A FRAMEWORK FOR IDENTIFYING AND EVALUATING ITEM ALTERATIONS DESIGNED FOR PERSISTENTLY LOW PERFORMING STUDENTS.

Jennifer L. Dunn, Melissa Fincher

TRI-STATE ENHANCED ASSESSMENT

The contents of this document were developed under Enhanced Assessment Grant #S368A060005 from the U.S. Department of Education. However, those contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal government or by the host for these Web materials, the National Center on Educational Outcomes and the University of Minnesota.

Partners:
Georgia Department of Education, Hawaii Department of Education, Kentucky Department of Education; National Center on Educational Outcomes; Southeast Regional Resource Center; University of Kentucky Human Development Institute; National Alternate Assessment Center; Expert interdisciplinary Review Panel from multiple states, universities, and advocacy organizations

A FRAMEWORK FOR IDENTIFYING AND EVALUATING ITEM ALTERATIONS DESIGNED FOR PERSISTENTLY LOW PERFORMING STUDENTS.

Jennifer L. Dunn

Measured Progress

& Melissa Fincher

Georgia Department of Education

19 March 2009

Author Note

This paper will be presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA. Correspondence concerning this manuscript should be addressed to Jennifer L. Dunn.100 Education Way, Dover, NH, 03820 E-mail: Dunn.Jennifer@measuredprogress.org

A FRAMEWORK FOR IDENTIFYING AND EVALUATING ITEM ALTERATIONS DESIGNED FOR PERSISTENTLY LOW PERFORMING STUDENTS.

Abstract

In order to build a technically sound assessment, it is imperative to understand who the target population is, what the students in the target population know and can do and how to measure those skills. This study outlines an empirical investigation of the issues involved in designing an assessment based on modified academic achievement standards. The empirical investigation was designed around three main goals: (1) to better understand the lowest performing students, (2) to carefully examine the qualities of their performance, and (3) to evaluate techniques designed to make the assessment more accessible to these students. The results offer preliminary insights into (1) how tracking student performance over time can identify persistently low performing students, (2) how traditional item statistics can be used to cautiously investigate item properties when applied to a unique population, and (3) guidelines for evaluating items on the general education assessment that have been altered to increase accessibility. This research will not only shed light on the types of things that persistently low performing students are able to do, but will provide information to the federal, state, and local policymakers regarding the development of modified achievement standards.

A FRAMEWORK FOR IDENTIFYING AND EVALUATING ITEM ALTERATIONS DESIGNED FOR PERSISTENTLY LOW PERFORMING STUDENTS.

Introduction

In the context of high expectations for all students, designing a fully inclusive assessment and accountability system is critical for ensuring that all children are able to show what they know in the grade-level standards-based curriculum. Despite the national No Child be Left Behind (NCLB) initiative, students who continually struggle to demonstrate grade level expectations remain across the country. These are students for whom new ways of thinking about assessment systems may be appropriate. This line of thinking led to the development of modified achievement standards, commonly called the 2% rule. In 2007, the U.S. Department of Education (USED) issued regulations allowing assessments based on modified achievement standards, and inclusion of up to 2% of all students in AYP determinations as proficient on such an assessment (U.S. Department of Education, 2007). USED further defined the population of students as special education students who could reach grade level expectations, but would likely not reach them during the assessment year.

The USED guidelines left states with a number of options on how to go about designing an assessment based on modified academic achievement standards. Some states decided to develop a new set of assessments, while others explored how the general assessments could be improved to address the needs of this population. Under current legislation, until a state has an assessment based on modified academic achievement standards in place, these students are required to take the general education assessment. Educators have been eager to investigate whether a few of the items on the general statewide assessment can be altered in a way that would allow students who persistently

failed to meet the proficiency requirements better demonstrate their knowledge and skills (Filbin, 2008; Lazarus, Thurlow, Christensen, & Cormier, 2007). If these students are able to demonstrate what they know and can do on some but not all of the general assessment items, an empirical examination of their responses to those items, may help clarify the knowledge, skills and abilities of the students who continually struggle to meet proficiency. In addition, by identifying the types of items these students are able to answer and those they struggle with, a better understanding of their specific learning needs may be obtained. In theory, this information could then be used to target the test questions toward the students' strengths in establishing an assessment based on modified achievement standards.

The purpose of this study was to (1) establish an operational definition for students who continually struggle to meet proficiency (2) develop a set of empirically-based criteria to evaluate item properties as they apply to these struggling students, and (3) apply the criteria to evaluate the impact of the item alterations designed to increase accessibility for these students.

Defining the Population

The assessments based on modified academic achievement standards are intended for a subgroup of students who are covered under IDEA, 2004 legislation, but do not meet the participation criteria for the alternate assessment designed for students with significant cognitive disabilities. This subgroup typically comprises a population of students who have a wide range of diverse skills and abilities, whose disability has prevented them from achieving grade-level proficiency on the general education assessment and who will not reach grade-level achievement in the same timeframe as

other students. The Council of Chief State School Officers (CCSSO) attempted to improve this definition by describing these students as students whose level of performance was too high for the alternate assessment, but whose results on the general assessment were so low, no meaningful information was available for instructional programming (ASES, 2005). Unfortunately, attempts to further the above descriptions through the use of state data have been met with varying degrees of success. The Colorado Department of Education probably made the most significant progress, when they attempted to identify students scoring in the lowest third of the lowest achievement level on the general assessment, and for whom the test did not indicate progress from one year to the next (Colorado Department of Education, 2005). Unfortunately, the primary findings in Colorado's study was that the students in the lowest third of the lowest achievement level changes from year to year, and that the general assessment cannot adequately assess the progress of these students.

In order to design superior assessment systems for these students, we must first understand who these students are, by going beyond the federal definition. Although the legislation is specifically targeted towards students with disabilities, this project encompasses consistently poorly performing students both with and without disabilities. By using a definition that expands the scope of the federal legislation, we hope to gain an understanding of the knowledge skills and abilities of all students who continually struggle to meet grade level expectations. Consequently for purposes of this study, persistently low performing (PLP) students have been defined as any student who scored in the lowest performance level on three consecutive annual grade-level assessments. By going beyond the federal definition, we hope to identify all students who continually fall

below grade level expectations, develop a procedure that states could use to identify these students, and gain a better understanding of what these students know and can do.

Item Identification

In theory, by examining the item responses of PLP students, the items they struggle with and the items they are able to answer can be identified. This can lead to an enhanced understanding of the students' knowledge, skills, and abilities, the identification of specific learning needs, and improved interventions. For example, traditional item analyses techniques applied to item responses of PLP students could be used to identify items that these students tend to answer correctly, and those they tend to answer incorrectly. Unfortunately, given the expected performance of PLP students, this simplistic analysis would likely identify the easy items as those they can answer and the difficult items as those they cannot. These results would not reflect the intention of the law, or the general beliefs of professionals who work with and research the PLP population. In order to be applicable, the analyses of PLP item responses needs to follow a slightly more sophisticated approach. This can be achieved by placing the analyses of the PLP responses in the context of the general population and by expanding the traditional item analyses to include the statistical properties of the incorrect options. By attempting to detect items of reasonable difficulty for the general population that PLP students are able to answer and easy items for the general population that PLP students struggle with, we hope to discover whether or not PLP students can go beyond the easiest items on the test. This underlying philosophy was used as a cornerstone for evaluating item analyses calculated using PLP student responses.

