined to determine the potential cause of the difference. The results verify the validity of plain language revision being a valid accommodation as there was no significant difference in performance for non-ELL students who did or did not receive this accommodation. Thus, the accommodation can be offered to limited English proficient students

> "without the fear of providing them with an unfair advantage. Since linguistic simplification is able to reduce the level of English language proficiency needed to comprehend a test item, it is likely that it can reduce the role of language proficiency in achievement test scores, generally" (pp. 19-20).

The study also provided evidence that linguistic simplification of items will not impact score comparability. This adds evidence to the body of evidence needed for determining construct validity, although the researchers caution that in some instances, content and language interact to such a degree in items that if simplified, the content would also be simplified.

Even making minor changes in the wording of a test item may effect student performance (Cummins et al., 1988). Rakow and Gee (1987) point out the importance of distinguishing between readability formulas used for extended passages of text and the potential misuse of those formulas in determining readability levels of assessment items. Test items, in general, are much too short to employ the use of many of the readability measures available. An assumption behind the determination of readability levels is that easier vocabulary words and shorter sentences will make a passage more readable. Word count often plays a factor in determining overall readability as well. Even at best, however, readability formulas are imprecise measures. With an absence of appropriate readability measures for test items, general principles of plain language

and universal design must come into play. In essence, the focus turns to clear and precise language, presented in the most simplistic language possible and appropriate. Ultimately, Rakow and Gee (1987) contend, clarity of language in test items has a greater potential to provide students an opportunity to demonstrate content knowledge.

Several studies using released NAEP items compare student performance on original NAEP mathematics items versus plain language parallel versions of the same content area assessment. In the plain language versions, the math task and content specific terminology were retained. Students performed better on the plain language versions of the assessments. Features of the more linguistically complex items that posed barriers for students included passive voice constructions, and unfamiliar or uncommon vocabulary as well as longer item stems (Abedi & Lord, 2001; Abedi et al., 1998; Abedi et al., 1997).

Abedi, Courtney and Leon (2003) studied 8th grade students using a variety of accommodations including plain language revision and only the accommodation of plain language revisions narrowed the performance gap between ELLs and English proficient students while customized English dictionary and bilingual/English glossary did not. Studies consistently showed that the greater the language demand in an

assessment, the greater the performance gap between ELL and fluent English speakers.

Abedi and Lord (2001) found that plain language revision of test items benefited not only ELLs, but students in low level math classes and low SES students as well. In another study, Rivera and Stansfield (2001) also investigated the validity of accommodations. This study involved a separate plain language version of the Delaware state assessment that was administered to ELL and non-ELL students. A separate form, rather than a full length operational test was used, as the researchers feared providing students an unfair advantage with the plain language version. Although the sample size of ELLs was too small for results to be reported, they were able to determine that the accommodation of plain language revision did not impact the performance of fluent English speakers, thus providing evidence that plain language revision is valid and not a threat to score comparability.

Abedi (2007) contends that plain language revision is the most promising of all language related accommodations as it does not effect the validity of the assessment, yet narrows the performance gap between ELL and non-ELL students.

Guidelines for Plain Language Revision

Kopriva (2000) recommends the following guidelines on which to focus to improve access to test materials for ELLs.

Item Stems. Item stems and sentences need to be kept brief and straightforward, utilizing a subject-verb-object structure. Sentence length can also impact student performance, as longer sentences tend to be more linguistically complex in structure and syntax (Abedi et al., 1997).

Paragraph Structures. Paragraph structures should be consistent, and while the rules of good writing ask for variety in sentence structures and complexity of language, for standardized test items this should be avoided.

Active Voice. Present tense should be used as much as possible, and while sometimes the use of active voice may increase the number of words in a sentence; this is still preferable to passive constructions. In one study, 8th grade ELL and non-ELL students were given mathematics items with and without passive voice construction. Students in average math classes scored higher on the tests where non-passive structures were employed (Abedi et al., 1997).

Wording. Consistency in wording is also important, in that paraphrasing or using synonymous terminology within an item should be avoided. Utilize high frequency words, and avoid words with double

meanings or colloquialisms. Abedi, Lord and Plummer (1997) found that on a math test with items that were comparable in difficulty, 8th grade students scored better on items that utilized vocabulary that was used more frequently and was thus more familiar. Similarly, they found that language minority students performed better on math tests where shorter words were used than on items that used longer words. Longer words tended to be less frequently used. Also important is avoiding words that have unusual spellings such as “trough” or “feign.”

Primary considerations for plain language revision include ensuring construct relevance, in that the items measure what they intend to measure and minimizing skills that the item is not directly measuring; using clear and readable text that includes commonly used words; including grade level appropriate vocabulary which minimizes unnecessary verbiage; using only content related technical terms and abbreviations; writing with a sentence complexity that is appropriate for the grade level, and the task or question (Thompson et al., 2005).

Eliminating Construct Irrelevant Variance

When an assessment requires highly developed reading skills, students with reading disabilities or ELLs may be unfairly disadvantaged as their ability to demonstrate content area understanding is impeded by reading skills required to access the content of the assessment item.

When this occurs, the measure of a students ability in mathematics or science is confounded by the additional, unintended measure of reading (Hanson, Hayes, Schriver, LeMahieu, & Brown, 1998). Abedi et al. (2001) identified that student who were better readers, as measured by reading test scores, also had higher math scores in a NAEP item study. Text comprehension processes impact dramatically a student's ability to respond to mathematics word problems. Namely, the student must be able to map and decipher linguistic characteristics into their knowledge about mathematics. This is problematic on many levels, but particularly troublesome in terms of construct validity, as developing linguistic skills can impede or hinder a student's ability to employ known problem solving strategies. Studies confirm that children find mathematical word problems difficult in part because they have difficulty interpreting key words and phrases, not because of a lack of conceptual knowledge required to solve a certain mathematics problem. For students with conceptual schemata necessary to solve a given mathematics problem, the textual structures can greatly determine student success, as students with more robust linguistic backgrounds are less likely to be hindered by linguistic structures, while students with weaker linguistic backgrounds are more likely to be confounded by the text itself, regardless of conceptual knowledge of the related content (Cummins et al., 1988).

By incorporating elements of plain language into content area assessment item development, the language demands on the student are reduced, thus construct irrelevant variance is also minimized (Hanson et al., 1998). Hanson et al. (1998) examined original and plain language versions of math and science items in the Delaware state assessment utilizing student "think alouds" to guide the revision of items. The study also employed item reviews by an expert panel of judges to examine and rate the content and skills of the items on each form of the test. Reviews of both the original and plain language versions of the assessment were compared to determine if the content had changed as the item was revised for plain language. The researchers found the think aloud process and input from the students to be valuable, as students identified areas in need of revision that the item reviewers had not initially identified. As reviewers' ratings were compared, only three of the revised items were identified as not measuring the same content and/or skills.

A 2003 NCEO study examined whether eliminating construct irrelevant materials in assessments yielded a more accurate measure of student performance. The instrument was designed following the requirements of universal design to determine whether it provided greater access to students who required accommodations. Two different studies were actually used to test the research questions. The first study used two

different forms of identical test items; one group received items revised using the principles of universal design, the other group receiving the original, unrevised items. This study found that as long as the construct was maintained, the design of the test impacted student performance. On the revised versions, all subgroups of students performed better. The second study included an examination of student experiences and their perceptions while they took the two different versions of the assessment. Here, students indicated that the design elements such as readability and recognizable materials were important for the students to be able to perform well (Johnstone, 2003).

Differential Item Functioning

The main question explored by DIF is, “do students who receive a similar test score on a particular test respond differently to an individual item from the test, and does this difference appear to be systematic by group?” and then “is this difference due to irrelevant characteristics, such as racial/ethnic group membership, or is the difference due to content central to the construct?” (Kopriva, 2000). An assessment item is intended to measure only one construct. If an item contains DIF, that item measures more than the intended construct.

While DIF analyses can be used to identify differences in item responses caused by construct-irrelevant variance Kopriva (2000) urges another step in validating DIF results, that items are further examined to determine the source of the variance. The advantage of using DIF is that it can examine differences item by item, and provide greater opportunity to examine potential sources of variance within each item.

Ockey (2007) examined whether or not math word problems inherently exhibit DIF against ELLs, that is if non-ELLs outperform ELLs in math word problems. The researcher used extant data from an earlier Abedi study involving 20 released NAEP items, ten of which were revised for plain language. Of the 1,174 students (mostly from the eighth grade) that participated in the original study, 372 were classified as ELL. Two different DIF detection methods were utilized, the Bilog-MG since it is effective at identifying uniform DIF using an IRT approach, and the Mantel-Haenszel (MH) because it is a commonly used, nonparametric technique which uses the total score across items. Bilog-MG detects DIF using a three parameter IRT model for multiple groups using the BILOG IRT analysis software. The Mantel-Haenszel method uses a Chi-square test to compare the reference group and the focal group (subgroup) at matched ability levels. The researcher found that non-ELLs did outperform ELLs in math word problems by nearly 3 points on the 20 point scale for

the assessment. Only one of the original items (not revised for plain language) exhibited DIF against ELLs in either of the techniques used. From the results of this study, Ockey suggests that math word problems do not inherently contain DIF against ELLs. Ockey does acknowledge, however, that if every item performs differentially against a subgroup, in this case ELLs, then the techniques used to identify DIF would not be effective, and no items would be identified for DIF. If every item included in this study performed differentially against ELLs, then the measures used to identify DIF would not be effective. This may be the case given the large performance difference between the two groups in this study.

ELLs in Colorado

English Language Learners in Colorado include students who are categorized as non-English proficient (NEP), limited English proficient (LEP), as well as students categorized as fluent English proficient (FEP) in monitored year one (M1) or monitored year two (M2) status. According to the English Language Acquisition Unit State of the State Report (2007), in 2003-2004, the number of total ELLs was 59,309. By 2006-2007, the total number of ELLs had risen to 100,039. In 2006-2007 almost 45% of ELLs in Colorado were designated as LEP, 39.4% were categorized as FEP, and 16% as NEP. Of the ELLs in Colorado, only 1% are Native American, 2.1% are black, 5.2% are white, 8% are Asian, and almost 84%

are Hispanic. Almost 9% of ELLs are categorized as having a disability. These dually identified students comprise 1.28% of the total student population in Colorado.

Growing Population

In the last twelve years, the number of ELLs enrolled in Colorado schools has increased over 352% (Unit of Student Assessment & English Language Acquisition Unit, 2007). The majority of ELLs are enrolled in grades pre-K and K-4, with over 69% of ELLs living in a metropolitan area. Of all ELLs in 2006-2007, 77% were eligible for free and reduced lunch. Only 6.1% of ELLs moved to a different school district between the years 2004 and 2007(Unit of Student Assessment & English Language Acquisition Unit, 2007).

Colorado Accommodations

Colorado provides for many different accommodations for English language learners in addition to the accommodations available to students with disabilities (Colorado Department of Education, 2007). The barriers students experience in demonstrating their content knowledge can often be mediated with an accommodation, but accommodations that have been traditionally available on standardized assessments have been focused on providing access for students with disabilities. Students learning English as their second language need targeted accommodations necessitating a

different approach to designating and providing appropriate accommodations, as learning a second language is not a disability. Accommodations can be used to better identify what students know and are able to do relative to the constructs being measured. The goal of any accommodation is not to give an unfair advantage to students, but rather an accommodation should “level the playing field” for all learners.

It is Colorado’s assumption that providing accommodations from the list of approved standard accommodations introduces no threat to validity or comparability of the items as they are intended to “level the playing field” for students needing the accommodation (Abedi et al., 2003). Provision of approved accommodations for students who require such to access the content of assessments will actually increase the validity of the results of the CSAP and other state assessments (Colorado Department of Education, 2007, 2007) . Additional accommodations may be allowable on state assessment provided the student has a disability documented on an Individual Education Plan (IEP) or 504 (Colorado Department of Education, 2007). Thoughtful consideration must be made before any accommodation is provided on a large scale assessment as the accommodation may in fact change the nature of the item or construct to such a degree that accommodated or modified items would no longer be deemed comparable (Johnstone, 2003). Colorado requires that all

decisions regarding the provision of accommodations for students are made by a team of educators and include the parent. For an accommodation to be effective, it cannot be introduced on the day of the assessment for only that purpose (CDE, 2007; Colorado Department of Education, 2007, 2007)

In September of 2007, the Colorado Accommodations Manual for ELLs (Colorado Department of Education, 2007) was published, providing information on research-based, effective accommodations appropriate for ELLs that could be used in instruction as well as accommodations that could be used on state assessments. In previous years, all accommodations, including those for ELLs, were under the purview of the special education team at the Colorado Department of Education. Several accommodations appropriate for ELLs have been allowed on the CSAP including extended timing, word to word dictionaries, oral scripts, and teacher-read directions. Both oral scripts and teacher-read directions are developed and published in English. Mathematics, writing and science oral scripts are provided to districts in a Spanish translation. If needed, districts may translate teacher-read directions as well as oral scripts into any languages needed. Very few ELLs, however, receive any accommodations for CSAP. For instance, in 2007, 39% of NEPs taking the mathematics CSAP did not receive any accommodation. Sixty-eight

percent of LEPs taking the 2007 Mathematics CSAP in grades 3-10 did not receive any accommodation (Lefly, 2007).

Spanish Language Versions. The CSAP forms are developed and published in English for all content areas. There is a provision in the Colorado Revised Statutes for Spanish language versions of reading and writing, and this has resulted in the Lectura and Escritura, which are Spanish language reading and writing assessments for grades three and four (22-7-409 (1)). They are not translated versions of CSAP; they are unique Spanish language reading and writing assessments developed as accommodations for students in specific programs. To be eligible to take this assessment, students must be receiving instruction and assessments in Spanish and enrolled in a dual language program. There is not a Spanish language version of the mathematics or science CSAP. For NEP and LEP students not instructed in Spanish, the provision of accommodations may be essential in accurately determining their content knowledge (Colorado Department of Education, 2007).

Colorado Revisions for Universal Design

The state of Colorado includes educators in many phases of the assessment cycle, including input during framework and item specification development, new item development, and scoring for constructed response items. Within this group of educators are included not only

representatives from the diverse geographic regions of the state, but also educators who have received specific training in working with students with disabilities, gifted and talented students, and English language learners. Inclusion of these, along with content area experts, is essential in ensuring that the assessments are indeed accessible for the different student populations, and that items are bias free (Kopriva, 2000). The input of these teams of educators assists in Colorado's work to ensure that items represent the principles of universal design or are revised to meet these standards.

The National Center on Educational Outcomes presents seven elements of universal design as it applies to large-scale assessments (Johnstone, 2003). Each element relates to the revisions Colorado has included in its development process and goals for revisions under Plain language and Universal design. How Colorado addresses each of the elements follows.

Inclusive assessment population. Colorado Revised statues require that all students enrolled in a public school in grades 3 through 12 participate in the state assessment. This, along with NCLB requirements, ensures that every student enrolled in public school receives the opportunity to demonstrate their skills and knowledge relative to the Colorado Model Content Standards.