Item responses are traditionally evaluated with classical item statistics and item response theory statistics, namely through difficulty and discrimination indices. It seemed appropriate to use these statistics as a starting point for evaluating the PLP student responses. Classical item statistics provide insight into the functionality of the items for the population from which the responses were drawn. Traditionally, acceptable items are expected to have difficulty values ranging from 0.35 to 0.95, positive discrimination indices, and negative distracter point biserials. In contrast, items with parameters outside of these ranges are considered un-ideal. Similar criteria can be applied to classical item statistics calculated using PLP student responses. Items with traditionally acceptable properties represent the types of questions PLP students are able to answer, while items with unacceptable properties represent those that PLP students struggle with. In addition, by identifying items where a particular incorrect option is selected more frequently than the correct response, further insights may be gained into the misunderstandings of the PLP students.

Unlike classical item statistics, IRT statistics require more stringent assumptions about the population distribution, but offer the advantage of placing the items and estimates of student abilities on the same scale. Not only are PLP students unlikely to meet these assumption (they are not normally distributed) but by definition there is also a mismatch between the ability of the population and the difficulties of the test items. In order to get reasonable IRT item parameter estimates for the PLP population, careful attention must be given to the location of the items in relation to the population. In theory, the PLP population should have a greater likelihood of answering the items located below the proficiency cut. That is, the easier items are better matched to the

abilities of the PLP population, and estimates for these items will be more readily obtained. However, because a primary goal in developing an assessment targeting modified achievement standards is to ensure that the resulting items do not simply form an easier test, it is critical that more difficult items are also considered. This can be achieved by examining the item parameters estimated for the general population. By considering items below the proficiency cut point, those slightly above the proficiency cut point, and those identified as having reasonable PLP classical statistics, we have identified items the PLP population is likely to answer and increased the probability that reasonable IRT parameters will be estimated using PLP student responses. Traditionally, acceptable items are expected to have positive discrimination indices and difficulties falling between -4 and +4. In theory, items with characteristics falling within these ranges represent the types of questions PLP students are able to answer, while items with characteristics that exceed these values represent those that PLP students struggle with.

Evaluating Alterations

The identification of the types of items PLP students are able to answer and those they struggle with can lead to a better understanding of the specific learning needs of these students. Items that PLP students struggle with may benefit from alterations while items that PLP students are able to answer may help to inform those alterations. Targeting the test questions toward the PLP students' strengths may represent an efficient way of establishing an assessment based on modified achievement standards. In the long run, this information could then inform item development and ultimately increase the accessibility of the test.

Although items are revised and enhanced regularly during test development, altering items to improve accessibility for the PLP population involves slightly different processes. First, the items being altered or improved have already been through strict item development criteria for the general population. It is critical that any further revisions or enhancements do not alter the construct being assessed. Careful attention should be given to how the changes influence performance of both the PLP and general populations. Second, it is critical that the impact of the revisions and enhancements be evaluated. A significant amount of time and effort is required for item revisions and enhancements, particularly when it involves working with teacher committees. In theory, the item improvements should lead to increased accessibility, and thus, improve the performance of the PLP students. Whether or not the revisions and enhancements improve the functionality of the items for the population of interest should be tested empirically.

The objectives of this study can be classified into three broad categories (1) defining the PLP population, (2) general assessment item identification for the PLP students, and (3) evaluating alterations to items on the general assessment. Each of these broad categories can be further broken down into specific research questions.

Defining the population

- What proportion of students can be classified as PLP?
- What are the demographics of the PLP population?
- Are the proportions and demographics consistent across content areas and cohorts?

Item Identification

- Are there items on the general assessment that PLP students can answer?
- What are the knowledge, skills and abilities of the PLP students?
- Are there items on the general assessment that PLP students struggle with?
- What are the knowledge, skills and abilities that PLP students struggle with?

Evaluating Alterations

- Can the items on the general assessment that the PLP students struggle with be altered so that PLP students can better demonstrate what they know and can do?

By investigating these questions, this study will help to (1) establish an operational definition for PLP students, (2) develop a set of empirically-based criteria to evaluate item properties as they apply to PLP students, and (3) apply the item identification techniques to evaluate the impact of item alterations. It is hoped that this study will improve our understanding of the students who continually struggle to meet proficiency requirements while outlining a set of systematic procedures that can be enhanced by other States as they develop modified assessments.

Sources of Information

The student enrollment file for the 2005/2006 school year from a southeastern state was used as a baseline, representing all students enrolled in grade 5 (N = 126,587) and grade 8 (N = 130,710). Mathematics and reading assessment scores from 2006, 2005 and 2004 were merged with the enrollment file using state student identification numbers.

Any records in the assessment files that were not in the 2006 baseline file were excluded from the final file. The item data for the 2006 statewide mathematics (n_{items} = 70) and reading (n_{items} = 40) assessments administered to all grade 5 and grade 8 students under standardized conditions was used for further item level analyses.

Methods and Data Sources

Defining the population

A group of students, identified as PLP were selected from the population of students that took each assessment. PLP students were defined as any student classified as non-proficient on three previous statewide assessments in the subject area of interest. Although, the operational definition of PLP is subject specific, the number and percent of students identified in both subjects was also examined.

The demographics of the general and PLP populations were examined in reading and math for both cohorts. We focused on the following demographic categories in our analyses: gender, race, free/reduced lunch, migrant status, students with disabilities, and limited English proficiency.

Item Identification

Classical statistics were calculated for all of the 2006 reading and mathematics items using PLP student responses. Items meeting the traditional criteria ($0.35 > p < 0.95$ & $r > 0$) were classified as potentially effective while items falling outside of these ranges were considered potentially problematic. Items with a distracter discrimination greater than that of the correct response or with a popular incorrect option ($p_{distracter} >$

0.35) were considered troublesome for the population at hand and thus classified as problematic.

The data for the entire population were calibrated using the 3PL IRT model as implemented in MULTILOG. The guessing parameter was of particular importance in this study since our goal was to identify items that PLP students can do, not those that they get correct by chance. These item parameters were used to locate items below and within a standard deviation of the proficiency cut.

The responses of the PLP population to items located below the proficiency cut, within a standard deviation above the proficiency cut, and those identified as having reasonable PLP classical statistics were calibrated using the 3PL IRT model as implemented in MULTILOG. Any items with positive discriminations and difficulties ranging from -/+ 4 were classified as potential. Any items with negative discriminations and difficulties exceeding -/+ 4 were classified as problematic. IRT item parameters and classical statistics were calculated for the final pool of items classified as potential using responses of the PLP population.

Evaluating Alterations

In conjunction with the technical analyses outlined above, all of the test items were examined by content experts to determine potential barriers to demonstrating proficiency. Findings of the cognitive and technical reviews were then triangulated in order to select items that could be altered in hopes of better measuring the knowledge, skills and abilities of the PLP students.

Seventy items were selected for alteration: 20 designed to measure grade 5 mathematicts, 20 designed to measure grade 8 mathematics, 17 designed to measure grade 5 reading and 13 designed to measure grade 8 reading. Alterations included both revision and enhancement. Revisions included using simplified language in item stems and response options; adding graphics and visuals; eliminating extraneous information; reformating of items; and grouping items measuring similar concepts and skills. Enhancements included providing scaffold assistance to students in the form of key definitions, reminders, and customized graphic organizers. All revisions and enhancements were carefully scruntized to ensure the construct of interest remained intact.

An unaltered version of each item and an altered version of each item were administered to approximately 5,000 students through one of two counterbalanced forms in each grade subject combination. If one form contained the altered version of the item, the other form contained the original version of the item. The number of altered items was balanced across the two forms. Eight to ten items, previously identified to have reasonable characteristics for both the general population and the PLP population were included on both forms to serve as a link between forms. Because the same items were used on both forms, the Stocking-Lord transformation allowed the forms to be placed on the same scale, thus allowing direct comparisons between the altered and unaltered version of each item. The transformation constants were calculated using the parameters estimated from the general population.

A total of four CTT and IRT item statistics were calculated for each item: (1) unaltered general, (2) altered general, (3) unaltered PLP, (4) altered PLP. The intention

was to compare the statistics of each altered item to the statistics of the item in its original form. Ideally, the altered form of an item will demonstrate improved statistics using the PLP population in comparison to the unaltered form, but similar statistics using the general population. The item identification criteria were applied to the pilot data. The impact of the item alterations were examined graphically.

Results

Defining the Population

Defining PLP students as any student scoring in the lowest performance level in all three assessments led to the identification of three percent of the grade five students in reading and four percent in math. Four percent of the grade eight students were identified as PLP in reading and nine percent in math. The number of students identified as PLP in both subjects represented approximately two percent of the total grade population (Table 1).

INSERT TABLE 1 ABOUT HERE

The operational definition of PLP is highly dependent on match rates. Many students may have been eliminated from the potential pool of low performing students because they did not have test scores for all three assessments (2006, 2005, and 2004). Since low performing students tend to be more mobile, there was some concern that a large portion of students would not be identified as PLP simply because of mobility. As a result, the number of students who did not exceed the lowest performance level and for whom there were only two scores was carefully investigated (Table 2).

INSERT TABLE 2 ABOUT HERE

Results indicate that altering the definition from scoring in the lowest performance category for three consecutive years to two consecutive years would increase the number of PLP students by four to nine percent depending on the grade and subject. Of the students in the enrolment file, scores for 79% of grade 5 students and 76% of grade 8 students were found in all three files. Given expected levels of state-wide mobility, these rates are quite acceptable. Of the students for whom scores in all three files could not be found, 1% of the grade 5 students had scores in 2005 only, 7% had scores in 2006 only, 4% had scores in 2004 and 2005, 6% had scores in 2005 and 2006, and 1% had scores in 2004 and 2005. In grade 8, 1% had scores in 2004 only, 1% had scores in 2005 only, 6% had scores in 2006 only, 5% had scores in 2004 and 2005, 6% had scores in 2005 and 2006 and 1% had scores in 2004 and 2006. Overall, the three year definition appears to be reasonable from both a data management viewpoint and the perspective that it gives educators time to address the needs of new students, before they are identified as persistently low performers.

The demographics of the students classified as PLP and the general population are outlined in Table 3. A disproportionately high number of males were identified as PLP relative to the general population. This trend was particularly true in reading. The PLP group included a larger proportion of African American and Hispanic students and a lower proportion of white students in both reading and mathematics, although the trend was more evident in reading. The PLP group also included meaningfully larger proportions of students who are eligible for the free/reduced lunch program.

INSERT TABLE 3 ABOUT HERE

Not surprisingly, the PLP populations included substantially more students with disabilities than the general population, to the point where slightly more than half of the

students identified as PLP were also identified as having a disability. More surprising is the fact that slightly less than half of the PLP population was not identified as having a disability. This result conflicts with the federal definition of the 2% population that requires students to be identified as having a disability in order to be eligible for inclusion in the modified assessment.

Item Identification

The average item characteristics of the items identified as potential and problematic for the PLP population are outlined in Tables 4 (difficulty) and 5 (discrimination). The numbers displayed in both tables are based on estimates calculated using the general population. The items classified as potential tend to represent the easier items from the original test. In contrast, the items classified as problematic tend to be the more difficult items on the test. It is not surprising that the items suited for the PLP students tend to be the easier items while the items they struggle with tend to be more difficult. Although the items classified as potential are slightly less discriminating than those on the whole test, while the problematic items are slightly more discriminating, the difference does not appear to be meaningfully significant.

INSERT TABLES 4 AND 5 ABOUT HERE

The distribution of items by content strand for the items classified as potential is displayed in Tables 6 & 7 for math and reading respectively. While the balance of representation seems to be fairly consistent with the total test for grade 8 reading and mathematics, the same cannot be said for grade 5. Grade 5 PLP students appear better able to answer the computation and estimation questions but have difficulty with the problem solving and statistics and probability questions. In reading, the PLP students

appear to have more success with the items requiring the students to read for information and less success with the items that involve reading for comprehension.

INSERT TABLES 6 AND 7 ABOUT HERE

The distribution of items by content strand for the items classified as problematic for PLP students is displayed in Tables 8 & 9 for math and reading respectively. Overall, the balance of representation of the problematic items appears to approximately parallel that of the total test, with the exception of grade 8 mathematics. In this latter case, it appears that the PLP students have difficulty with the problem solving, as eight of the fourteen problem sovling items were flagged as problematic.

INSERT TABLES 8 AND 9 ABOUT HERE

Evaluating Alterations

The number of students who took each form of the pilot test and the percent of PLP students is outlined in Table 10. It appears from these numbers that a larger number of PLP students participated in the pilot study than would have had the participation been random. This over-representation was intentional and is likely due to the fact that the educators responsible for selecting the students were aware of the study's purpose.

INSERT TABLE 10 ABOUT HERE

The number of altered items on the pilot test forms identified as having reasonable item characteristics is outlined in Table 11. The first column outlines the number of items for each grade content combination that was altered. The second column outlines the number of altered items identified as having reasonable item characteristics. The classification of these items in their original state is outlined in columns three and four. Across all of the grade contents, more than half of the altered items are classified as

potential. Unfortunately, some of these items were also potential in their original form, which means that at this point, our conclusions about the alterations are limited to that they did no harm. Of particular interest are the classification results of grade 8 reading where all of the altered items were classified as potential.

INSERT TABLE 11 ABOUT HERE

Unfortunately, not all of the item alterations had a positive effect. The number of items identified as having problematic item characteristics post alteration is outlined in Table 12. This table mirrors that of Table 11, with the exception that it illustrates the ineffective alterations. Although, more of the altered items were classified as potential (Table 11), some of the alterations changed an item classified as potential in its original form to a classification of problematic in its altered form. Seven of 70 items moved from a classification of potential pre-alteration, to a classification of problematic post alteration while 3 items moved from being neither potential nor problematic to being problematic after alteration. In contrast, 9 of 70 items moved from a classification of problematic pre-alteration to a classification of potential post-alteration while 10 of 70 items moved from no classification pre-alteration to a classification of potential post-alteration. The classification of the remaining 41 items was unaffected by the item alterations.

INSERT TABLE 12 ABOUT HERE

In an attempt to understand the impact of the item alterations on the items whose classification did not change, the item alterations were further explored graphically. Figure 1 outlines the effects of the item alterations in grade 5 math for the entire population (solid line), and the PLP (dashed lines) population. The altered item parameters are illustrated with a blue line while the non-altered items are illustrated with

a grey line. An ideal item change is one where there is very little difference between the solid lines but large differences between the dashed lines. Within these differences, the blue dashed line should have a steeper curve and possibly be shifted to the left of the grey dashed curve. An increase in the slope indicates the item is more discriminating after the revision/enhancement and a shift to the left indicates more PLP students were able to answer the item correctly. Similar comparisons for the altered items in grade 5 math, grade 8 reading and grade 8 math are displayed in Figures 2 through 4, respectively.