Precisely defined constructs. In designing the CSAP items, developers and reviewers pay particular attention to maintaining fidelity of the construct. Items are reviewed on several occasions for construct alignment. They are also reviewed with a specific focus on minimizing potential barriers, including linguistic barriers in content area assessments. By removing these barriers, results provide a more accurate view as to areas of the educational system in need of improvement.

Accessible, non-biased items. Quantitative and qualitative analysis for bias is done on an annual basis for every item that appears on each CSAP. This is an important part of the data analysis as the state seeks to minimize any disadvantage an item may have presented to a test taker. If an item presents a significant amount of bias upon review, it may be suppressed from score calculations and either revised or removed entirely from the item bank. Along with the content alignment review, a comprehensive bias review is also conducted on every new item. This is done early in the development process, employing the expertise of Colorado educators who are not only representatives of all the geographic regions of the state, but who also bring to the review an eye for a particular subgroup of students, such as visually impaired, or English

language learners. The input of these educators is considered as items are revised and refined for the operational assessment.

Amenable to accommodations. Much of the research relative to the validity of accommodations, and particularly accommodations for ELLs, comes after NCLB deadlines have passed requiring states to have content standards based assessments developed and operational. For that reason, it may be necessary to retrofit assessments to ensure maximum accessibility, as is the case with the CSAP revisions for plain language for the 2007 assessment. While adhering to the principles of universal design in the development of assessments may increase access to the assessment items, it does not necessarily eliminate altogether the need for accommodations. It is important, however, to consider the elements of universal design in new item development to reduce, as much as possible, a condition under which accommodations could diminish the comparability of items because the item itself was not conducive to administration under accommodated conditions.

Simple, clear and intuitive instructions and procedures. In order to ensure that students understand the content of the item, and how they are in turn to respond to the stimulus in an item, clear directions are essential. Revision of instructions for CSAP items included simplifying instructions by reducing unnecessary verbiage, directing students' attention directly to

important information such as tables or graphs, bulleting multi-step directions or processes, and ensuring that directions appeared for every item.

Maximum readability and comprehensibility. This includes measures that calculate a level of readability and comprehensibility of a text or passage, as well as the organization and clarity of ideas. While the readability of longer pieces of text can be easily identified, determining the readability of a short item that has only one or two sentences is more problematic. In order to improve readability, Gaster and Clark (1995) recommend using simple, commonly used words; defining any technical terms; breaking down compound, complex sentences into many shorter sentences; introducing one idea or fact at a time, ensuring that ideas are presented and progress in a logical manner; and establishing clear noun-pronoun relationships. Readability is calculated for CSAP text and passages, and individual words are often examined for grade-level appropriateness. Whenever possible, for the mathematics and science assessments, non-content related words are used only if they are at least two grade levels below in terms of readability to ensure maximum access. Footnotes or glossary boxes provide definitions for non-content related words and terms as well. Ensuring maximum readability is one of the

primary considerations guiding the revisions regarding plain language on CSAP.

Maximum legibility. In terms of assessment, maximum legibility relates to contrast, type size, spacing between letters, words and lines, the design of the typeface or font, margin alignments, line length, and use of blank space throughout the document. All of these elements have been incorporated into the style and revisions of the CSAP to ensure legibility. Additionally, superfluous art has been removed or replaced by art that enhances the meaning of the passage or item, or graphics that enhance student understanding of a text passage or item.

Gap in the Literature

As Francis et al. (2006) pointed out; previous studies on plain language revision have had some methodological concerns, namely small sample sizes, which require one to question the validity of plain language as an accommodation. Colorado is in a unique situation in that the items revised for plain language have been operationalized in all forms of the test. Every student in the state who has taken the mathematics CSAP has had the opportunity to respond to the plain language revised items, therefore, state-wide data exists relative to student performance on revised as well as unrevised items.

For this study, state-wide CSAP data were used, providing one of the largest sample sizes of any plain language revision study to date. This study is important as it examines the validity of plain language revision as a development tool used to minimize threats to validity due to linguistic barriers in math assessment items. It also provides information relative to the types of plain language revisions used most frequently.

Chapter Three

METHODOLOGY

The purpose of the study was to examine how Colorado Student Assessment Program mathematics assessment items, once revised for plain language, compare to the original versions of the items and determine if plain language revision of items allows better access for English Language Learners. The study examined the following questions:

Research Questions

1. Does the cognitive complexity (DOK) of items change after plain language revisions?
2. Is item difficulty impacted by plain language revision?
3. Is item discrimination impacted by plain language revision?
4. Do CSAP mathematics items revised for plain language provide increased access to assessment items by reducing differences on individual item performance as measured by Differential Item Functioning (DIF) for all groups of ELLs (NEP, LEP and FEP)?
5. Are there certain plain language revisions (context, graphics, vocabulary/wording, sentence structure, and format/style) of

mathematics assessment items on CSAP that reduce DIF and effect difference in item difficulty?

CSAP Development and Administration

The CSAP is administered each spring to every third through tenth grade student enrolled in public school in Colorado. No students are exempt from taking the assessment, so the data gathered includes students of all ability levels and language proficiency levels. Every other year the core test form for the CSAP is repeated, with 25% of the test form being refreshed each administration.

When items are developed for the CSAP, the Colorado Model Content Standards and assessment frameworks are analyzed by a team of content and assessment experts at CTB who then develop preliminary items based on the item specifications. Item specifications, which are developed collaboratively by CDE assessment staff, CTB content and assessment experts, and Colorado educators with content area expertise, outline specifically what can and cannot be assessed at each grade level for each content area for each benchmark, and how these components of the assessment frameworks should be assessed. After development and a first round of revisions, preliminary items are taken to Colorado stakeholders (teachers, curriculum experts, higher education personnel) for input and review. In these reviews, participants examine every

proposed new item for grade level appropriateness, alignment to the Colorado Model Content Standards and Assessment Frameworks, level of Depth of Knowledge, and potential bias. This input is used by CDE and CTB to further refine items. Final versions of items are used to augment the item pool for each content area and grade level assessment. Besides new items, CTB pulls items from a pool of previously used items. Many of these items are revised to better align with the Colorado Model Content Standards.

Each test form must meet the requirements of the blueprint. The blueprint specifies the percentage of items from each test that come from each standard (Table 1). This is consistent from year to year, and new item development is determined, in part, by the number of items needed to meet the blueprint requirements.

Table 1.
CSAP Mathematics Blueprint

CSAP Mathematics Blueprint					
Percentage of Content Standard Representation by Grade Level					
	Standard 1	Standard 2	Standard 3	Standards 4/5	Standard 6
Grade 3	20%	25% (Combined 2&3)		35%	20%
Grade 4	20%	15%	15%	30%	20%
Grade 5	20%	20%	20%	20%	20%
Grade 6	20%	20%	20%	25%	15%
Percentage of Content Standard Representation by Grade Level					
	Standard 1/6	Standard 2	Standard 3	Standards 4/5	
Grade 7	30%	20%	20%	30%	
Grade 8	25%	25%	20%	30%	
Grade 9	20%	30%	25%	25%	
Grade 10	20%	30%	25%	25%	

New and revised items are taken annually to the Content Validity and Alignment Review where Colorado educators (teachers, curriculum experts, and higher education faculty) representing all geographic areas of the state review items for content validity, quality, appropriateness, and presence of bias. Specifically, each content and grade span team reviews items for accessibility and grade level appropriateness, as well as content and scoring guide accuracy. Particular attention is paid to ensuring alignment to one assessment objective in the Colorado Model Content Standards. Participants identify their alignment choice, as well as the depth of knowledge (DOK) required by each assessment item. Any inconsistencies are documented. This sometimes results in suggestions for revision to assure better alignment. This process is essential in building

evidence of content validity. In addition, these committees review items for plain language and universal design. This input is collected and used to further revise and refine items to contribute to the item pool from which operational tests are constructed.

In addition to new multiple choice and constructed response items, previously used multiple choice items are used as an anchor set in order for the tests to be equated from year to year. Within each grade and content area form of the assessment, 17 to 25 items are selected in anchors across all of the content standards. Other previously used multiple choice and constructed response items comprise the remainder of the test form. Out of the 60 items in each test form, 45 items are multiple choice, and 15 items are constructed response items ranging from two to four points. Only 25% of each assessment is made up of new items.

CSAP Scoring

The CSAP is scored using Item Response Theory (IRT) as a means by which both multiple choice and constructed response items can be scored using the same scale. CTB's PARDUX software is used to implement the IRT model. A three parameter logistic model is used for all multiple choice items. A two parameter partial credit model is used for constructed response items (CTB McGraw-Hill, 2007). IRT models probability of an individual with a certain ability level answering an item

correctly. Item characteristic curves are then generated that represent three parameters, the item discrimination (slope), the item difficulty which also relates to student ability, and also the probability that a student with no ability would answer an item correctly, such as might happen when a student is guessing on multiple choice items. IRT is based on the assumption that any variance in item responses is due to one latent trait. IRT assumes unidimensionality of the items (Lord, 1980). P-values are used from classical test theory to represent item difficulty throughout this study.

Colorado Revisions for PL and UD in 2007

In 2006, after new items had been reviewed for content validity, bias, accessibility and appropriateness, further work was undertaken by CTB content experts as well as assessment experts from the Colorado Department of Education to incorporate universal design principles and plain language revisions. In terms of universal design, Colorado incorporated the following aspects in item revision: confirming direct match of assessment item to one assessment objective, simplifying directions, building consistency across items and assessments, using simplified fonts and increasing white space within the assessment documents. Specifically for the plain language revisions, Colorado paid particular attention to

accessibility in vocabulary, style of discourse, sentence structure and format of items in order to clarify the item content.

Procedure

A post-hoc analysis of CSAP data from 2005 and 2007 was conducted to explore the research questions. The 2005 and 2007 forms were equated using anchor items. Items revised for plain language in grades 3-10 mathematics CSAP were categorized by type of linguistic revision using five categories of revision including context, graphics, vocabulary/wording, sentence structure, and format/style (Sato, 2007). Item statistics were generated and then examined for any change in DIF, item difficulty or item discrimination between the administrations due to plain language revision. The unrevised items for each assessment were examined for significant differences in item difficulty, discrimination and measure of DIF. Depth of Knowledge (DOK) for items in each category of revision was evaluated to determine if items changed DOK rating due to plain language revision.

The numbers of items included in the study are outlined by grade in Table 2. Every grade (3-10) included items revised for plain language as well as unrevised items that were used as controls.

Table 2.
Number of Items Analyzed by Grade and by Item Type

		Control Item	Revised item	Group Total
GRADE LEVEL	3	13	11	24
	4	17	12	29
	5	16	27	43
	6	8	34	42
	7	10	35	45
	8	7	38	45
	9	5	35	40
	10	7	28	35
Group Total		83	220	303

Categorization of Items

Items were categorized based on type of plain language or universal design revision. Sato (2007) presented five strategies for revising items for plain language based on the current research. Those were context, graphics, vocabulary/wording, sentence structure and format/style.

- Context of the item should be familiar to students without presenting cultural or linguistic bias
    - For the CSAP, items categorized as “context” had either context added to clarify the item, or in some instances extraneous context removed from the item.
- Item graphics need to be relevant and related to the intended construct; they must support a student’s understanding of the content. Consistency in labeling and format is also important

- Items include high frequency, common and familiar words; technical terms are relevant and defined if appropriate; ambiguous or multiple meaning words are avoided as are words with irregular spelling, and when used, formal proper names are appropriate and relevant
- Sentence structure maintains clear noun-pronoun relationships and grammatical structures are clear including use of present tense, and active verb forms
- Item format includes the organization of the problem and ensuring that directions are clear, long statements are shortened and the order of operations is clearly evident.

Many items included more than one type of revision. Very few items were revised for graphics or context, while many items were revised for a combination of format and vocabulary/wording revisions or format and sentence structure revisions. Table 3 identifies the predominant categories of revision by grade.

Table 3.
Revised Items by Category and Grade

		Revision Category							
		Wording	Format & Style	Sentence Structure	Graphics	Format / Sentence	Format / Wording	Other multiple	**Totals**
GRADE LEVEL	3	7	3			1			11
	4	4	4	1		1	2		12
	5	15	4	1		1	3	3	27
	6	25	2		1	1	3	2	34
	7	17	7	4		2	5		35
	8	19	7	1		3	5	3	38
	9	23	3		1	3	3	2	35
	10	17	1	1		8	1		28
Totals		127	31	8	2	20	22	10	220

Examples of items revised for each category are included in Appendix B.

Depth of Knowledge

Depth of Knowledge (DOK) for the 2005 items was documented during the Content Alignment and Validity Study. This study took place in April of 2005 with 120 participants representing classroom teachers, school administrators, curriculum experts, and higher education and business leaders. Seventy individuals worked on the math assessments grades 3-10. Along with examining DOK for the assessment frameworks and the assessment items themselves, participants aligned assessment items with assessment objectives. The data were then analyzed using the

Webb Alignment Tool for categorical concurrence, DOK consistency, range of knowledge, and balance of representation (Webb, 1997).

Depth of knowledge relates to the cognitive complexity of the assessment items and the cognitive demand placed on the student for a given item. DOK is categorized into four levels. Level one is recall, level two is demonstrating a skill/concept using two or more steps, level three is strategic thinking which requires reasoning, analyzing and an item that may have more than one answer. Level four is extended thinking, which requires not only an extended response, but an extended period of time over which the investigation is conducted (Webb, 1997). On CSAP, levels one through three are represented (Appendix B). CTB item writers determined the DOK for the 2007 items when they were written. CDE assessment staff validated the DOK assignment of items. Since neither the construct nor the assessment task of an item should change as a result of plain language, the DOK should not change with these revisions. By analyzing any changes in item difficulty, discrimination, and DIF for subgroups (particularly for ELLs) from one administration to the other the effects of plain language revision may be revealed. Additionally, the examination of DOK provides an additional perspective regarding the perception of items pre and post revision relative to cognitive complexity.

Item Difficulty

Examining item difficulty provides insight as to whether or not students of a particular subgroup (in this case ELLs) perform better on one form of the item versus another. A 1984 court ruling (Golden Rule Insurance Company and the Illinois Department of Education/ETS) specified that a difference of 0.15 or more in p-value between majority and minority subgroups was evidence the item was biased. For this study, p-values as a measure of item difficulty were examined for the 2005 and 2007 administrations. Item difficulty was calculated as part of the classical item statistics. On CSAP, item difficulty for multiple choice items is the percent of students that answered an item correctly. For constructed response items, item difficulty indicates the mean percent of the possible maximum score. Any item responses left blank were treated as incorrect responses (CTB McGraw-Hill, 2007). Both multiple choice and constructed response items were analyzed as a part of this study. Item difficulty on CSAP is derived using the IRT model and CTB's PARDUX (Burket, 1993), however p-values were used in this study as this is the common way item difficulty for the CSAP is communicated to stakeholders. One sample t-tests were used to identify any significant differences in mean item difficulty between the two administrations for each of the subgroups, with 2005 mean p-values serving as the population value.

Item Discrimination

Item discrimination is a probability that assesses the likelihood that an item is correctly answered by a higher ability student and answered incorrectly by a lower ability student. Item discrimination is one of the three parameter estimates generated from the three parameter logistic IRT model, and one of the two parameters from the two parameter model used for constructed response items The a-parameter (item discrimination) value relates to the slope of the line in the IRT model, where a high value represents a steep slope and a more sharply discriminating item, while low values represent a flatter slope. Item discrimination is independent of item difficulty. Item parameters for CSAP are generated using PARDUX. One sample t-tests were used to identify any significant differences in average item discrimination between the two administrations for the items, with 2005 means serving as the population value.

Differential Item Functioning (DIF)

Every item contains a measure of DIF, the unexpected difference between the observed and expected performance on a particular item for a particular subgroup. In other words, DIF is the measure of the degree to which members of the focal group performed better or worse than expected on a particular item. A negative value indicates an item as performing against the focal group, while a positive value indicates an item

as favoring a particular focal group. In this study, DIF was assessed across performance groups.

The technique used to identify DIF was the Linn–Harnisch procedure (1981) which uses the information provided by the IRT model, but does not require as many cases as other DIF detection methods. Other DIF methods require more than 1,000 cases in each comparison group to be considered reliable. In CTB McGraw-Hill's application of the Linn-Harnisch method, an item with an expected/observed mean difference of >0.10 or standardized measure of $\geq$ 2.85 (99% confidence interval), where the proportion of the students in the focal group perform better or worse on a particular item, is flagged as containing significant DIF and may be removed from score calculation (CTB McGraw-Hill, 2007).

The item parameters for *a* (item discrimination), *b* (item difficulty), and *c* (guessing) are estimated based on data from the total sample of valid examinees. The sample is then divided into specific subgroups. The members in each group are then sorted into deciles based on location of scale score. The expected proportion correct for each group is compared to the observed (actual) proportion correct for the group. In looking at the differences between the observed and expected performance of the focal group, DIF provides differences by ability decile within groups.

The proportion of people in decile g who are expected to answer item i correctly is where n_g is the number of examinees in decile g.

$$P_{ij} = P_{ig}(\theta) = \frac{1}{n_g} \sum_{j \in g} P_{ij}(\theta),$$

The formula to compute the proportion of students expected to answer item i correctly (over all deciles) for a group, is given by

$$P_{i.} = P_i(\theta). = \frac{\sum_{g=1}^{10} n_g P_{ig}(\theta)}{\sum_{g=1}^{10} n_g}.$$

The corresponding observed proportion correct for examinees in a decile (O_{ig}) is the number of examinees in decile g who answered item i correctly divided by the number of people in the decile (n_g). That is,

$$O_{ig} = \frac{\sum_{j \in g} u_{ij}}{n_g},$$

where u_{ij} is the dichotomous score for item i for examinee j.
The corresponding formula to compute the observed proportion answering each item correctly (over all deciles) for a complete gender group is given by

$$O_{i.} = \frac{\sum_{g=1}^{10} n_g O_{ig}}{\sum_{g=1}^{10} n_g}.$$

After the values are calculated for these variables, the difference between the observed proportion correct (for gender) and expected proportion correct can be computed. The decile group difference (D_{ig}) for observed and expected proportion correctly answering item i in decile g is

$$D_{ig} = O_{ig} - P_{ig},$$

and the overall group difference (D_i) between observed and expected proportion correct for item i in the complete group (over all deciles) is

$$D_{i\cdot} = O_{i\cdot} - P_{i\cdot} \ .$$

(CTB McGraw-Hill, 2007, pp. 77-78).

These indices indicate to what degree members of a particular subgroup within a decile perform better or worse on a specific item based on the performance of other members of the entire subgroup. Differences for decile groups provide an index for each of the ten regions on the score (θ) scale. The decile group difference (D_{ig}) can be either positive or negative. Use of the decile group differences as well as the overall group difference allows one to detect items that give a large positive difference in one range of θ and a large negative difference in another range of θ, yet have a small overall difference (CTB McGraw-Hill, 2007, pp. 77-78).

While items included in this study were not flagged for DIF in 2005, content editors and item reviewers considered the format and/or linguistic complexity of the item to have the potential to introduce barriers for students, thus requiring revision prior to the next administration. While different DIF detection methods may yield varying levels of sensitivity in identifying DIF (Abedi, 2008; Ockey, 2007), using the same method should not result in tremendous variance in the measure of DIF.

After DIF for each item from 2005 and 2007 for each subgroup was calculated, these values were compared at each grade level using one sample t-tests.

Anchor Items

Unrevised anchor items along with additional core selection computation items for each test were used as controls in this study. The control items were used to determine if any change between the two years could be attributed to an overall change in performance versus a change in performance due to plain language revision of the items themselves. Purely computational items repeated across test forms were included as a part of the control set as they had no linguistic attributes and could represent differences in math ability without interference of linguistic structures.

Setting

Colorado is very diverse geographically. The Eastern half of the state is arid, high desert, and primarily agricultural. The Denver Metro area and other metropolitan areas such as Colorado Springs, and Fort Collins lie in the center of the state, along the 1-25 corridor and at the foothills of the Rocky Mountains. The Mountain areas hold another set of diverse communities, from small rural towns to exclusive ski resort areas. The

western slope of Colorado includes towns like Grand Junction, where agriculture and ranching are prevalent.

Population

It is required by Colorado State Statute (C.R.S. 22-7-409) that all students enrolled in a Colorado public school in grades 3-10 take the CSAP. Colorado, therefore, has about 98% of the students taking the assessment each year. Approximately 450,000 students in grades 3-10 took the CSAP in 2005 (Table 4). For the 2007 CSAP, roughly 460,000 students participated in grades 3-10 (Table 5).

Table 4.
2005 CSAP Participants Demographic Profile

2005	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10
Total N	55539	55399	55910	57072	58363	58069	59906	53562
Female	27132	26934	27264	27807	28554	28317	29151	26325
Male	28388	28465	28644	29264	29807	29745	30752	27236
American Indian/Alaska Native	670	628	675	697	746	714	701	610
Asian/Pacific Islander	1896	1895	1727	1652	1765	1706	1741	1718
Black (not Hispanic)	3371	3352	3433	3537	3487	3436	3733	2952
Hispanic	15630	15173	15206	15251	14941	14408	14496	11081
White (not Hispanic)	33953	34351	34865	35934	37423	37799	39233	37198
IEP	5762	6048	6024	5817	5673	5447	5426	4546
504 Plan	170	234	328	471	532	624	583	625
NEP	1757	1270	1121	1162	1309	1379	1449	1074
LEP	5109	4519	3883	3017	2504	2120	1897	1366
FEP	2316	2851	3412	3817	3765	3475	3306	2679
Free and Reduced Lunch Eligible	21341	20667	20635	20320	19361	18177	15911	11549

Table 5.
2007 CSAP Participants Demographic Profile

2007	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10
Total N	58080	56799	56958	56711	57153	58162	61012	56416
Female	28463	27884	27812	27580	27877	28341	29852	27643
Male	29607	28908	29138	29124	29273	29821	31150	28761
American Indian/Alaska Native	673	640	705	642	720	713	758	642
Asian/Pacific Islander	2086	2069	1979	1996	1851	1747	1875	1806
Black (not Hispanic)	3547	3252	3443	3392	3514	3607	3838	3437
Hispanic	17347	16369	15937	15481	15486	15373	15800	13173
White (not Hispanic)	34422	34465	34888	35195	35581	36721	38738	37354
IEP	5826	5785	5895	5776	5501	5158	5175	4527
504 Plan	183	321	361	516	580	722	737	776
NEP	3121	1842	1724	1303	1148	1000	1254	1023
LEP	5713	5461	4622	3859	3537	3155	2908	2360
FEP	2018	2912	3408	3846	4038	4128	3938	3150
Free and Reduced Lunch Eligible	22936	21580	21236	20374	20163	19650	17833	14477

Subgroups

For this study, both ELL and non-ELLs were groups used to examine item performance. However, as Ockey (2007) found, examining ELL versus non-ELL performance is not sufficient, which may be especially important when investigating for DIF, as there is such a tremendous variance of language proficiency in the ELL subgroup. For that reason, this study examined item performance for NEPs, LEPs, and FEPs as well as for native English speakers (non-ELLs).

Data Analysis

Mean DOK values for control items and plain language revised items were calculated for 2005 and 2007. Chi square was used to examine the association between the 2005 and 2007 ratings.

Item discrimination values, as measured by the a-parameter were calculated for control items and plain language revised items for 2005 and 2007. One sample t-tests were used to evaluate the differences in item discrimination between 2005 and 2007 for both control items and plain language revised items, using the 2005 means as the population values.

Item difficulty, represented by p-value, was calculated for both years of administration (2005 and 2007) for the three ELL categories (NEP, LEP and FEP) as well as non-ELLs for each item using CTB's PARDUX (Burket, 1993).

Item difficulty, as represented by p-value, is the proportion of students responding correctly to an item. A one sample t-test was used to evaluate the differences in item difficulty for each group of ELLs as well as non-ELLs on both the control and the plain language revised items between 2005 and 2007 in grades 3-10.

DIF was calculated for both years of administration (2005 and 2007) for the three ELL categories (NEP, LEP and FEP) as well as non-ELLs by performance group for each item using CTB's PARDUX (Burket, 1993). A one sample t-test was used to evaluate the differences in DIF for each group of ELLs as well as non-ELLs on both the control and the plain language revised items between 2005 and 2007 in grades 3-10.

Additionally, items were categorized into type of linguistic change, "vocabulary/wording", "sentence structure", "format/style" and combined categories of "vocabulary/wording and format", "sentence structure and format" as well as an "other multiple" category. Item difficulty and DIF measurements were used to examine the differences between these groups of revised items between 2005 and 2007. A one-way analysis of variance was used to test if the type of revision resulted in a significant difference in item difficulty or difference in DIF between 2005 and 2007, as well as Tukey's post hoc analysis to compare the revision categories.

Chapter 4

RESULTS

This chapter reports the results of data analyses addressing the following research questions:

Research Questions

1. Does the cognitive complexity (DOK) of items change after plain language revisions?
2. Is item difficulty impacted by plain language revision?
3. Is item discrimination impacted by plain language revision?
4. Do CSAP mathematics items revised for plain language provide increased access to assessment items by reducing differences on individual item performance as measured by Differential Item Functioning (DIF) for all groups of ELLs (NEP, LEP and FEP)?
5. Are there certain plain language revisions (context, graphics, vocabulary/wording, sentence structure, and format/style) of mathematics assessment items on CSAP that reduce DIF and effect difference in item difficulty?

Depth of Knowledge

For the first question, “Does the cognitive complexity (DOK) of items change after plain language revisions?” Chi-square was used to examine the change in DOK level classification between 2005 and 2007 for both control and revised items. For the control items rated level one in 2005, five fewer were rated the same way in 2007. Ninety percent of items retained level one rating. Two more items were rated a DOK of 2, while 93% of items retained the level two rating. Three items that had been rated a level one DOK in 2005 were rated a level three in 2007, in this level, 40% of items stayed a level three (Table 6).

Table 6.
DOK for Control Items in 2005 and 2007

	Observed N 2005	Observed N 2007	Change
DOK Level 1	50	45	-5
DOK Level 2	31	33	+2
DOK Level 3	2	5	+3
Total	*83*	*83*	
Chi-square	42.241	30.458	
Df	2	2	
Asymp. Sig.	<.001	<.001	

For the revised items, ten items that had been rated DOK Level 1 in 2005 were rated DOK Level 3 in 2007. Only 75% of items retained a level one rating, while 85% of level three items retained a level three rating. The items were rated DOK Level 2 in 2005 and 2007 (Table 7).

Table 7.
DOK for Revised Items in 2005 and 2007

	Observed N 2005	Observed N 2007	Change
DOK Level 1	40	30	-10
DOK Level 2	124	124	0
DOK Level 3	56	66	+10
Total	220	220	
Chi-square	54.255	61.345	
Df	2	2	
Asymp. Sig.	.001	.001	

The chi-square test indicates a significant association between the 2005 and 2007 ratings. The null hypothesis was not rejected since there was a significant association between 2005 and 2007 ratings.

Item Difficulty

In examining whether item difficulty was impacted by plain language revision, item difficulty values for 2005 and 2007 were examined by grade level and language proficiency level. Within each grade level, item difficulty values for each language proficiency subgroup (NEP, LEP, FEP and Non ELL) were used. Means for control and revised items for 2005 and 2007 were calculated separately. One sample t-tests were used to examine the difference in item difficulty between 2005 and 2007 for both control and plain language revised items within each language proficiency subgroup for each grade level. While some skewness existed for each of the data distributions, it was within the range of normality and the data could be used for this analysis (Box, 1953).

Grade 3 Item Difficulty Differences

Third grade item difficulties appeared to have a relatively normal distribution. The 2007 item difficulties for the LEP group had a slight negative skew; NEP in both 2005 and 2007 exhibited a positive skew, while both Non ELL and FEP data were skewed negatively (Table 8). The mean 2005 item difficulty values for each subgroup were used as the population means in the one sample t-test (Table 8).

Table 8.
Grade 3 Revised and Control Items by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2005 Std. Deviation	Skewness 2005	2007 Mean	2007 Std. Deviation	Skewness 2007
NEP *(Non English Proficient)*	Control	13	.64812	.167040	.442	.64172	.169923	.224
	Revised	11	.52273	.207224	-.662	.50510	.209159	.209
LEP *(Limited English Proficient)*	Control	13	.71348	.161145	.048	.78182	.135651	-.511
	Revised	11	.59182	.228954	-.830	.66175	.198910	-.633
FEP *(Fluent English Proficient)*	Control	13	.84835	.113750	-.930	.87356	.098153	-1.230
	Revised	11	.74008	.199486	-1.235	.76964	.159329	-.854
Non ELL *(Non English Language Learner)*	Control	13	.85406	.108261	-.994	.86913	.098297	-1.123
	Revised	11	.75209	.188218	-1.190	.77029	.155728	-.856

In grade 3, item difficulty was not statistically significantly different for any language proficiency subgroup between 2005 and 2007 (Table 9). The null hypothesis for 3rd grade was not rejected.

Table 9.
One Sample t-Tests for Grade 3 Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Mean Difference
NEP *(Non English Proficient)*	Control	.64812	-.136	12	.894	-.006402
	Revised	.52273	-.280	10	.785	-.017633
LEP *(Limited English Proficient)*	Control	.71348	1.816	12	.094	.068341
	Revised	.59182	1.166	10	.271	.069933
FEP *(Fluent English Proficient)*	Control	.84835	.926	12	.373	.025212
	Revised	.74008	.615	10	.552	.029564
Non ELL *(Non English Language Learner)*	Control	.85406	.553	12	.590	.015073
	Revised	.75209	.388	10	.706	.018204

Grade 4 Item Difficulty Differences

Within fourth grade, item difficulties appeared to have a relatively normal distribution. For LEP, the data exhibits a slight positive skew in both 2005 and 2007, although it would still be considered within a normal distribution. Likewise, NEP also exhibit a positive skew, and there are some outliers in the 2007 NEP data set that additionally contribute to the positive skew. Control items for all language proficiency subgroups exhibit some negative skew. The mean 2005 item difficulty values for each subgroup were used as the population mean in the one sample t-test (Table 10).