INSERT FIGURES 1 THOUGH 4 ABOUT HERE

For the most part, it appears that the alterations had either a positive or no effect on slightly more than half of the items in grade 5 and grade 8 math, on two thirds of the items in grade 5 reading, and on all of the items in grade 8 reading. The alterations also appear to have had a much greater impact on the PLP responses than on those of the general population. This trend appears to be particularly true in math, where there is very little difference between the black and blue (solid lines) item curves for the general population but greater differences between those for the PLP population (dashed lines). In contrst, in grade 5 reading, much larger differences between the item curves for the general population are evident indicating that the item alterations assisted the general population as well as the PLP population. Interstingly, in grade 8 reading, where the item alterations had the greatest effect for the PLP population, with a few exceptions, the alterations did not tend to affect item properties for the general population. The item alterations in grade 8 reading appear to have been more successful than the other grade contents. Of the four grade content areas, grade 8 reading was the only place where passages were “segmented” as an alteration.

In grade 8 reading, two of the passages were segmented on one form but not on the other. Although some of the passage items were also revised or enhanced on one of the two forms (making it impossible to isolate the effect of segmenting the passage from the item revision/enhancement), seven items did not change from one form to the other. Because the items did not change, any differences between the item properties on the two forms can be attributed to changes in the format of the passage. The IRT characteristics of these items are displayed in Figure 5. Comparisons among the seven items reveals that although an improvement was not evident across all of the items for the PLP population, non of the items appeared to become meaninfully worse. That is, the PLP students were able to answer the items in both cases, indicating that the segmenting, at worst, "did no harm". In addition, the properties of four of the seven items were very similar for the general population despite segmenting. Although this trend did not occur for all items and must be interpreted with caution, passage segmenting may represent an alteration that is benificial for PLP students but has a limited impact on the general population.

INSERT FIGURE5 ABOUT HERE

Discussion

This study was founded on the principle that systematic investigation of performance will identify options that will lead to improved assessment of students who continually struggle to meet the proficiency requirements while maintaining high expectations. The empirical investigation was designed around three main goals: (1) to better understand the lowest performing students, (2) to carefully examine the qualities of their performance, and (3) to evaluate assessment techniques designed to make learning and assessment more accessible to all students.

Persistently low performing students were defined as any students who scored in the lowest performance level in three consecutive assessments. The persistently low performing population consisted of a larger proportion of males, African Americans, students eligible for the free/reduced lunch program and students with mild intellectual disabilities than the full population. In reading, but not math, a larger proportion of LEP students were also classified as persistently low performing. Of critical importance in these findings is that PLP students are not exclusively special education students.

The statistical approaches identified items the PLP students were able to answer correctly, indicating that they do have some knowledge of the grade level content. In addition, certain trends were identified in terms of the types of items they were able or unable to answer. In math, PLP students tended to have an easier time with the computation questions while struggling with the problem solving questions. In reading, PLP students tended to have an easier time with information-based questions while struggling with comprehension questions.

Despite the significant amount of careful thought and attention that was given to the types of revisions and enhancements made to the items, the effect was somewhat inconclusive. While the majority of the changes had a positive impact on the ability of the PLP students to respond to the items, some of the changes appear to have made the items more difficult. Although the overall effectiveness of the revisions and enhancements is inconclusive, item alterations should not be eliminated as an approach for developing assessments based on modified achievement standards. The use of "segmenting" as a passage alteration showed promising results that merit further investigation. Perhaps more can be learned about alterations designed to increase accessibiltiy for the PLP

population through a think aloud study designeed to understand the students thought processes as they are answering the items.

This study was in part proposed as an empirical approach for identifying items that could be better designed to assess the knowledge and skills of the 2% population. In theory, by identifying items that this population can answer, educators can learn how to adjust inaccessible items to better target the students skills. Given that the federal guidelines for modified academic achievement standards are targeted towards students with IEPs, the differences between the PLP students with and without IEPs merits further investigation.

The differential item functioning (DIF) between the two groups of students was examined for the items using a two step process, utilizing both the Mantel-Haenszel (Holland and Thayer, 1988) and standardization (Dorans and Kulick, 1986) procedures. Both of these procedures calculate the difference in item performance for groups of students matched for achievement on the total test. In the first step, the Mantel-Haenszel procedure is used to identify items that show statistically significant DIF. However, because of the large number of students on which the calculations tend to be based, the majority of items tend to indicate a statistically significant difference between the focal and reference groups. (Note that this issue is not specific to Mantel-Haenszel calculations. Large enough sample sizes will indicate that even trivially small results are statistically significant. For this reason, in the second step of the process, the standardization procedure is used to categorize items according to the amount of DIF detected.) The number of items identified with DIF is outlined in Table 13.

INSERT TABLE 13 ABOUT HERE

According to the results depicted in Table 13, very few of the items are showing signs of DIF. Four of the items are showing low DIF, and only one item shows high DIF. Further investigation into the grade 5 reading item with high DIF revealed an advantage to the students identified as having a disability. Overall, there does not appear to be meaningful differences between the item performance of PLP students who have been identified as having a disability and those who have not. When the item is accessible to these students, their disability does not appear to be influencing students' capability of responding to the item. This finding has some policy implications for the general legislation, namely that we may want to look beyond students with IEPs when designing inclusive assessments.

Limitations

This study offers a series of technical guidelines for identifying a specialized population, investigating the item characteristics as they apply to the population, and evaluating the effectiveness of item alterations designed to increase the accessibility for said population. Traditional statistics, calculated using the student responses of the PLP population, were able to provide insight into items that this population of students was able to answer and those they were not. It was assumed that insight into the knowledge, skills, and abilities of the PLP students could be gained through careful examination of the proportions of PLP students who selected the correct option and the relationship between each item score and the total score. In contrast, it was also assumed that insights into the knowledge, skills and abilities lacking in the PLP students could be gained by examining the predominant distracters selected by all of the PLP students and the high ability PLP students. The classical statistics used to make these assumptions are generally

considered appropriate regardless of the underlying distribution, although the interpretation of the statistic changes as the distribution changes. More specifically, the interpretation of item difficulty (the proportion of students who get the item correct) needs to be adjusted accordingly (to the proportion of PLP students who get the item correct) when calculated using only PLP students. For example, we would expect the items to be more difficult for the PLP students, resulting in lower item difficulty statistics for the PLP students than for the general population students.

IRT parameters were also used to examine the knowledge, skills and abilities of the PLP students. A very liberal range of acceptable parameters was used to identify items. In contrast to the classical statistics, IRT requires that certain assumptions be met, one of which is the assumption of a normal underlying distribution. The PLP population clearly violates this assumption, indicating that any interpretations based on the IRT parameters must be made with caution. Although matching the student abilities to the items (selecting items below the proficiency cut) was an attempt to address and correct for the lack of normal distribution, the correction was likely not sufficient.