Table 10.
Grade 4 Revised and Control Items by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2005 Std. Deviation	Skewness 2005	2007 Mean	2007 Std. Deviation	Skewness 2007
NEP *(Non English Proficient)*	Control	17	.64715	.176703	-.574	.64167	.169200	-.676
	Revised	12	.48028	.220834	.820	.46853	.238760	.901
LEP *(Limited English Proficient)*	Control	17	.69701	.202577	-1.107	.75459	.181090	-1.157
	Revised	12	.54046	.226440	.526	.61587	.208561	.512
FEP *(Fluent English Proficient)*	Control	17	.81670	.168764	-1.453	.84499	.145934	-1.475
	Revised	12	.71205	.162939	.061	.76189	.141974	.107
Non ELL *(Non English Language Learner)*	Control	17	.82948	.155834	-1.579	.84525	.142412	-1.516
	Revised	12	.74414	.145649	-.131	.76917	.133735	.105

In grade 4, item difficulty was not statistically significantly different for any language proficiency subgroup between 2005 and 2007 (Table 11). The null hypothesis for 4th grade was not rejected.

Table 11.
One Sample t-Tests for Grade 4 Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Mean Difference
NEP *(Non English Proficient)*	Control	.64715	-.134	16	.895	-.005480
	Revised	.48028	-.171	11	.868	-.011754
LEP *(Limited English Proficient)*	Control	.69701	1.311	16	.208	.057576
	Revised	.54046	1.252	11	.236	.075408
FEP *(Fluent English Proficient)*	Control	.81670	.799	16	.436	.028292
	Revised	.71205	1.216	11	.249	.049841
Non ELL *(Non English Language Learner)*	Control	.82948	.457	16	.654	.015774
	Revised	.74414	.648	11	.530	.025032

Grade 5 Item Difficulties

For fifth grade, item difficulty appeared to have a relatively normal distribution. For all language proficiency groups there was a slight negative skew for control items, as compared to a slight positive skew for NEP and LEP on the revised items. The mean 2005 item difficulty values for each subgroup were used as the population means in the one sample t-test (Table 12).

Table 12.
Grade 5 Revised and Control Items by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2005 Std. Deviation	Skewness 2005	2007 Mean	2007 Std. Deviation	Skewness 2007
NEP *(Non English Proficient)*	Control	16	.59075	.213940	-.876	.52955	.210775	-.571
	Revised	27	.48349	.131559	.995	.42385	.142867	.979
LEP *(Limited English Proficient)*	Control	16	.63351	.218883	-1.070	.66733	.215418	-1.133
	Revised	27	.54584	.137921	.815	.59783	.141633	.632
FEP *(Fluent English Proficient)*	Control	16	.76315	.200751	-1.715	.78411	.182683	-1.680
	Revised	27	.70574	.121808	.272	.73774	.116718	.109
Non ELL *(Non English Language Learner)*	Control	16	.77781	.185712	-1.827	.79248	.167674	-1.760
	Revised	27	.72992	.108893	.246	.74700	.108969	.115

In grade 5, item difficulty was statistically significantly different for LEP on revised items between 2005 and 2007, $p = .039$ with an effect size of 0.626756 (Table 13). There was no significant difference for LEP on the control items. Relative to item difficulty for LEP in 5^{th} grade, the null

hypothesis was rejected. The null hypothesis was not rejected for any other language proficiency levels.

Table 13.
One Sample t-Tests for Grade 5 Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Effect Size (*r*)	Mean Difference
NEP *(Non English Proficient)*	Control	.59075	-1.161	15	.264		-.061197
	Revised	.73774	.001	26	1.000		.000002
LEP *(Limited English Proficient)*	Control	.63351	.628	15	.540		.033815
	Revised	.48349	-2.169	26	.039*	0.626756	-.059639
FEP *(Fluent English Proficient)*	Control	.76315	.459	15	.653		.020956
	Revised	.70574	1.425	26	.166		.032002
Non ELL *(Non English Language Learner)*	Control	.77781	.350	15	.731		.014671
	Revised	.72992	.814	26	.423		.017069

*p** = <.05

Grade 6 Item Difficulties

In sixth grade, item difficulty appeared to have a relatively normal distribution. All language groups, however, did have some outliers on both ends of the distribution for control items. The means of 2005 item difficulty values for each subgroup were used as the population means in the one sample t-test (Table 14).

Table 14.
Grade 6 Revised and Control Items by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2005 Std. Deviation	Skewness 2005	2007 Mean	2007 Std. Deviation	Skewness 2007
NEP *(Non English Proficient)*	Control	8	.39552	.213853	.130	.37404	.210643	.579
	Revised	33	.44311	.166617	.238	.42043	.149576	.224
LEP *(Limited English Proficient)*	Control	8	.41816	.238698	.215	.47088	.230575	.210
	Revised	33	.48777	.176910	.124	.52769	.171866	.087
FEP *(Fluent English Proficient)*	Control	8	.54612	.238884	-.048	.61340	.210133	-.088
	Revised	33	.64944	.173879	-.091	.61420	.165951	-.243
Non ELL *(Non English Language Learner)*	Control	8	.62565	.218514	-.339	.67228	.184841	-.263
	Revised	33	.67326	.160582	-.267	.69273	.156297	-.364

In grade 6, item difficulty was not statistically significantly different for any language proficiency subgroup between 2005 and 2007 (Table 15). The null hypothesis for 6th grade was not rejected.

Table 15.
One Sample t-Tests for Grade 6 Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Mean Difference
NEP *(Non English Proficient)*	Control	.39552	-.288	7	.781	-.021478
	Revised	.44311	-.871	32	.390	-.022681
LEP *(Limited English Proficient)*	Control	.41816	.647	7	.538	.052715
	Revised	.48777	1.334	32	.192	.039917
FEP *(Fluent English Proficient)*	Control	.54612	.906	7	.395	.067284
	Revised	.64944	.001	32	1.000	.000003
Non ELL *(Non English Language Learner)*	Control	.62565	.713	7	.499	.046625
	Revised	.67326	.715	32	.480	.019466

Grade 7 Item Difficulties

Seventh grade item difficulty appeared to have a relatively normal distribution. All subgroups tended toward a slight positive skew except revised items for Non ELL in 2007, which exhibited a slight negative skew. The mean 2005 item difficulty values for each subgroup were used as the population mean in the one sample t-test (Table 16).

Table 16.
Grade 7 Revised and Control Items by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2005 Std. Deviation	Skewness 2005	2007 Mean	2007 Std. Deviation	Skewness 2007
NEP *(Non English Proficient)*	Control	10	.38137	.221386	1.031	.39600	.220443	.841
	Revised	35	.34396	.192670	.701	.33285	.192414	.618
LEP *(Limited English Proficient)*	Control	10	.41692	.238832	.714	.44613	.238503	.535
	Revised	35	.36997	.206305	.726	.40502	.214808	.637
FEP *(Fluent English Proficient)*	Control	10	.51172	.222438	.412	.53950	.217880	.309
	Revised	35	.47343	.208205	.446	.50431	.215632	.239
Non ELL *(Non English Language Learner)*	Control	10	.60328	.194755	.155	.62725	.186198	.076
	Revised	35	.54722	.199234	.108	.57131	.201273	-.008

In grade 7, item difficulty was not statistically significantly different for any language proficiency subgroup between 2005 and 2007 (Table 17). The null hypothesis for 7th grade was not rejected.

Table 17.
One Sample t-Tests for Grade 7 Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Mean Difference
NEP *(Non English Proficient)*	Control	.38137	.210	9	.838	.014625
	Revised	.34396	-.342	34	.735	-.011112
LEP *(Limited English Proficient)*	Control	.41692	.387	9	.708	.029209
	Revised	.36997	.965	34	.341	.035055
FEP *(Fluent English Proficient)*	Control	.53950	.001	9	1.000	-.000005
	Revised	.47343	.847	34	.403	.030873
Non ELL *(Non English Language Learner)*	Control	.60328	.407	9	.693	.023973
	Revised	.57131	.001	34	1.000	-.000010

Grade 8 Item Difficulties

Eighth grade item difficulty also appeared to have a relatively normal distribution. There is a slight positive skew for the NEP data, and a slight negative skew for Non ELLs due to some outliers. The mean 2005 item difficulty values for each subgroup were used as the population mean in the one sample t-test (Table 18).

Table 18.
Grade 8 Revised and Control Items by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2005Std. Deviation	Skewness 2005	2007 Mean	2007 Std. Deviation	Skewness 2007
NEP *(Non English Proficient)*	Control	7	.31416	.2100079	.704	.31464	.183075	.697
	Revised	38	.28985	.183473	1.074	.26265	.179660	.981
LEP *(Limited English Proficient)*	Control	7	.35746	.236097	.401	.36652	.234780	.589
	Revised	38	.32945	.198870	.779	.34156	.210026	.710
FEP *(Fluent English Proficient)*	Control	7	.46114	.261481	.015	.45803	.253453	-.098
	Revised	38	.42946	.213194	.367	.43834	.218353	.270
Non ELL *(Non English Language Learner)*	Control	7	.54290	.253466	-.343	.54687	.250137	-.457
	Revised	38	.49783	.209547	.078	.51134	.211732	.001

In grade 8, item difficulty was not statistically significantly different for any language proficiency subgroup between 2005 and 2007 (Table 19). The null hypothesis for 8th grade was not rejected.

Table 19.
One Sample t-Tests for Grade 8 Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Mean Difference
NEP *(Non English Proficient)*	Control	.31416	.007	6	.995	.000483
	Revised	.28985	-.933	37	.357	-.027198
LEP *(Limited English Proficient)*	Control	.35746	.102	6	.922	.009056
	Revised	.32945	.355	37	.724	.012111
FEP *(Fluent English Proficient)*	Control	.46114	-.032	6	.975	-.003113
	Revised	.42946	.251	37	.804	.008877
Non ELL *(Non English Language Learner)*	Control	.54290	.042	6	.968	.003973
	Revised	.49783	.393	37	.696	.013512

Grade 9 Item Difficulties

Ninth grade item difficulty in 2005 and 2007 across language proficiency levels had a relatively normal distribution. Data for NEP and LEP do show a slight positive skew. The means of 2005 item difficulty values for each subgroup were used as the test values in the one sample t-test (Table 20).

Table 20.
Grade 9 Revised and Control Items by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2005 Std. Deviation	Skewness 2005	2007 Mean	2007 Std. Deviation	Skewness 2007
NEP *(Non English Proficient)*	Control	5	.25319	.125939	.845	.25507	.117832	.878
	Revised	35	.21646	.178687	1.144	.20455	.171973	1.237
LEP *(Limited English Proficient)*	Control	5	.28103	.129307	.605	.29248	.129718	.844
	Revised	35	.25879	.201503	1.227	.26845	.202445	1.207
FEP *(Fluent English Proficient)*	Control	5	.34435	.153391	.234	.35479	.167938	.467
	Revised	35	.33908	.207625	1.076	.35252	.211256	.935
Non ELL *(Non English Language Learner)*	Control	5	.41767	.1711211	-.486	.43124	.188019	.110
	Revised	35	.41075	.202615	.838	.42603	.200571	.700

In grade 9, item difficulty was not statistically significantly different for any language proficiency subgroup between 2005 and 2007 (Table 21). The null hypothesis for 9th grade was not rejected.

Table 21.
One Sample t-Tests for Grade 9 Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Mean Difference
NEP *(Non English Proficient)*	Control	.25319	.036	4	.973	.001882
	Revised	.21646	-.410	34	.685	-.011909
LEP *(Limited English Proficient)*	Control	.28103	.197	4	.853	.011454
	Revised	.25879	.282	34	.779	.009665
FEP *(Fluent English Proficient)*	Control	.34435	.139	4	.896	.010444
	Revised	.33908	.376	34	.709	.013444
Non ELL *(Non English Language Learner)*	Control	.41767	.161	4	.880	.013572
	Revised	.42603	.001	34	1.000	-.000002

Grade 10 Item Difficulties

Tenth grade item difficulty in 2005 and 2007 across language proficiency levels had a relatively normal distribution. There is a slight positive skew to the data for all language proficiency subgroups. The mean 2005 item difficulty values for each subgroup were used as the population mean in the one sample t-test (Table 22).

Table 22.
Grade 10 Revised and Control Items by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2005 Std. Deviation	Skewness 2005	2007 Mean	2007 Std. Deviation	Skewness 2007
NEP *(Non English Proficient)*	Control	7	.31522	.114684	.653	.28799	.082938	.598
	Revised	28	.20641	.143430	.800	.19434	.134431	.533
LEP *(Limited English Proficient)*	Control	7	.24128	.129221	.743	.34756	.127595	.848
	Revised	28	.24128	.153838	.932	.25506	.155445	.601
FEP *(Fluent English Proficient)*	Control	7	.43050	.154393	.405	.43129	.151928	.704
	Revised	28	.31008	.159034	.747	.32936	.168039	.608
Non ELL *(Non English Language Learner)*	Control	7	.51087	.152328	.147	.51770	.153748	.395
	Revised	28	.38871	.171621	.699	.40265	.174092	.539

In grade 10, item difficulty was not statistically significantly different for any language proficiency subgroup between 2005 and 2007 (Table 23). The null hypothesis for 10th grade was not rejected.

Table 23.

One Sample t-Tests for Grade 10 Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Mean Difference
NEP *(Non English Proficient)*	Control	.31522	-.869	6	.418	-.027233
	Revised	.20641	-.475	27	.639	-.012071
LEP *(Limited English Proficient)*	Control	.24128	.469	27	.643	.013783
	Revised	.24128	.469	27	.643	.013783
FEP *(Fluent English Proficient)*	Control	.43050	.014	6	.989	.000793
	Revised	.31008	.607	27	.549	.019278
Non ELL *(Non English Language Learner)*	Control	.51087	.118	6	.910	.006830
	Revised	.38871	.424	27	.675	.013937

Within each individual grade level, item difficulty between 2005 and 2007 did not change for any language proficiency subgroups except for 5th grade, where the item difficulty for revised items within the LEP subgroup was significantly different than 2005 item difficulty, with control group item difficulty in the LEP subgroup not changing significantly.

Combined Grades Results

When examining item difficulty for language proficiency subgroups for all grades combined, the mean 2005 item difficulty value for the language proficiency subgroups were used as the population mean in the one sample t-test (Table 24).

Table 24.
Combined Grades Item Difficulty for Revised and Control Items by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2007 Mean	2007 Std. Deviation
NEP *(Non English Proficient)*	Control	83	.50188	.48397	.226369
	Revised	219	.34520	.32296	.198166
LEP *(Limited English Proficient)*	Control	83	.54288	.58263	.256039
	Revised	219	.38864	.52433	.242444
FEP *(Fluent English Proficient)*	Control	83	.65750	.68155	.246751
	Revised	219	.52433	.41971	.232010
Non ELL *(Non English Language Learner)*	Control	83	.70062	.71815	.213251
	Revised	219	.55917	.57691	.220946

For all grades combined, the 2007 revised items in the LEP group were statistically significantly different, p=.049 with an effect size (r) of 0.36798 (Table 25).

Table 25.
One Sample t-Tests for Combined Grades Item Difficulty for Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Effect Size (*r*)	Mean Difference
NEP *(Non English Proficient)*	Control	.50188	-.721	82	.473		-.017909
	Revised	.34520	-1.661	218	.098		-.022242
LEP *(Limited English Proficient)*	Control	.54288	1.414	82	.161		.039749
	Revised	.38864	1.982	218	.049*	0.36798	.031075
FEP *(Fluent English Proficient)*	Control	.68155	.001	82	1.000		-.000005
	Revised	.52433	.001	218	1.000		-.000004
Non ELL *(Non English Language Learner)*	Control	.70062	.749	82	.456		.017533
	Revised	.55917	1.188	218	.236		.017743

*p**<.05

In looking at the mean item difficulty in 2005 and 2007 on both control and revised items for each of the language proficiency levels, the LEP, FEP and Non ELL subgroups performed better in 2007 on both control and revised items. As item difficulty represents the proportion of students that answer a given item correctly, NEP students did not perform as well on the control and revised items in 2007 compared with 2005 (Tables 26 and 27).

Table 26.
Minimum, Maximum, Mean, and Standard Deviation of Control Items Combined Grades Item Difficulty

	N	Minimum	Maximum	Mean	Std. Deviation
2005 Item Difficulty for LEPs	83	.075	.965	.54288	.251104
2007 Item Difficulty for LEPs	83	.078	.963	.58263	.256039
2005 Item Difficulty for NEPs	83	.094	.931	.50188	.232759
2007 Item Difficulty for NEPs	83	.099	.933	.48397	.226369
2005 Item Difficulty for Non ELLs	83	.162	.984	.70062	.221248
2007 Item Difficulty for Non ELLs	83	.172	.981	.71815	.213251
2005 Item Difficulty for FEPs	83	.117	.986	.65750	.252386
2007 Item Difficulty for FEPs	83	.101	.984	.68155	.246751
Valid N (listwise)	83				

Table 27.
Minimum, Maximum, Mean, and Standard Deviation of Revised Items Combined Grades Item Difficulty

	N	Minimum	Maximum	Mean	Std. Deviation
2005 Item Difficulty for LEPs	219	.023	.942	.38864	.220299
2007 Item Difficulty for LEPs	219	.046	.965	.41971	.232010
2005 Item Difficulty for NEPs	219	.027	.913	.34520	.204167
2007 Item Difficulty for NEPs	219	.013	.916	.32296	.198166
2005 Item Difficulty for Non ELLs	219	.109	.975	.55917	.221086
2007 Item Difficulty for Non ELLs	219	.111	.981	.57691	.220946
2005 Item Difficulty for FEPs	219	.076	.972	.49977	.236803
2007 Item Difficulty for FEPs	219	.079	.981	.52433	.242444
Valid N (listwise)	219				

Item Discrimination

To examine the question “Is item discrimination impacted by plain language revision?” item discrimination (a-parameter) values for 2005 and 2007 were examined. Means for control and revised items for 2005 and 2007 were calculated as seen in Table 28. A one sample t-test was used to examine the difference in item discrimination between 2005 and 2007 for both control and plain language revised items.

Table 28.
Minimum, Maximum, Mean, and Standard Deviation of Item Discrimination for Control and Revised Items by Year

	Type of Item	N	Minimum	Maximum	Mean	Std. Deviation
2005 Item Discrimination	Control	83	.2700	1.6300	.859347	.2858990
	Revised	220	.2300	2.2400	.979000	.3459536
2007 Item Discrimination	Control	83	.2491	1.9943	.864485	.3263286
	Revised	220	.2282	1.8250	.956568	.3286850

For the control items, the mean item discrimination from 2005 (.859347) was used as the population mean. The 2007 mean item discrimination was not statistically significantly different from the 2005 mean item discrimination (Table 29).

Table 29.
One Sample t-test for Item Discrimination for Revised and Control Items

	Type of Item	Test Value	t	df	Sig. (2-tailed)	Mean Difference
2007 Item Discrimination	Control	.859347	.143	82	.886	.0051383
	Revised	.979000	-1.012	219	.313	-.0224317

For the revised items, the mean item discrimination value from 2005 (.979000) was used as the population mean. The 2007 mean item discrimination was not statistically significantly different from the 2005 mean item discrimination (Table 29). Based on the mean discrimination parameter value for revised items in both 2005 (.979000) and 2007 (.956568), the revised items became less discriminating after revision for plain language (Table 28), while the control items became more discriminating (Table 28). The null hypothesis was not rejected.

Differential Item Functioning (DIF)

DIF is frequently used in *post hoc* item analyses to determine if any items presented bias to test examinees. In exploring whether CSAP mathematics items revised for plain language provide increased access to assessment items by reducing differences on individual item performance as measured by Differential Item Functioning (DIF) for all groups of ELLs (NEP, LEP and FEP), DIF values for 2005 and 2007 within ability

groupings within each grade level for each language proficiency subgroup (NEP, LEP, FEP and Non-ELL) were used.

Means of DIF values (the difference between observed and expected performance) for control and revised items for 2005 and 2007 were calculated. One sample t-tests were used to examine the difference in DIF between 2005 and 2007 for both control and plain language revised items for each grade level. The means of 2005 DIF values for each subgroup were used as the population means in the one sample t-tests.

Grade 3 DIF Results

In grade 3, mean DIF values were not statistically significantly different between 2005 and 2007 for any language proficiency subgroup for either control or revised items (Table 31). DIF values on revised items for FEPs and NEPs as well as DIF values on control items for LEPs changed from slightly disfavoring those subgroups to slightly favoring those subgroups (Table 32).

Table 30.
Grade 3 DIF Differences for Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	N	Mean	Std. Deviation	Std. Error Mean
NEP *(Non English Proficient)*	Control	13	.002154	.0443506	.0123006
	Revised	11	.000727	.0612145	.0184569
LEP *(Limited English Proficient)*	Control	13	.000615	.0251149	.0069656
	Revised	11	.007273	.0239462	.0072200
FEP *(Fluent English Proficient)*	Control	13	.001154	.0134465	.0037294
	Revised	11	.004636	.0186937	.0056364
Non ELL *(Non English Language Learner)*	Control	13	-.000615	.0083918	.0023275
	Revised	11	.000273	.0120257	.0036259

Table 31.
One Sample t-Tests for Grade 3 DIF Differences

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Mean Difference
NEP *(Non English Proficient)*	Control	.006154	-.325	12	.751	-.0040002
	Revised	-.004818	.300	10	.770	.0055453
LEP *(Limited English Proficient)*	Control	-.002769	.486	12	.636	.0033844
	Revised	.002455	.667	10	.520	.0048177
FEP *(Fluent English Proficient)*	Control	.000615	.144	12	.888	.0005388
	Revised	.002636	1.290	10	.226	.0072724
Non ELL *(Non English Language Learner)*	Control	-.000308	-.132	12	.897	-.0003067
	Revised	.001818	-.426	10	.679	-.0015453

Table 32.
Grade 3 Revised and Control Items DIF Means by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2005 Std. Deviation	2007 Mean	2007 Std. Deviation
NEP *(Non English Proficient)*	Control	13	.006154	.0525909	.002154	.0443506
	Revised	11	-.004818	.0554163	.000727	.0612145
LEP *(Limited English Proficient)*	Control	13	-.002769	.0359355	.000615	.0251149
	Revised	11	.002455	.0425802	.007273	.0239462
FEP *(Fluent English Proficient)*	Control	13	.000615	.0167309	.001154	.0134465
	Revised	11	-.002636	.0207859	.004636	.0186937
Non ELL *(Non English Language Learner)*	Control	13	-.000308	.0116432	-.000615	.0083918
	Revised	11	.001818	.0114614	.000273	.0120257

Grade 4 DIF Results

In grade 4, DIF for control items for FEP was statistically significantly different at p =.017 (Table 34). Control items for LEP and FEP went from slightly disfavoring those subgroups to slightly favoring, while revised items for NEP and LEP went from slightly favoring to slightly disfavoring in 2007 (Table 35).

Table 33.
Grade 4 DIF Differences for Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	N	Mean	Std. Deviation	Std. Error Mean
NEP *(Non English Proficient)*	Control	17	.008294	.0583511	.0141522
	Revised	12	-.012167	.1149180	.0331740
LEP *(Limited English Proficient)*	Control	17	.006941	.0185690	.0045037
	Revised	12	-.014583	.0695511	.0200777
FEP *(Fluent English Proficient)*	Control	17	.003647	.0115214	.0027943
	Revised	12	.004000	.0438551	.0126599
Non ELL *(Non English Language Learner)*	Control	17	-.001118	.0056000	.0013582
	Revised	12	.001750	.0119478	.0034490

Table 34.
One Sample t-Tests for Grade 4 DIF Differences

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Effect Size (*r*)	Mean Difference
NEP *(Non English Proficient)*	Control	.001294	.495	16	628		.0070001
	Revised	.004083	-.490	11	.634		-.0162497
LEP *(Limited English Proficient)*	Control	-.001059	1.776	16	.095		.0080002
	Revised	.005000	-.975	11	.350		-.0195833
FEP *(Fluent English Proficient)*	Control	-.003824	2.674	16	.017*	-0.169362	.0074711
	Revised	.004750	-.059	11	.954		-.0007500
Non ELL *(Non English Language Learner)*	Control	.003824	1.993	16	.0.64		.0027069
	Revised	.001667	.024	11	.981		.0000830

*p** = <.05

Table 35.
Grade 4 Revised and Control Items DIF Means by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2005 Std. Deviation	2007 Mean	2007 Std. Deviation
NEP *(Non English Proficient)*	Control	17	.001294	.0652799	.008294	.0583511
	Revised	12	.004083	.0833748	-.012167	.1149180
LEP *(Limited English Proficient)*	Control	17	-.001059	.0351292	.006941	.0185690
	Revised	12	.005000	.0322969	-.014583	.0695511
FEP *(Fluent English Proficient)*	Control	17	-.002118	.0207361	.003647	.0115214
	Revised	12	.004750	.0216842	.004000	.0438551
Non ELL *(Non English Language Learner)*	Control	17	-.003824	.0108065	-.001118	.0056000
	Revised	12	.001667	.0077381	.001750	.0119478

Grade 5 DIF Results

In grade 5, mean DIF values were not statistically significantly different between 2005 and 2007 for any language proficiency subgroup for either control or revised items (Table 37). Control items for LEP and FEP went from slightly disfavoring to slightly favoring those subgroups in 2007, while revised items for NEP and control and revised items for Non ELLs went from slightly favoring to slightly disfavoring those subgroups in 2007 (Table 38).

Table 36.
Grade 5 DIF Differences for Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	N	Mean	Std. Deviation	Std. Error Mean
NEP *(Non English Proficient)*	Control	16	.027188	.0415070	.0103767
	Revised	27	-.025667	.0810826	.0156043
LEP *(Limited English Proficient)*	Control	16	.002625	.0267753	.0066938
	Revised	27	.003296	.0446635	.0085955
FEP *(Fluent English Proficient)*	Control	16	.000438	.0158155	.0039539
	Revised	27	.010148	.0364034	.0070058
Non ELL *(Non English Language Learner)*	Control	16	-.000188	.0046793	.0011698
	Revised	27	-.000630	.0072703	.0013992

Table 37.
One Sample t-Tests for Grade 5 Differences

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Mean Difference
NEP *(Non English Proficient)*	Control	.017625	.922	15	.371	.0095625
	Revised	.006259	-2.046	26	.051	-.0319257
LEP *(Limited English Proficient)*	Control	-.001750	.654	15	.523	.0043750
	Revised	.01085	-.879	26	.388	-.0075537
FEP *(Fluent English Proficient)*	Control	-.002500	.743	15	.469	.0029375
	Revised	.006407	.534	26	.598	.0037411
Non ELL *(Non English Language Learner)*	Control	.000875	-.908	15	.378	-.0010625
	Revised	.000963	-1.138	26	.265	-.0015926

Table 38.
Grade 5 Revised and Control Items DIF Means by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2005 Std. Deviation	2007 Mean	2007 Std. Deviation
NEP *(Non English Proficient)*	Control	16	.017625	.0717634	.027188	.0415070
	Revised	27	.006259	.0670451	-.025667	.0810826
LEP *(Limited English Proficient)*	Control	16	-.001750	.0311330	.002625	.0267753
	Revised	27	.010185	.0377003	.003296	.0446635
FEP *(Fluent English Proficient)*	Control	16	-.002500	.0167531	.000438	.0158155
	Revised	27	.006407	.0284446	.010148	.0364034
Non ELL *(Non English Language Learner)*	Control	16	.000875	.0125053	-.000188	.0046793
	Revised	27	.000963	.0156733	-.000630	.0072703

Grade 6 DIF Results

In grade 6, DIF mean differences between 2005 and 2007 for all language proficiency subgroups showed statistically significant differences. For the control items, LEP, FEP and NON ELL were significant at p=001, p=.004, and p=.018 (Table 40). For LEP and FEP, the items went from slightly favoring to slightly disfavoring those subgroups in 2007 (Table 41). The control items for Non ELLs went from slightly disfavoring to slightly favoring in 2007. For the revised items, the DIF mean differences for NEP and LEP were statistically significant at *p*=.021, and p=.007 (Table 40). The revised items for LEP and FEP both went from disfavoring to favoring those subgroups (Table 41).

Table 39.
Grade 6 DIF Differences for Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	N	Mean	Std. Deviation	Std. Error Mean
NEP *(Non English Proficient)*	Control	8	.011750	.0430672	.0152266
	Revised	34	.018294	.0734395	.0125948
LEP *(Limited English Proficient)*	Control	8	-.015250	.0203733	.0072030
	Revised	34	.013147	.0490684	.0084152
FEP *(Fluent English Proficient)*	Control	8	-.010000	.0154087	.0054478
	Revised	34	.007676	.0364565	.0062522
Non ELL *(Non English Language Learner)*	Control	8	.001500	.0043753	.0015469
	Revised	34	-.001794	.0087656	.0015033

Table 40.
One Sample t-Tests for Grade 6 DIF Differences

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Effect Size (*r)*	Mean Difference
NEP *(Non English Proficient)*	Control	.032625	-1.371	7	.213		-.0208750
	Revised	.012118	2.415	33	.021*	-0.198776	.0304121
LEP *(Limited English Proficient)*	Control	.024500	-5.518	7	.001*	0.479132	-.0397500
	Revised	.011088	2.880	33	.007*	-0.246607	.0242351
FEP *(Fluent English Proficient)*	Control	.013500	-4.314	7	.004*	0.341448	-.0235000
	Revised	-.004882	2.009	33	.053		.0125585
Non ELL *(Non English Language Learner)*	Control	.003250	3.071	7	.018*	-0.301890	.0047500
	Revised	.000529	-1.545	33	.132		-.0023231

*p** = <.05

Table 41.
Grade 6 Revised and Control Items DIF Means by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2005 Std. Deviation	2007 Mean	2007 Std. Deviation
NEP	Control	8	.032625	.0452988	.011750	.0430672
	Revised	34	-.012118	.0764729	.018294	.0734395
LEP	Control	8	.024500	.0472894	-.015250	.0203733
	Revised	34	-.011088	.0461246	.013147	.0490684
FEP	Control	8	.013500	.0430681	-.010000	.0154087
	Revised	34	-.004882	.0241080	.007676	.0364565
Non ELL	Control	8	-.003250	.0096622	.001500	.0043753
	Revised	34	.000529	.0076524	-.001794	.0087656

Grade 7 DIF Results

Grade 7 control item DIF mean differences between 2005 and 2007 for LEP, FEP and NON ELL showed statistically significant differences at p=.008, p=.022, and p=.003 (Table 43). For NEP, revised items went from disfavoring the subgroup in 2005 to slightly favoring in 2007. Control items for LEP went from favoring the subgroup to disfavoring, while revised items went from disfavoring in 2005 to favoring in 2007. Control items for Non ELL went from slightly disfavoring to slightly favoring the subgroup in 2007 (Table 44).