Regardless of the specific statistics (IRT or classical), the interpretations linked to the statistics should be done cautiously. This study did not investigate the stability of the statistics, and consequently, they should not be interpreted as truth. For example, the statistics calculated using the PLP students might vary substantially had the sample been randomly split in two, or had a different group of PLP students been used in the estimation. Many of the values encountered in this study would traditionally be seen as problematic, bearing further investigation. For example, an IRT difficulty value of greater than ten (of which there were a few) would typically be interpreted as an item that did not

converge, for which the parameters cannot be considered valid. There may be more error than interpretable value in many of the statistics. Investigating the stability of these statistics across samples of PLP students would be a worthwhile endeavor that would help to verify (or negate) the interpretations made in this study.

Educational or scientific importance of the study

This study offers a series of technical guidelines for investigating item characteristics for a specialized population. Although the stability of these results merits further investigation, it offers preliminary insights into (1) how tracking student performance over time can lead to an operational definition of persistently low performing students, (2) how traditional item statistics can be used to cautiously investigate item properties when applied to a unique population, and (3) guidelines for evaluating items on the general education assessment that have been altered to increase accessibility. This research has not only shed light on the types of things that PLP students are able to do but should inform federal, state, and local policy makers regarding the development of modified achievement standards. The technical analyses specified here could provide a framework for states to use in the future when designing their own modified assessments.

References

American Institute of Research (AIR) (2000). *Effects of item scaffolding on student responses: A cognitive laboratory study*. Washington, DC: AIR for Institutes for Research for the National Assessment Governing Board in support of contract #RJ97153001.

Filbin, J. (2008). *Lessons from the initial peer review of alternate assessments based on modified achievement standards.* Washington, DC: Office of Elementary and Secondary Education, U.S. Department of Education

Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners* (Technical Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Johnstone, C., Liu, K., Altman, J., & Thurlow, M. (2007). *Student think aloud reflections on comprehensible and readable assessment items: Perspectives on what does and does not make an item readable* (Technical Report 48). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Lazarus, S. S., Thurlow, M. L., Christensen, L. L., & Cormier, D. (2007). *States' alternate assessments based on modified achievement standards (AA-MAS) in 2007* (Synthesis Report 67). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

U.S. Department of Education (2007, July 20). *Modified Academic Achievement Standards: Non-regulatory Guidance.* Washington, DC: Office of Elementary and Secondary Education, U.S. Department of Education. (Available at: http://www.ed.gov/admins/lead/account/saa.html#regulations).

Table 1. Identified as Persistently Low Performing

Grade	Subject	Number of test takers	Persistently Low Performing	
			N	%
5	Reading	115,415	3,953	3
	Math	115,826	4,976	4
8	Reading	116,501	5,042	4
	Math	116,733	12,238	9

Table 2. Baseline Match Rates

			Not Matched					
Grade	N	Matched	2004	2005	2006	2004 & 2005	2005 & 2006	2004 & 2006
5	126,587	79%	0%	1%	7%	4%	6%	1%
8	130,710	76%	1%	1%	6%	5%	6%	1%

Table 3. Persistently Low Performing Demographics

	Baseline				**Grade 5**				**Grade 8**			
	Grade 5		Grade 8		Reading		Math		Reading		Math	
	N	%	N	%	N	%	N	%	N	%	N	%
Gender												
Females	61,802	49%	63,895	49%	1,432	36%	2,085	41%	1,780	35%	5,469	45%
Males	64,785	51%	66,815	51%	2,584	64%	2,990	59%	3,272	65%	6,783	55%
Race												
Asian/Pacific Islands	3,612	3%	3,479	3%	42	1%	25	0%	59	1%	65	1%
Black	48,769	39%	52,520	40%	2,343	58%	3,066	60%	3,205	63%	7,797	64%
Hispanic	11,492	9%	10,017	8%	641	16%	526	10%	619	12%	996	8%
Indian/Native	156	0%	215	0%	3	0%	4	0%	3	0%	9	0%
Multi-Racial	3,336	3%	2,651	2%	65	2%	96	2%	48	1%	141	1%
White	59,195	47%	61,828	47%	922	23%	1,358	27%	1,118	22%	3,244	26%
Free/Reduced Lunch												
Not Identified	62,242	49%	68,197	52%	745	19%	1,054	21%	998	20%	3,043	25%
Free/Reduced Lunch	64,345	51%	62,513	48%	3,271	81%	4,021	79%	4,054	80%	9,209	75%
Migrant												
Not Identified	126,118	100%	130,287	100%	3,969	99%	5,035	99%	5,015	99%	12,200	100%
Migrant	469	0%	423	0%	47	1%	40	1%	37	1%	52	0%
Students with Disabilities												
Not Identified	109,337	86%	114,254	87%	2,035	51%	2,174	43%	2,531	50%	7,110	58%
SWD	17,250	14%	16,456	13%	1,981	49%	2,901	57%	2,521	50%	5,142	42%
Limited English Proficiency												
Not Identified	120,838	95%	126,291	97%	3,524	88%	4,703	93%	4,548	90%	11,620	95%
LEP	5,749	5%	4,419	3%	492	12%	372	7%	504	10%	632	5%

Table 4. The mean item difficulty of identified items across grade contents.

	Whole Test			Potential Items			Problematic Items		
	Number of Items	Difficulty *M*	*SD*	Number of Items	Difficulty *M*	*SD*	Number of Items	Difficulty *M*	*SD*
Grade 5 Math	70	-0.67	0.97	30	-1.43	0.79	25	-0.33	0.59
Grade 8 Math	70	-0.55	0.97	38	-1.24	0.72	21	0.14	0.44
Grade 5 Reading	40	-0.43	0.77	14	-1.04	0.39	22	-0.33	0.40
Grade 8 Reading	40	-0.70	0.94	15	-1.38	0.97	14	-0.74	0.39

Table 5. The mean item discrimination of identified items across grade contents.

	Whole Test			Potential Items			Problematic Items		
	Number of Items	Discrimination *M*	*SD*	Number of Items	Discrimination *M*	*SD*	Number of Items	Discrimination *M*	*SD*
Grade 5 Math	70	0.88	0.29	30	0.77	0.21	25	0.93	0.27
Grade 8 Math	70	0.92	0.30	38	0.82	0.27	21	1.10	0.30
Grade 5 Reading	40	0.82	0.32	14	0.81	0.25	22	0.84	0.38
Grade 8 Reading	40	0.78	0.32	15	0.82	0.34	14	0.72	0.30

Table 6. Mathematics balance of representation for the total test and the items classified as effective for PLP students

	Grade 5		Grade 8	
	Total Test (70)	Potential Items (30)	Total Test (70)	Potential Items (38)
Computation & Estimation	21%	33%	10%	11%
Geometry & Measurement	17%	20%	20%	21%
Number Sense & Numeration	20%	23%	14%	13%
Patterns & Relationships/Algebra	11%	13%	20%	26%
Problem Solving	20%	7%	20%	11%
Statistics & Probability	10%	3%	16%	18%

Table 7. Reading balance of representation for the total test and the items classified as effective for PLP students

	Grade 5		Grade 8	
	Total Test (40)	Ineffective Items (14)	Total Test (40)	Ineffective Items (15)
Functional & Media Literacy	15%	7%	18%	13%
Information	30%	43%	43%	40%
Literacy Comprehension	35%	25%	25%	27%
Skills and Vocabulary	20%	14%	15%	20%

Table 8. Mathematics balance of representation for the total test and the items classified as problematic for PLP students