Table 42.
Grade 7 DIF Differences for Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	N	Mean	Std. Deviation	Std. Error Mean
NEP *(Non English Proficient)*	Control	10	.013000	.0447064	.0141374
	Revised	35	.016229	.0616155	.0104149
LEP *(Limited English Proficient)*	Control	10	-.009300	.0228135	.0072143
	Revised	35	.010486	.0319755	.0054048
FEP *(Fluent English Proficient)*	Control	10	-.017400	.0162015	.0051234
	Revised	35	.006171	.0346168	.0058513
Non ELL *(Non English Language Learner)*	Control	10	.001700	.0028694	.0009074
	Revised	35	-.002200	.0059399	.0010040

Table 43.
One Sample t-Tests for Grade 7 DIF Differences

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Effect Size *(r)*	Mean Difference
NEP *(Non English Proficient)*	Control	.004300	.615	9	.554		.0087000
	Revised	-.000457	1.602	34	.118		.0166856
LEP *(Limited English Proficient)*	Control	.015100	-3.382	9	.008*	0.422281	-.0244000
	Revised	-.000457	2.025	34	.051		.0109427
FEP *(Fluent English Proficient)*	Control	-.003300	-2.752	9	.022*	0.321542	-.0141000
	Revised	.001286	.835	34	.410		.0048854
Non ELL *(Non English Language Learner)*	Control	-.002000	4.078	9	.003*	-0.226417	.0037000
	Revised	-.00057	-1.623	34	.114		-.0016300

*p** = <.05

Table 44.
Grade 7 Revised and Control Items DIF Means by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2005 Std. Deviation	2007 Mean	2007 Std. Deviation
NEP	Control	10	.004300	.0463131	.013000	.0447064
	Revised	35	-.000457	.0581582	.016229	.0616155
LEP	Control	10	.015100	.0291755	-.009300	.0228135
	Revised	35	-.001714	.0360780	.010486	.0319755
FEP	Control	10	-.003300	.0244861	-.017400	.0162015
	Revised	35	.001286	.0288011	.006171	.0346168
Non ELL	Control	10	-.002000	.0108832	.001700	.0028694
	Revised	35	-.000057	.0128153	-.002200	.0059399

Grade 8 DIF Results

In grade 8, control item DIF mean differences between 2005 and 2007 for LEP showed statistically significant differences at p=.008 (Table 46). Revised items for Non ELL went from slightly favoring the subgroup in 2005 to slightly disfavoring the subgroup in 2007 (Table 47).

Table 45.
Grade 8 DIF Differences for Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	N	Mean	Std. Deviation	Std. Error Mean
NEP *(Non English Proficient)*	Control	7	-.019429	.0513999	.0194273
	Revised	38	.002974	.0655725	.0106373
LEP *(Limited English Proficient)*	Control	7	-.032000	.0487374	.0184210
	Revised	38	.008737	.0525165	.0085193
FEP *(Fluent English Proficient)*	Control	7	-.027857	.0420295	.0158856
	Revised	38	.006447	.0373455	.0060582
Non ELL *(Non English Language Learner)*	Control	7	-.008714	.0230197	.0087006
	Revised	38	-.003000	.0129990	.0021087

Table 46.
One Sample t-Tests for Grade 8 DIF Differences

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Effect Size *(r)*	Mean Difference
NEP *(Non English Proficient)*	Control	-.034857	.794	6	.457		.0154284
	Revised	.007211	-.398	37	.693		-.0042373
LEP *(Limited English Proficient)*	Control	.039000	-3.854	6	.008*	-0.069221	-.0710000
	Revised	.014184	-.639	37	.527		-.0054472
FEP *(Fluent English Proficient)*	Control	-.031571	.234	6	.823		.0037139
	Revised	.018211	-1.942	37	.060		-.0117636
Non ELL *(Non English Language Learner)*	Control	-.010714	.230	6	.826		.0019997
	Revised	.001632	-2.197	37	.034*	0.180022	-.0046320

*p** = <.05

Table 47.
Grade 8 Revised and Control Items DIF Means by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2005 Std. Deviation	2007 Mean	2007 Std. Deviation
NEP *(Non English Proficient)*	Control	7	-.034857	.0450349	-.019429	.0513999
	Revised	38	.007211	.0590278	.002974	.0655725
LEP *(Limited English Proficient)*	Control	7	-.039000	.0520897	-.032000	.0487374
	Revised	38	.014184	.0476456	.008737	.0525165
FEP *(Fluent English Proficient)*	Control	7	-.031571	.0446500	-.027857	.0420295
	Revised	38	.018211	.0333541	.006447	.0373455
Non ELL *(Non English Language Learner)*	Control	7	-.010714	.0230775	-.008714	.0230197
	Revised	38	.001632	.0123012	-.003000	.0129990

Grade 9 DIF Results

In grade 9, mean DIF values were not statistically significantly different between 2005 and 2007 for any language proficiency subgroup for either control or revised items (Table 49). Control items for NEP went from slightly disfavoring to slightly favoring the subgroup in 2007. Revised items for LEP and FEP went from slightly favoring the subgroups in 2005 to slightly disfavoring the subgroups in 2007 (Table 50).

Table 48.
Grade 9 DIF Differences for Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	N	Mean	Std. Deviation	Std. Error Mean
NEP *(Non English Proficient)*	Control	5	.002000	.0532118	.0237971
	Revised	35	.000257	.0520906	.0088049
LEP *(Limited English Proficient)*	Control	5	-.004800	.0236685	.0105849
	Revised	35	-.001771	.0350202	.0059195
FEP *(Fluent English Proficient)*	Control	5	-.012200	.0151063	.0067557
	Revised	35	-.001029	.0317828	.0053723
Non ELL *(Non English Language Learner)*	Control	5	.000200	.0040249	.0018000
	Revised	35	.000571	.0066168	.0011184

Table 49.
One Sample t-Tests for Grade 9 DIF for Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Mean Difference
NEP *(Non English Proficient)*	Control	-.022000	1.009	4	.370	.0240000
	Revised	.001886	-.185	34	.854	-.0016289
LEP *(Limited English Proficient)*	Control	-.007800	.283	4	.791	.0030000
	Revised	.003343	-.864	34	.394	-.0051144
FEP *(Fluent English Proficient)*	Control	-.004600	-1.125	4	.324	-.0076000
	Revised	.000600	-.303	34	.764	-.0016286
Non ELL *(Non English Language Learner)*	Control	.001200	-.556	4	.608	-.0010000
	Revised	.000429	.127	34	.899	.0001424

Table 50.
Grade 9 Revised and Control Items DIF Means by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2005 Std. Deviation	2007 Mean	2007 Std. Deviation
NEP *(Non English Proficient)*	Control	5	-.022000	.0344166	.002000	.0532118
	Revised	35	.001886	.0422595	.000257	.0520906
LEP *(Limited English Proficient)*	Control	5	-.007800	.0165892	-.004800	.0236685
	Revised	35	.003343	.0445427	-.001771	.0350202
FEP *(Fluent English Proficient)*	Control	5	-.004600	.0101390	-.012200	.0151063
	Revised	35	.000600	.0335701	-.001029	.0317828
Non ELL *(Non English Language Learner)*	Control	5	.001200	.0086718	.000200	.0040249
	Revised	35	.000429	.0085965	.000571	.0066168

Grade 10 DIF Results

Grade 10 mean DIF values were not statistically significantly different between 2005 and 2007 for any language proficiency subgroup for either control or revised items (Table 52). Control items for Non ELLs went from slightly disfavoring to slightly favoring the subgroup in 2007. Revised items for LEP and FEP went from slightly disfavoring the subgroups in 2005 to slightly favoring the subgroups in 2007 (Table 53).

Table 51.
Grade 10 DIF Differences for Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	N	Mean	Std. Deviation	Std. Error Mean
NEP *(Non English Proficient)*	Control	7	.012571	.0376026	.0142124
	Revised	28	-.002929	.0740050	.0139856
LEP *(Limited English Proficient)*	Control	7	.003857	.0161908	.0061196
	Revised	28	.001143	.0412299	.0077917
FEP *(Fluent English Proficient)*	Control	7	-.002286	.0152065	.0057475
	Revised	28	.001964	.0254260	.0048051
Non ELL *(Non English Language Learner)*	Control	7	-.000714	.0042706	.0016141
	Revised	28	.000250	.0093872	.0017740

Table 52.
One Sample T-Tests for Grade 10 DIF Differences for Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Mean Difference
NEP *(Non English Proficient)*	Control	.029429	-1.186	6	.280	-.0168576
	Revised	.004143	-.506	27	.617	-.0070716
LEP *(Limited English Proficient)*	Control	.008000	-.677	6	.524	-.0041429
	Revised	-.009500	1.366	27	.183	.0106429
FEP *(Fluent English Proficient)*	Control	.004143	-1.119	6	.306	-.0064287
	Revised	-.005964	1.650	27	.111	.0079283
Non ELL *(Non English Language Learner)*	Control	-.003857	1.947	6	.099	.0031427
	Revised	.003179	-1.651	27	.110	-.0029290

Table 53.
Grade 10 Revised and Control Items DIF Means by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2005 Std. Deviation	2007 Mean	2007 Std. Deviation
NEP *(Non English Proficient)*	Control	7	.029429	.0497790	.012571	.0376026
	Revised	28	-.009857	.0890737	-.002929	.0740050
LEP *(Limited English Proficient)*	Control	7	.008000	.0198494	.003857	.0161908
	Revised	28	-.009500	.0571253	.001143	.0412299
FEP *(Fluent English Proficient)*	Control	7	.004143	.0145422	-.002286	.0152065
	Revised	28	-.005964	.0299907	.001964	.0254260
Non ELL *(Non English Language Learner)*	Control	7	-.003857	.0103026	-.000714	.0042706
	Revised	28	.003179	.0148749	.000250	.0093872

Combined Grades DIF Results

In looking at all grades combined, differences in DIF were significant (.001) only for revised items in the FEP subgroup (Table 55), where the difference between the observed and expected performance (DIF) increased for that subgroup (Table 56). With all grades combined, control items for LEPs and revised items for Non ELLs went from slightly favoring to slightly disfavoring, while revised items for NEPs went from slightly disfavoring to slightly favoring that subgroup (Table 56).

Table 54.
Combined Grades DIF Differences for Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	N	Mean	Std. Deviation	Std. Error Mean
NEP *(Non English Proficient)*	Control	83	.009518	.0471966	.0051805
	Revised	220	.001814	.0708711	.0047781
LEP *(Limited English Proficient)*	Control	83	-.003229	.0269771	.0029611
	Revised	220	.005045	.0440666	.0029710
FEP *(Fluent English Proficient)*	Control	83	-.005325	.0200311	.0021987
	Revised	220	-.000991	.0092435	.0006232
Non ELL *(Non English Language Learner)*	Control	83	-.000795	.0084055	.0009226
	Revised	220	.005064	.0337694	.0022767

Table 55.
One Sample t-Tests for Combined Grades DIF Differences for Revised and Control Items by Language Proficiency Level

Language Proficiency Level	Type of Item	Test Value	t	df	Sig. (2-tailed)	Effect Size *(r)*	Mean Difference
NEP *(Non English Proficient)*	Control	.006506	.581	82	.563		.0030121
	Revised	-.000905	.569	219	.570		.0027186
LEP *(Limited English Proficient)*	Control	.00108	-1.455	82	.149		-.0043089
	Revised	.001432	1.216	219	.225		.0036135
FEP *(Fluent English Proficient)*	Control	-.002506	-1.282	82	.203		-.0028193
	Revised	.001127	-3.398	219	.001*	-0.034765	-.0021179
Non ELL *(Non English Language Learner)*	Control	-.002373	1.710	82	.091		.0015778
	Revised	.002845	.974	219	.331		.0022186

Table 56.
Combined Grades Revised and Control Items DIF Means by Language Proficiency Level by Year

Language Proficiency Level	Type of Item	N	2005 Mean	2005 Std. Deviation	2007 Mean	2007 Std. Deviation
NEP *(Non English Proficient)*	Control	83	.006506	.0575655	.009518	.0471966
	Revised	220	-.000905	.0657603	.001814	.0708711
LEP *(Limited English Proficient)*	Control	83	.000108	.0368729	-.003229	.0269771
	Revised	220	.001432	.0447287	.005045	.0440666
FEP *(Fluent English Proficient)*	Control	83	-.002506	.0258751	-.005325	.0200311
	Revised	220	.002845	.0299031	.005064	.0337694
Non ELL *(Non English Language Learner)*	Control	83	-.002373	.0123602	-.000795	.0084055
	Revised	220	.001127	.0117385	-.000991	.0092435

Based on the data, while there were significant differences in control item DIF for several grades in different language proficiency subgroups, there was no difference for plain language revised items. In examining DIF differences with all grades combined, however, DIF for the revised items for FEP were significantly different, with DIF increasing in 2007. Based on the data from the individual grade levels, the null hypothesis was not rejected.

Revision Category

Finally, to determine if certain plain language revisions (context, graphics, vocabulary/wording, sentence structure, and format/style) of mathematics assessment items on CSAP provide greater access, the differences

between 2005 and 2007 in mean item difficulty and mean DIF value for each of the language proficiency subgroups were calculated. For all grades combined, a one-way ANOVA was used to examine any significant difference for types of revisions for each language proficiency level. Each item was categorized into only one category. Tables 57 to 60 provide the minimum, maximum, mean, and standard deviation of combined grades differences in item difficulty and DIF by revision category for each of the language proficiency subgroups.