	Grade 5		Grade 8	
	Total Test (70)	Ineffective Items (30)	Total Test (70)	Ineffective Items (38)
Computation & Estimation	21%	8%	10%	10%
Geometry & Measurement	17%	20%	20%	14%
Number Sense & Numeration	20%	20%	14%	19%
Patterns & Relationships/Algebra	11%	12%	20%	14%
Problem Solving	20%	24%	20%	38%
Statistics & Probability	10%	16%	16%	5%

Table 9. Reading balance of representation for the total test and the items classified as problematic for PLP students

	Grade 5		Grade 8	
	Total Test (40)	Ineffective Items (14)	Total Test (40)	Ineffective Items (15)
Functional & Media Literacy	15%	14%	18%	29%
Information	30%	23%	43%	43%
Literacy Comprehension	35%	36%	25%	21%
Skills and Vocabulary	20%	27%	15%	7%

Table 10. Sample of Pilot students identified as Persistently Low Performing

Grade	Subject	Number of test takers		Persistently Low Performing			
		Form 1	Form 2	Form 1		Form 2	
				N	%	N	%
5	Reading	1,894	1,831	133	7	102	6
	Math	1,886	1,825	146	8	124	7
8	Reading	1,826	1,783	114	6	93	5
	Math	1,827	1,780	254	14	218	12

Table 11. Summary of effective item changes

	Number of items changed	Changed items classified as Potential	Original classification of changed items now classified as potential	
			Potential	Problematic
Grade 5: Math	20	10	6	4
Grade 8: Math	20	11	7	3
Grade 5: Reading	17	13	8	2
Grade 8: Reading	13	13	9	0

Table 12. Summary of ineffective item changes

	Number of items changed	Changed items classified as Problematic	Original classification of changed items now classified as problematic	
			Potential	Problematic
Grade 5: Math	20	8	2	3
Grade 8: Math	20	8	2	6
Grade 5: Reading	17	3	3	0
Grade 8: Reading	23	0	0	0

Table 13. A summary of Mantel-Haenszel DIF results comparing the performance of PLP students identified as having a disability to PLP students who have not been identified as having a disability

	Total Number of PLP Students	Number of PLP Students with IEPs	No DIF	Low DIF	High DIF	Total Number of Test Items
Grade 5 Math	4,976	2,876	29	1	0	30
Grade 8 Math	12,238	5,135	36	2	0	38
Grade 5 Reading	3,953	1,963	13	0	1	14
Grade 8 Reading	5,042	2,515	14	1	0	15

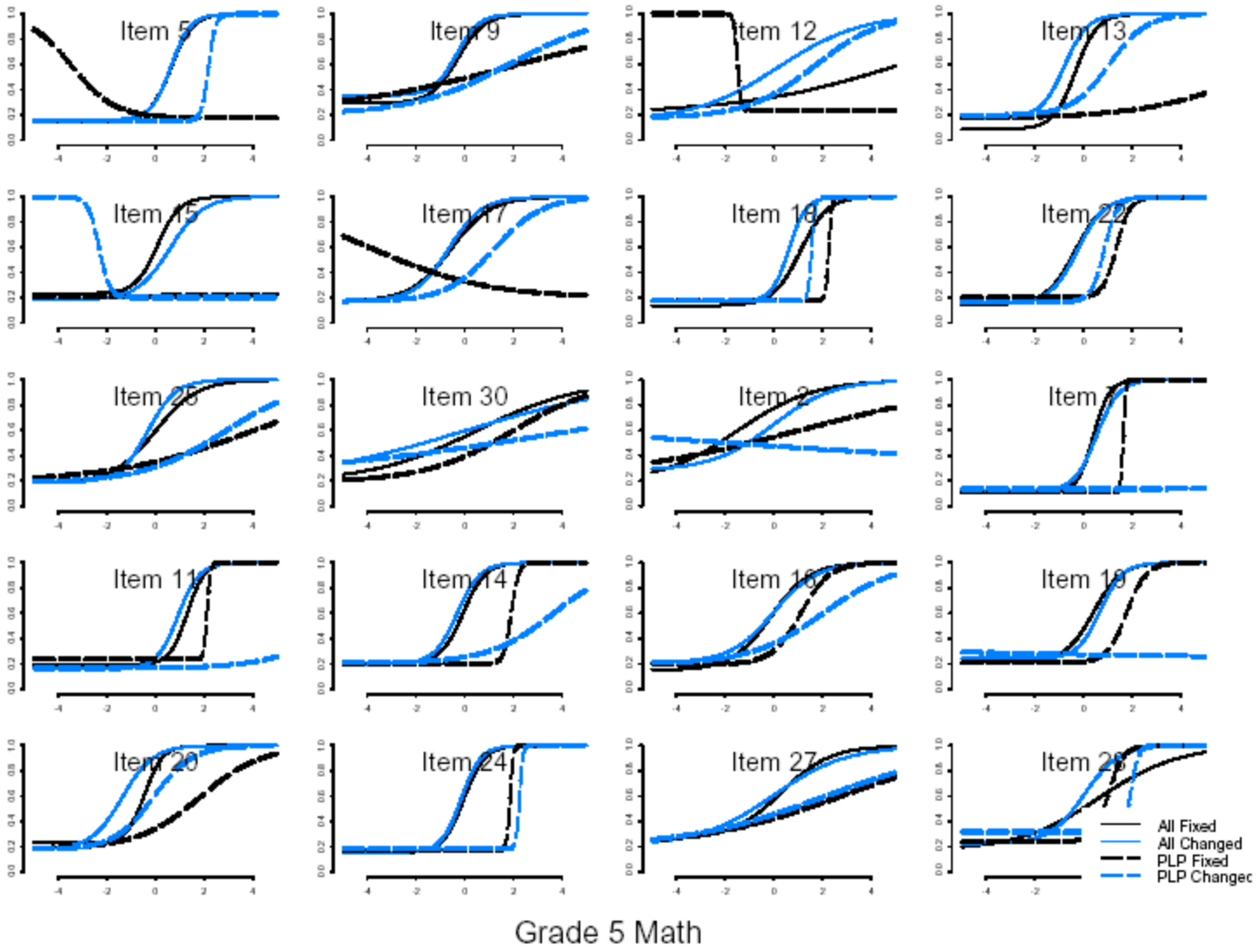

Figure 1. A comparison of item alterations for all students and persistently low performing students in grade 5 math.