Table 57.
Minimum, Maximum, Mean, and Standard Deviation of Combined Grades Differences in Item Difficulty and DIF by Revision Category for NEPs

		N	Mean	Std. Deviation	Min	Max
Difference in Item Difficulty for NEPs	No Revision	83	-.0179	.04409	-.12	.10
	Wording	126	-.0182	.03901	-.17	.06
	Format/Style	31	-.0270	.03208	-.11	.04
	Sentence Structure	8	-.0189	.01639	-.05	.01
	graphics	2	-.0299	.02832	-.05	-.01
	format/sentence	20	-.0153	.06170	-.13	.13
	format/wording	22	-.0444	.03372	-.14	.00
	Other multiple	10	-.0248	.03681	-.07	.04
	Total	302	-.0210	.04106	-.17	.13
Difference in DIF for NEPs	No Revision	83	.0030	.05035	-.11	.21
	Wording	127	.0070	.07087	-.25	.30
	Format/Style	31	.0048	.07276	-.13	.15
	Sentence Structure	8	.0213	.04694	-.04	.12
	graphics	2	-.0475	.07849	-.10	.01
	format/sentence	20	-.0015	.05467	-.15	.09
	format/wording	22	-.0182	.09042	-.26	.25
	Other multiple	10	-.0086	.07055	-.13	.11
	Total	303	.0028	.06601	-.26	.30

Table 58.
Minimum, Maximum, Mean, and Standard Deviation of Combined Grades Differences in Item Difficulty and DIF by Revision Category for LEPs

		N	Mean	Std. Deviation	Min	Max
Difference in Item Difficulty for LEPs	No Revision	83	.0398	.03655	-.07	.16
	Wording	126	.0304	.03862	-.05	.15
	Format/Style	31	.0390	.05144	-.12	.14
	Sentence Structure	8	.0165	.03418	-.02	.07
	graphics	2	.0197	.01340	.01	.03
	format/sentence	20	.0334	.08558	-.09	.25
	format/wording	22	.0186	.03819	-.10	.07
	Other multiple	10	.0515	.04780	.00	.15
	Total	302	.0335	.04428	-.12	.25
Difference in DIF for LEPs	No Revision	83	-.0033	.03254	-.12	.09
	Wording	127	.0029	.05138	-.19	.21
	Format/Style	31	.0096	.04677	-.11	.11
	Sentence Structure	8	.0016	.04202	-.04	.09
	graphics	2	.0225	.08697	-.04	.08
	format/sentence	20	.0070	.02532	-.05	.06
	format/wording	22	-.0055	.04759	-.09	.12
	Other multiple	10	.0050	.04014	-.05	.08
	Total	303	.0017	.04406	-.19	.21

Table 59.
Minimum, Maximum, Mean, and Standard Deviation of Combined Grades Differences in Item Difficulty and DIF by Revision Category for FEPs

		N	Mean	Std. Deviation	Min	Max
Difference in Item Difficulty for FEPs	No Revision	83	.0241	.03477	-.07	.17
	Wording	126	.0254	.03716	-.06	.17
	Format/Style	31	.0207	.04454	-.11	.11
	Sentence Structure	8	.0249	.03398	-.03	.07
	graphics	2	.0199	.00611	.02	.02
	format/sentence	20	.0232	.09121	-.08	.27
	format/wording	22	.0162	.03285	-.06	.08
	Other multiple	10	.0472	.04847	-.02	.15
	Total	302	.0244	.04269	-.11	.27
Difference in DIF for FEPs	No Revision	83	-.0028	.02324	-.11	.06
	Wording	127	.0013	.03380	-.14	.16
	Format/Style	31	.0040	.03479	-.08	.11
	Sentence Structure	8	.0100	.01973	-.02	.04
	graphics	2	.0150	.05233	-.02	.05
	format/sentence	20	.0035	.02184	-.03	.04
	format/wording	22	-.0047	.03235	-.07	.05
	Other multiple	10	.0124	.04556	-.08	.07
	Total	303	.0008	.03069	-.14	.16

Table 60.
Minimum, Maximum, Mean, and Standard Deviation of Combined Grades Differences in Item Difficulty and DIF by Revision Category for Non ELLs

		N	Mean	Std. Deviation	Min	Max
Difference in Item Difficulty for Non ELLs	No Revision	83	.0175	.02844	-.05	.18
	Wording	126	.0191	.02726	-.08	.13
	Format/Style	31	.0107	.03489	-.09	.08
	Sentence Structure	8	.0160	.03033	-.03	.05
	graphics	2	.0204	.00192	.02	.02
	format/sentence	20	.0200	.09041	-.07	.27
	format/wording	22	.0097	.02719	-.05	.06
	Other multiple	10	.0363	.03901	.00	.12
	Total	302	.0177	.03616	-.09	.27
Difference in DIF for Non ELLs	No Revision	83	.0016	.01001	-.05	.03
	Wording	127	-.0007	.01264	-.06	.05
	Format/Style	31	-.0036	.01228	-.04	.02
	Sentence Structure	8	-.0043	.00828	-.02	.01
	graphics	2	.0020	.00566	.00	.01
	format/sentence	20	-.0021	.01581	-.03	.03
	format/wording	22	-.0049	.01265	-.04	.02
	Other multiple	10	-.0084	.01673	-.05	.01
	Total	303	-.0011	.01231	-.06	.05

Type of plain language revision had no statistically significant effect on item difficulty or DIF for any of the language proficiency subgroups as significance levels were all greater than .05 (Table 61). A Tukey's post hoc analysis was also conducted to compare each revision category. No plain language revision category was significantly different relative to item difficulty differences or DIF differences when compared to other revision categories. The null hypothesis was not rejected.

Table 61.
ANOVA - Combined Grades Differences in Item Difficulty and DIF by Revision Category

		Sum of Squares	df	Mean Square	F	Sig.
Item Difficulty Difference NEP	Between Groups	.016	7	.002	1.360	.222
	Within Groups	.492	294	.002		
	Total	.508	301			
DIF Difference NEP	Between Groups	.022	7	.003	.700	.672
	Within Groups	1.294	295	.004		
	Total	1.316	302			
Item Difficulty Difference LEP	Between Groups	.016	7	.002	1.188	.309
	Within Groups	.574	294	.002		
	Total	.590	301			
DIF Difference LEP	Between Groups	.007	7	.001	.502	.833
	Within Groups	.579	295	.002		
	Total	.586	302			
Item Difficulty Difference FEP	Between Groups	.007	7	.001	.568	.782
	Within Groups	.541	294	.002		
	Total	.548	301			
DIF Difference FEP	Between Groups	.005	7	.001	.702	.670
	Within Groups	.280	295	.001		
	Total	.284	302			
Item Difficulty Difference Non ELL	Between Groups	.007	7	.001	.741	.637
	Within Groups	.387	294	.001		
	Total	.393	301			
DIF Difference Non ELL	Between Groups	.002	7	.000	1.692	.110
	Within Groups	.044	295	.000		
	Total	.046	302			

Chapter 5

DISCUSSION

One of the primary reasons to incorporate plain language revision into development of large scale assessments is to ensure optimum access to comprehensible content. The results of this study showed that for some ELLs, plain language revision may provide more access to assessment items; however, some item characteristics and measurements remained unchanged. Plain language revision alone is not enough to ensure that ELLs have equitable access to assessment items.

The first research question focused on Depth of Knowledge, and if those ratings are impacted by plain language revision. DOK should remain constant if the construct of the item remains intact; it was included as a part of this study to examine if perceptions of cognitive complexity are influenced by the linguistic complexity of an item rather than the assessment task. While DOK is not truly an indicator of access to the item, it was included as a part of this study as it is meant to be independent of item difficulty. It was supposed that added linguistic

complexity in the item stem would contribute to an artificially inflated DOK, and that plain language revision may cause the ratings to be lower.

Without having the same raters rate both forms of the items as a part of the same study, DOK change was difficult to examine. Even though a significant association exists between the 2005 and 2007 DOK ratings, in some cases, 25% to 40% of items changed ratings. DOK for the 2005 items was assigned by a panel of expert judges during Colorado's Content Validity and Alignment Study in 2005. The 2007 ratings were assigned by the content editors and item writers at CTB. Given the subjective nature of the DOK assignment and varying levels of training or experience in assigning DOK, changes in DOK ratings were not unanticipated. It is interesting to note that the DOK for both control items and plain language revised items was higher in 2007. These results must be viewed with caution as there is no indication that the DOK of the items actually changed, the raters were different and the method by with the DOK rating was obtained varied between 2005 and 2007 as well. Additional alignment studies on these items would be required to validate any DOK ratings. Because rating changed for both control and revised items, this is most likely due to differences in raters' perceptions or ability to identify correctly the cognitive complexity of an item rather than due to effects of plain language revision.

The second research question asked if item difficulty was changed due to plain language revision of items. The only grade in which this value changed significantly for any of the subgroups between 2005 and 2007 was 5th grade, where the item difficulty for revised items within the LEP subgroup was significantly different than 2005 item difficulty, with control group item difficulty in the LEP subgroup not changing significantly. In this instance, the LEP students performed better on the revised items.

In grade 5, where half of the control items were computation items, it was clear that ELLs did not perform as well on linguistic items as strictly computational items (Tables 66 and 67). Students were able to show their math ability on the computation items, whereas they were less successful in showing their math ability on the linguistically-based control items. While these two sets of items may have measured similar math concepts, the linguistic structure may have contributed to item difficulty. Given that, it is essential that accommodations are provided for students in order for their true content knowledge to be revealed on this large scale, paper and pencil test. Without access to the content, students may be placed in mathematics classes well below their true ability.

Table 62.
Mean, Maximum, Minimum and Standard Deviation of Computation Control Items in Grade 5

	N	Minimum	Maximum	Mean	Std. Deviation
2007 Item Difficulty for LEPs	8	.609	.911	.75030	.106653
2007 Item Difficulty for NEPs	8	.458	.812	.60452	.140207
2007 Item Difficulty for FEPs	8	.737	.957	.85865	.072663
2007 Item Difficulty for Non ELLs	8	.752	.947	.85913	.065553
Valid N (listwise)	8				

Table 63.
Mean, Maximum, Minimum and Standard Deviation of Linguistically Based Control Items in Grade 5

	N	Minimum	Maximum	Mean	Std. Deviation
2007 Item Difficulty for LEPs	8	.203	.888	.58435	.268940
2007 Item Difficulty for NEPs	8	.125	.753	.45459	.250393
2007 Item Difficulty for FEPs	8	.348	.934	.70957	.231374
2007 Item Difficulty for Non ELLs	8	.391	.927	.72584	.214000
Valid N (listwise)	8				

Because mean values were used to explore this question, individual items may have proved to provide increased access for language proficiency subgroups, however, within each grade the mean difference was not significant.

While not prevalent in every grade level, it does appear that plain language revision may increase access for LEPs as item difficulty values are higher after revision, indicating that more LEP students answered these items correctly. It may be that as LEPs are rapidly acquiring the

English language, they are more sensitive to linguistic changes than non English proficient students or students already proficient in English (FEPs and Non ELLs). Plain language revision for CSAP mathematics items is not alone sufficient to ensure that ELLs have access to comprehensible content.

The third question dealt with item discrimination, and if that item parameter is affected by plain language revision. While item discrimination (a -parameter) is meant to be independent of item difficulty (*b*-parameter), they are interrelated in that more difficult items do tend to discriminate more highly. In this study, while the differences in item discrimination between the two administrations were not statistically significant, the control items had a higher discrimination than the revised items in 2007, and a higher p-value. Likewise, the revised items, which had lower p-values, did not discriminate as highly. All language subgroups apart from NEPs performed better on both control and revised items in 2007 than in 2005 (Tables 25 and 26). While it appears that in this case that the higher discriminating items were also more difficult, this could also indicate that the items in 2005 discriminated based on linguistic ability, and once the linguistic complexity of the items was reduced, so too was ability of the item to discriminate. Because the Linn-Harnisch method examines the decile groups performance within the larger subgroup, it

would also be beneficial to utilize another methodology to examine DIF, one that compares different subgroups. Utilizing a different methodology may yield very different results.

The fourth research question focused on DIF across performance groups, and whether differences in DIF result from plain language revision. Overall, differences in DIF in this study were not impacted by plain language revision. Only in grade 6 was DIF for revised items statistically significantly different from 2005 to 2007 for NEP and LEP. DIF for control items for LEP was also statistically significantly different; however, the direction of the change is important to note. For LEP, the control items went from slightly favoring the subgroup in 2005, to slightly disfavoring the subgroup in 2007, while the revised items went from slightly disfavoring the subgroup in 2005 to slightly favoring the subgroup in 2007. While this provides some indication that the difference between the expected and observed performance for LEP in 6th grade may have been impacted by plain language revision, group differences may have contributed to the variance in DIF between the two years.

While using DIF to identify bias is a well established practice, new research indicates that other analyses, such as distractor analyses should be used to identify bias specifically for certain subgroups, including ELLs (Kopriva, Cameron, Carr, & Taylor, 2008). It would be important to look at

the difference in DIF values of all items over several administrations of the assessment in order to ascertain an expected and reasonable variance in those values.

The final research question looked at different types of plain language revision, and if any certain type of revision provided more access. Both the difference in item difficulty and difference in DIF for each of the language subgroups between 2005 and 2007 were used to explore this question. The data did not indicate that specific types of revisions provide greater access for ELLs than others. While the categories of revision may be helpful in learning about plain language revision, in reality most plain language revisions will incorporate many of the different categories. It is possible that revisions that fell into one category may not have been substantial enough to provide increased access, particularly those where only one or two words were eliminated from the directions or items stems.

While not every piece of evidence indicated that plain language revision benefits ELLs, there is a suggestion that for LEP students, plain language revision may be a key step in ensuring that assessment items are accessible, but plain language revision alone is not enough to ensure that ELL students have equitable access to mathematics assessment items.

Limitations

While the CSAP administration is standardized as much as possible, it is unlikely that every student that took the assessment did so under identical conditions, possibly contributing to score variance.

The data used in this study was generated by CTB/McGraw-Hill as a part of the Colorado Student Assessment Program contract. While the researcher conducted the comparative analyses, the data used for these analyses (item difficulty, DIF, and item discrimination) were provided by CTB/McGraw-Hill, and thus the ways in which the data could be viewed were limited. For example, p-values were provided but IRT difficulty estimates were not. The types of data available were also limited by contract, therefore, some data requests could not be fulfilled.

DOK for each math item on the 2005 assessment was identified as a part of Colorado's Content Validity and Alignment Study in spring of 2005 which involved over one hundred educators from around the state. For the 2007 assessment, item writers and content editors from CTB assigned DOK to the items during the development cycle. As the assignment of DOK is based on judgment, it is likely that some variance will occur in the rating of DOK while the actual cognitive complexity of the item has not changed even with revisions.

There are factors and confounding variables which cannot be controlled which may impact student performance on the assessment. Students may or may not receive accommodations for the assessment, which may introduce or eliminate barriers for individual students. Students who did receive accommodations, linguistic or otherwise were not removed from this sample; therefore, some students may have had better access to the items than others. Some students may have been provided a greater opportunity to learn than others.

Furthermore, as CSAP mathematics assessments contain many more linguistically based items than purely computational items, the linguistic barrier cannot totally be removed. For students with very little English language proficiency, it may be impossible to ascertain their true knowledge of mathematics on this assessment unless other accommodations are applied.

While the unrevised anchor items were used to determine if there were group differences that could explain any change in revised items, there may be other explanations besides plain language revision for changes in item performance.

In examining the types of plain language revision, some categories had only a few items to include as part of the analysis. In operational state assessments, revisions are based on need, not by category of revision. A

more robust sample of the different types of linguistic changes would be necessary to draw more definite conclusions.

Suggestions for Future Research

As it is clear that plain language revision alone may not provide sufficient access for ELLs, studies examining plain language revised items in conjunction with specific linguistic accommodations may yield information relative to access on these assessment items for ELLs.

A study where ELL students English language proficiency score was matched to their content area score would be helpful to more closely examine the LEP category and item difficulty for this particular subgroup after items are revised for plain language.

Continual examination is needed as to the accuracy of DOK ratings of assessment items. To better examine any changes in rating because of plain language revision, a within subjects design may better isolate these discrepancies, as the same rater would be rating both plain language revised as well as unrevised items. It would also be beneficial to examine in depth the perceptions of raters as they are rating items. This may lead to a greater understanding behind variances in rating and provide some insight into the actual value of such ratings.