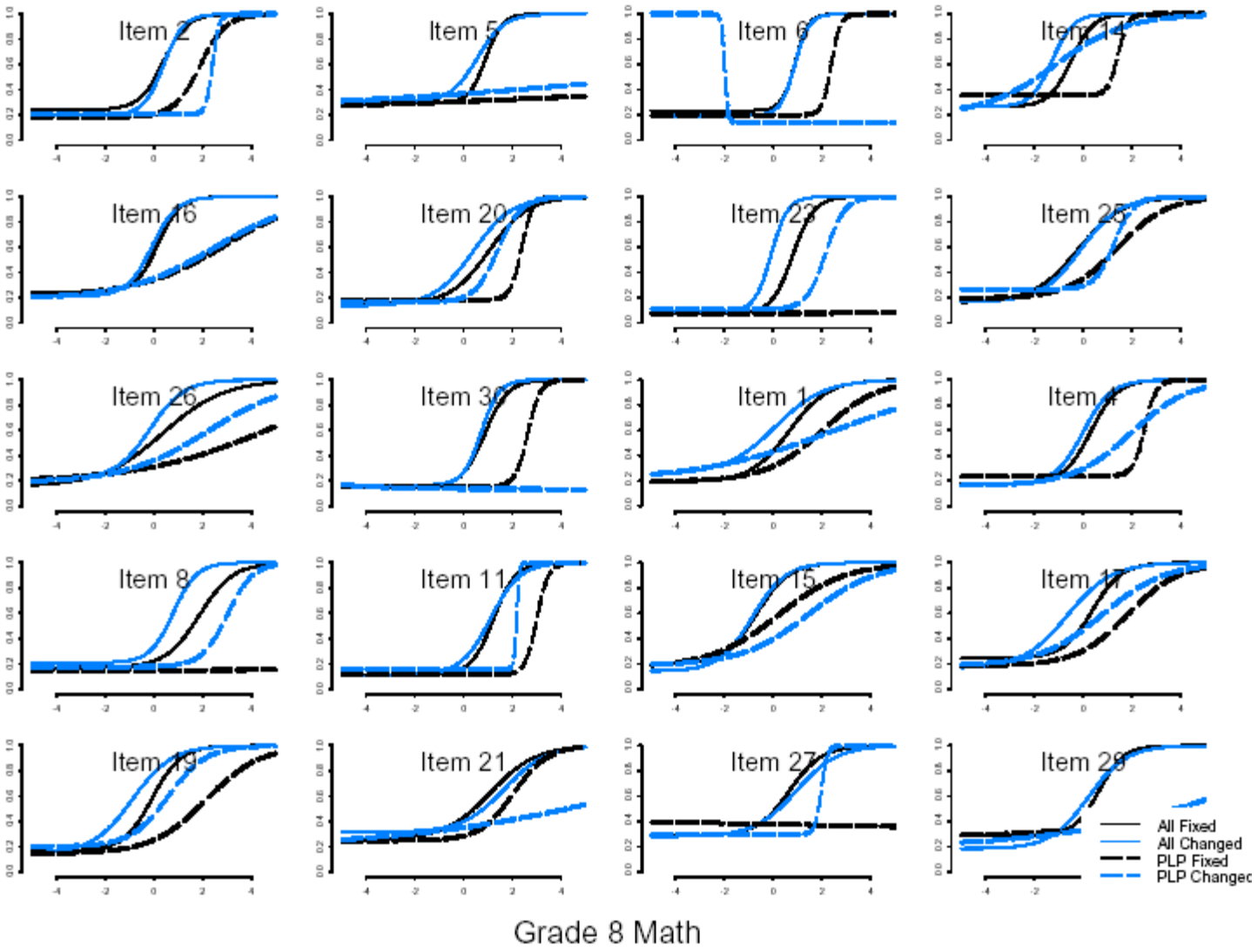

Figure 2. A comparison of item alterations for all students and persistently low performing students in grade 8 math.

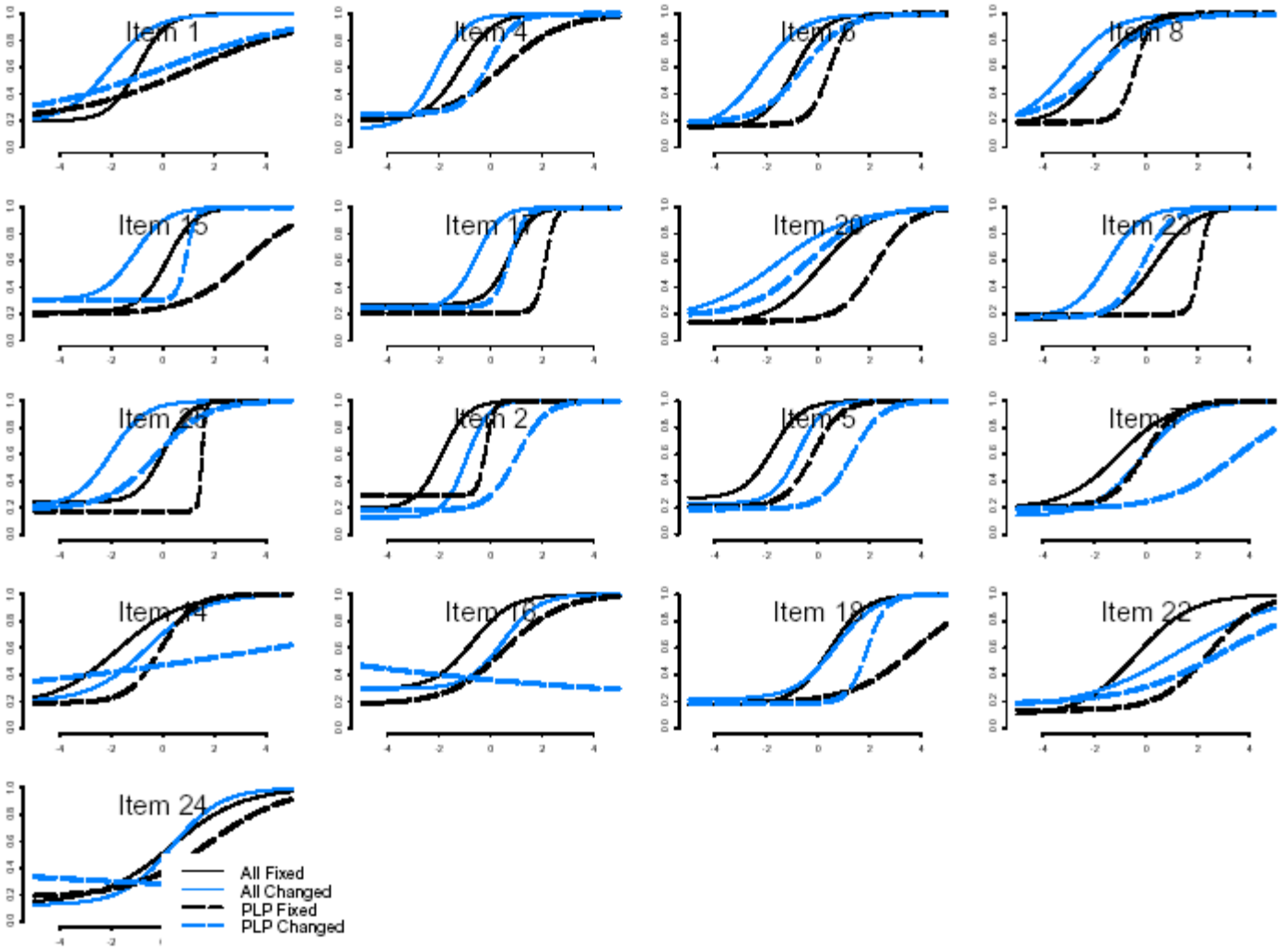

Figure 3. A comparison of item alterations for all students and persistently low performing students in grade 5 reading.