It would be beneficial to have an experimental study with equal numbers of items in the different revision categories so as to better understand if different types of plain language revision provide greater access.

A focused study on plain language revisions in conjunction with accommodations for NEP students would help add to the body of evidence for ways in which access for this group of ELLs could be enhanced. Gathering information relative to the student's educational background would also shed light on the impact of linguistic barriers versus lack of content knowledge.

A study analyzing distractor choices by subgroups would add to the information relative to item bias, possibly revealing patterns in types of multiple choice responses the different language proficiency subgroups choose. Further, examination of DIF by linguistic subgroup rather than performance grouping might be informative.

References

Abedi, J. (2007). *Language factors in the assessment of English language learners: The theory and principles underlying the linguistic modification approach*. Paper presented at the LEP Partnership Meeting.

Abedi, J. (2008). *Consistencies between results of DIF analyses by different approaches for ELLs in math and reading tests*. Paper presented at the National Council on Measurement in Education.

Abedi, J., Courtney, M., & Leon, S. (2003). *Effectiveness of validity of accommodations for English language learners in large-scale assessments* (CSE Report No. 608). Los Angeles: National Center for Research and Evaluation, Standards and Student Testing

Abedi, J., Courtney, M., & Leon, S. (2003). *Research-supported accommodation for English language learners* (CSE Technical report No. 56). Los Angeles: National Center for Research and Evaluation, Standards and Student Testing

Abedi, J., Courtney, M., Mirocha, J., Leon, S., & Goldberg, J. (2005). *Language accommodations for English language learners in large-scale assessments: Bilingual dictionaries and linguistic modification*

(CSE Report No. 666). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing; University of California.

Abedi, J., & Gandara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues an Practice*(Winter).

Abedi, J., Hofstetter, C., & Baker, E. (2001). *NAEP math performance and test accommodations: Interactions with student language background* (CSE Technical Report No. 536). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Lord, C. (2001). The language factor in mathematics assessments. *Applied Measurement in Education, 14*(2), 219-234.

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance*. Los Angeles, CA: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards and Student Testing.

Abedi, J., Lord, C., & Plummer, J. (1997). *Final report of language background as a variable in NAEP mathematics performance* (CSE Technical Report No. 429). Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards and Student Testing.

Abedi, J., Mirocha, J., & Leon, S. (2001). *Students' performance differences in standardized achievement tests and background factors: Analyses of extant data.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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

Box, G. E. P. (1953). Non-normality and tests for variance. *Biometrika, 40*, 318-335.

Brown, P. J. (1999). *Findings of the 1999 plain language field test*. Newark, DE: University of Delaware, Delaware Education Research and Development Center.

Burket, G. R. (1993). PARDUX (Version 1.7).

Butler, F. A., & Stevens, R. A. (1997). *Accommodations strategies for English language learners on large-scale assessments: Student characteristics and other considerations* (CSE Technical Report No. 448). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing; University of California.

CDE. (2007, 9/17/07). Colorado accommodations manual for English language learners. First Edition. Retrieved 10/1/07, from http://www.cde.state.co.us/cdeassess/documents/csap/manuals/2007/CO_ACCOMM_MANUAL_ELL_091707.pdf

CDE. (2008). CSAP Released Items. Retrieved May 25, 2008, from http://www.cde.state.co.us/cdeassess/released_items.html

CDE, & ELAU. (2007). English language in Colorado : A state of the state. Retrieved 11/01/07, from http://www.cde.state.co.us/cde_englihs/download/Resources-Links/astateoftheState.pdf

Colorado Department of Education. (2007). Colorado accommodations manual. First Edition. Retrieved 11/1/07, 2007, from

http://www.cde.state.co.us/cdeassess/documents/csap/manuals/2007/CO_%20ACCOMM_MANUAL_10172007.pdf

Colorado Department of Education. (2007, 9/17/07). Colorado accommodations manual for English language learners. First Edition. Retrieved 10/1/07, from http://www.cde.state.co.us/cdeassess/documents/csap/manuals/2007/CO_ACCOMM_MANUAL_ELL_091707.pdf

Colorado Department of Education. (2008). CSAP Released Items. Retrieved May 25, 2008, from http://www.cde.state.co.us/cdeassess/released_items.html

Connell, B. R., Jones, M., Mace, R., Mueller, J., Mullick, A., Ostroff, E., et al. (1997, 4/1/97). The principles of universal design. Version 2.0. Retrieved November 1, 2007

CTB McGraw-Hill. (2007). *Colorado Student Assessment Program Technical Report 2007*.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20*, 405-438.

Garcia, G. E. (1991). Factors influencing the English reading test performance of Spanish-speaking Hispanic children. *Reading research Quarterly, 26*(4).

Gaster, L., & Clark, C. (1995). *A guide to providing alternate formats* (No. ERIC Document No. ED 405689). West Columbia, SC: Center for Rehabilitation Technology Services.

Hanson, M. R., Hayes, J. R., Schriver, K., LeMahieu, P. G., & Brown, P. J. (1998). *A plain language approach to the revision of test items*. Paper presented at the American Educational Research Association.

Johnson, H. L. NCLB standards and assessment decision letter. In C. D. o. Education (Ed.) (Colorado's Peer Approval Letter ed.): United States Department of Education.

Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (NCEO Technical Report No. 37). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Jones, D. D. (2007). *Forward thinking: The voice and future of CDE*. Denver: Colorado Department of Education.

Kiplinger, V. L., Haug, C. A., & Abedi, J. (2000). *A math assessment should test math, not reading: One state's approach to the problem*. Paper presented at the 30th annual National Conference on Large-Scale Assessments.

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

Kopriva, R. J., Cameron, C., Carr, T., & Taylor, M. (2008). *The limits of DIF: Why this item evaluation tool is flawed for English learners, hearing impaired, and students with learning disabilities*. Paper presented at the National Council of Measurement in Education.

La Celle-Peterson, M. W., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review, 64*(1), 55-75.

Lefly, D. (2007). Spreadsheet of accommodations data. In H. Baker (Ed.).

Lefly, D., & Karkee, T. (2007). *Modification of items under universal design: Technical issues and impact on items and students.* Paper presented at the National Large Scale Assessment Conference, Nashville, TN.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*(2), 109-118.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erldaum.

The Math Forum @ Drexel. (2008). Retrieved May 25, 2008, from http://mathforum.org/dr.math/

Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13 - 103). Phoenix, AZ: Oryx Press.

No Child Left Behind Act of 2001, Public Law No. 107-110 (2002).

Ockey, G. J. (2007). Investigating the validity of math word problems for English language learners with DIF. *Language Assessment Quarterly, 4*(2), 149-164.

Rakow, S. J., & Gee, T. C. (1987). Test science, not reading. *Science Teacher, 54*(2), 28-31.

Rivera, C., & Stansfield, C. W. (2001). The effects of linguistic simplification of science test items on performance of limited English proficient and monolingual English-speaking students.

Journal. Retrieved from http://ceee.gwu/Products_asess.Accountability/simplification.pdf

Sato, E. (2007). *A guide to linguistic modification: Increasing English language learner access to academic content*. Paper presented at the LEP Partnership.

Stevens, R. A., Butler, F. A., & Castellon-Wellington, M. (2000). *Academic language and content assessment: Measuring the progress of English Language Learners* (CSE Technical Report No. 552). Los Angeles: National Center for Research on Evaluation, Standards and Student Testing; University of California.

Stevens, R. A., Butler, F. A., & Castellon-Wellington, M. (2000). *Academic language and content assessment: measuring the progress of English language learners* (CSE Technical Report No. 552). Los Angeles: National Center for Research on Evaluation, Standards and Student Testing; University of California.

Thompson, S. J., Johnstone, C. J., Anderson, M. E., & Miller, N., A. (2005). *Considerations for the development and review of universally designed assessments* (Technical Report No. 42).

Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report No. 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Unit of Student Assessment, & English Language Acquisition Unit. (2007). English language in Colorado: A state of the state. Retrieved 11/01/07, from http://www.cde.state.co.us/cde_englihs/download/Resources-Links/astateoftheState.pdf

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research Monograph No. 6): National Institute for Science Education; University of Wisconsin-Madison; Council of Chief State School Officers, Washington, DC.

Notes

[1]This is one of the current research projects under the auspices of the Center for Research on Evaluation, Standards and Student Testing (CRESST). CRESST will be working in Colorado, among other locations, studying the effects of accommodations for varying proficiency levels of ELLs in addition to non-ELL students. If non-ELL students' performance improves, then the effectiveness and validity of that accommodation must be questioned.

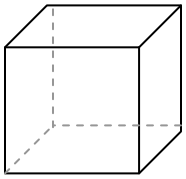
Appendices

Appendix A.
Examples of Different Levels of DOK in Items

DOK 1 - Recall

This level may have students performing a routine operation such as a computation problem, or recalling facts or formulas. Problems at this level are solved with routine procedures with clearly defined steps.

Study the figure.



How many edges does the figure have?

The student has to recall what an edge is and then count the number of edges. This is a DOK of 1.

DOK 2 – Skill/Concept

This levels requires some mental processing beyond a habitual response and will require students to decide on how to approach a given task or problem.

A student has 64 pieces of candy. He kept 12 pieces of candy. The student then divided the candy equally between four of his friends. How many pieces of candy did each of his friends get?

The student is required to do computation for this problem, but it is a multi-step problem in which the student must decide how to approach the computation. This item is a DOK 2.

DOK 3 - Strategic Thinking

This level requires student to reason, plan, and use evidence in solving problems and drawing conclusions. Many level three items will require students to explain or justify their reasoning.

Study the Table.

Farmer's Market Flower Basket Sales

Number of flower baskets	20	25	30	35
Amounts of Money	$120	$150	$180	

P

art A - Complete the table to show the amount of money Ann receives for selling 35 flower baskets.

Part B - On the lines below, explain the rule used in the pattern.

Part C - Ann pays $45 each day for a place at the farmers' market. One day, she sold 8 flower baskets. Did Ann receive enough money to pay for her place that day? Show your work and explain your reasoning, and write your answer on the line (CDE, 2008; Colorado Department of Education, 2008).

This item is a DOK three, as the student must choose use information from the table to devise a strategy to solve a problem and justify their reasoning.

Appendix B.
Examples Linguistic Change Categories

Context

In this item, we want to find out if the student knows how to select a scientific tool for the appropriate purpose. In this case, the context of the type of experiment is unnecessary and may pose additional linguistic barriers for some students. This is an example of an item that could be revised for context.

> Veronica and her friend Tami are conducting an experiment to determine the growth of bacteria in different liquids. They are using three different types of liquids and will conduct the experiment over a period of three weeks. Veronica and Tami will record the algae growth on a daily basis.
>
> Which tool should the student use to measure the amount of liquid needed for the experiment?
>
> Thermometer
>
> Graduated Cylinder
>
> Balance
>
> Petri Dish

The item could be revised as follows:

> A student needs to measure liquid for an experiment. Which tool would be the best to use?
>
> Thermometer
>
> Graduated Cylinder
>
> Balance
>
> Petri Dish

By removing some of the original context, the construct becomes clearer. There are instances where context needs to be added to an item to provide clarity for the student as to the required task.

Graphics

In this item, information is given which is not essential in solving the problem; however it is sometimes important for the student to be able to identify the values needed to solve the problem. Having several values in the item stem, however, can be confusing so this is an example of an item that could be revised for graphics.

> The park forestry department often keeps track of tree growth within the park. A particular white pine close to a highway had been measured since 1962. In 1962 this white pine was 16 meters tall. It was 20 meters tall in 1965. It is now 34 meters tall. How much has it grown since 1962? ("The Math Forum @ Drexel," 2008).

The item could be revised as follows:

Study the table.

White Pine Growth

Year	Height
2002	16 meters
2005	20 meters
2008	34 meters

How much has the white pine grown since 2002?

By adding the table we have made the important information easier to navigate in this item.

Vocabulary/Wording

For CSAP, when items were categorized as vocabulary/wording most often complicated proper names were changed to more generic names.

Tonya had 8 red blocks and 2 blue blocks in a bag. Bryant pulls out one block. What is the probability the block will be blue?

This item could be revised as follows:

A student has 8 red blocks and 2 blue blocks in a bag. She pulls out one block. What is the probability the block will be blue?

Sentence Structure

This item was written in passive voice. To increase access the item was revised to active voice, keeping the construct intact.

> On Sunnybrook Farm there are 144 chickens. Six chickens are placed in each cage. In total, how many cages are needed for the 144 chickens? ("The Math Forum @ Drexel," 2008).

This item could be revised as follows:

> A farmer has 144 chickens. He puts 6 chickens in each cage. How many cages does he need?

Other items coded as revised for sentence structure need clarified pronoun relationships, moving subjects and objects within the sentence so they were closer to one another and putting modifiers next to the words they modify.

Format/Style

Items with a great deal of extraneous language in the stem were coded as format/style.

> Bob and Sam drove a yellow Mercedes convertible from New York to Los Angeles, a distance of 3900 miles. They stopped for coffee 28 times, and stopped to use restrooms 19 times. They spent $125 on food, and listened mostly to the Beach Boys and Nirvana while driving. They were stopped for speeding twice, but were able in

> each case to talk the police officer out of giving them a ticket. The price of gas varied between $1.29 and $1.56 per gallon. If the trip took 78 hours, what was their average speed? ("The Math Forum @ Drexel," 2008).

This item could be revised as follows:

> A person drove a distance of 3900 miles. The trip took 78 hours. What was the average speed?

Revisions included taking out entire sentences, and using shorter sentences in the directions. Some revisions called for using bulleted lists, increasing the white space around the items, or using plain simple fonts or the use of boldface in type and headings.

Appendix C.
IRB Approval Letter

University of Denver

Sylk Sotto-Santiago, MBA
Manager, Regulatory Research Compliance Tel: 303-871-4052

Certification of Human Subjects Approval

April 30, 2008
To,
Holly Baker

Subject **Human Subject Review**
TITLE: Using Plain Language in the Colorado Student Assessment Program
IRB#: 2008-0628
Expedited Application

Dear Baker,

The Institutional Review Board for the Protection of Human Subjects has reviewed the above named project. The IRB has approved this project through an expedited review process effective null. This approval is effective for twelve months. We will be sending you a continuation application reminder for this project. This form must be completed and returned to the Office of Research and Sponsored Programs if the project is to be continued.

NOTE: Please add the following information to any consent forms, surveys, questionnaires, invitation letters, etc you will use in your research as follows: This survey (consent, study, etc.) was approved by the University of Denver's Institutional Review Board for the Protection of Human Subjects in Research on null. This information must be updated on a yearly basis, upon continuation of your IRB approval for as long as the research remains active.

The Institutional Review Board appreciates your cooperation in protecting subjects and ensuring that each subject gives a meaningful consent to participate in research projects. If you have any questions regarding your obligations under the Assurance, please do not hesitate to contact us.

Sincerely yours,

Dennis Wittmer, PhD
Chair, Institutional Review Board
for the Protection of Human Subjects

Approval Period:	04/08/2008 through 04/07/2009
Application Type - Review Type:	EXPEDITED - Designated Review - NEW
Expedited Under ParaGraph:	5
Funding:	**SPO:**
Investigational New Drug :	
Investigational Device:	
Assurance Number:	00004520, 00004520a