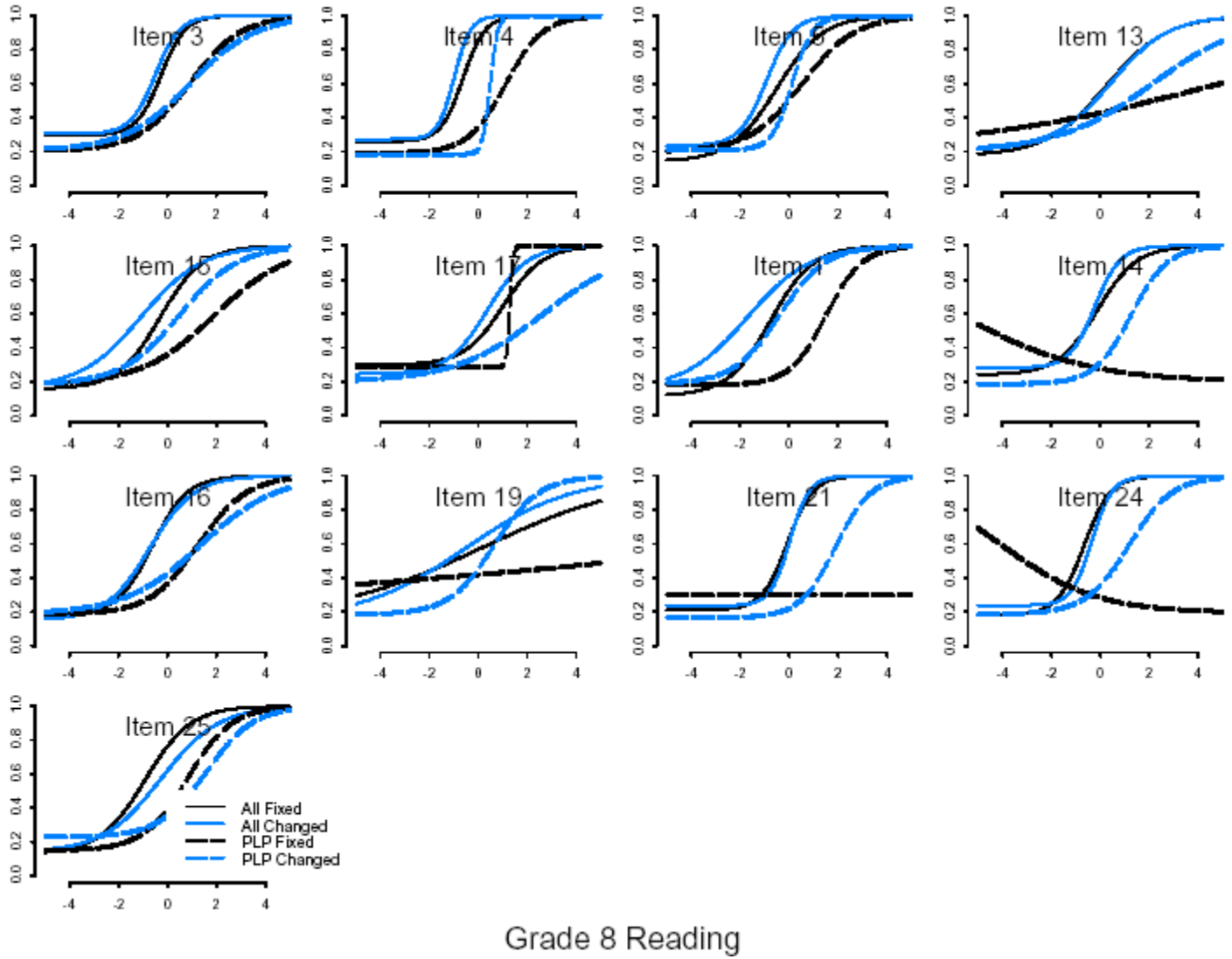

Figure 4. A comparison of item alterations for all students and persistently low performing students in grade 8 reading.

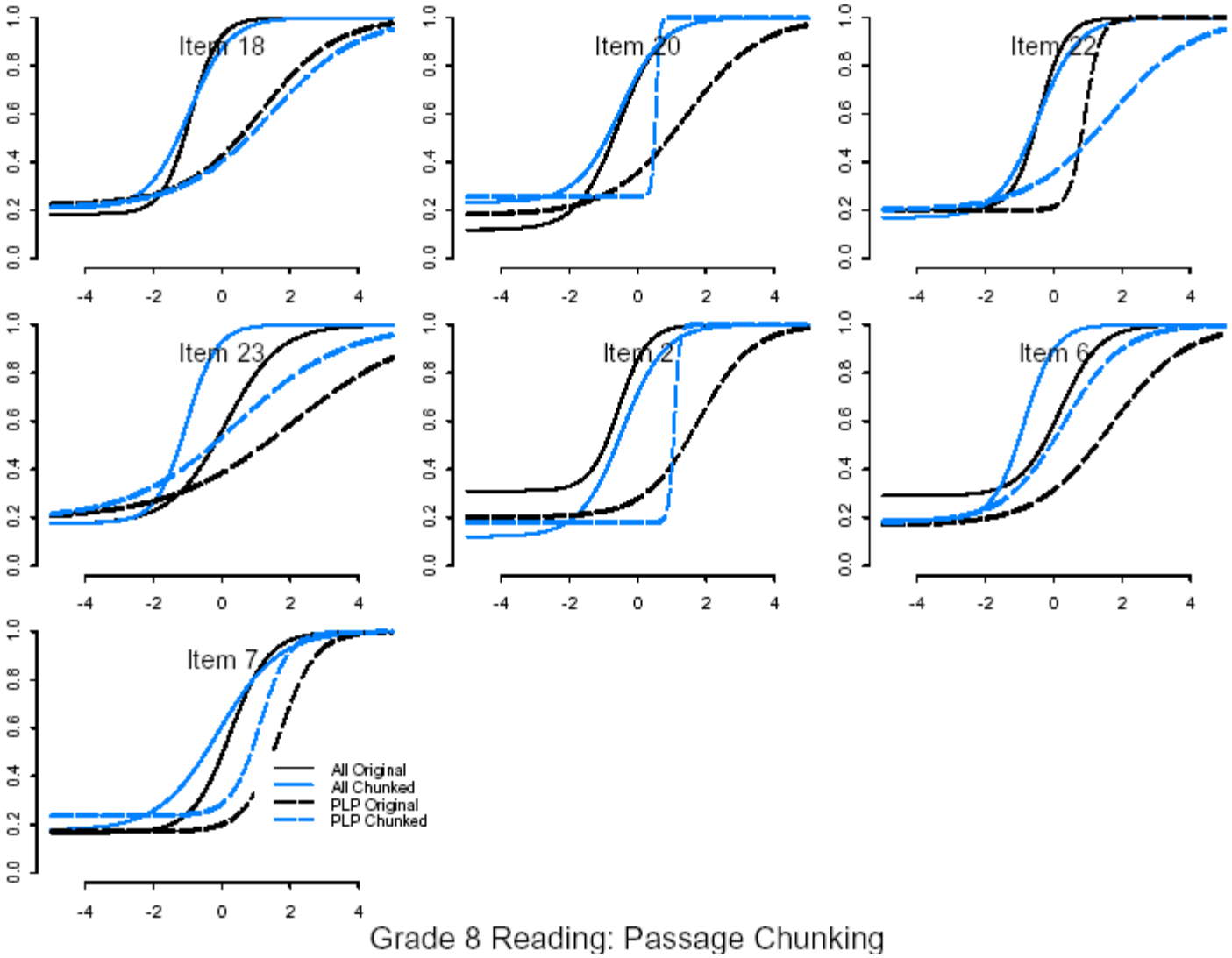

Figure 5. A comparison of the effect of passage segmenting in grade 8 reading for all students and persistently low performing students in grade 8 reading.