Running Head: Accommodation Beliefs

Using DIF Analyses to Examine Several Commonly-Held Beliefs about Testing Accommodations for Students with Disabilities

Sara E. Bolt

National Center on Educational Outcomes (NCEO)

University of Minnesota, Twin Cities

Paper presented at the annual conference of the

National Council on Measurement in Education (NCME)

San Diego, CA

April 13, 2004

Abstract

In order to examine the extent of data-based support for several commonly-held opinions about testing accommodations for students with disabilities, a series of DIF analyses were conducted across three statewide achievement tests. First, the belief that the read-aloud accommodation is less appropriate for reading tests than for non-reading tests was examined by comparing the extent of DIF present on reading and math tests among students with disabilities receiving this accommodation. Second, the belief that accommodations are more appropriate for students with sensory and physical disabilities than for students with cognitive disabilities was examined by comparing the extent of DIF present across test administrations for these two groups of students. Finally, many people have argued that accommodations fall on a continuum in terms of the degree to which they affect the validity of a test. In order to investigate this belief, the extent of DIF present for several different accommodated administrations, some of which are commonly considered appropriate and some of which are commonly considered inappropriate, was compared. Results provide some support for the commonly-held beliefs, although results were not always consistent across datasets. The results also point to the challenge of appropriately assessing the skills and knowledge of students with disabilities using currently available assessments.

The No Child Left Behind Act of 2001 emphasizes the importance of improving educational results for *all* students by including through the specific inclusion of certain student subgroups in large-scale assessment and accountability systems. In the past, students with disabilities were a subgroup of students commonly excluded from these systems (Heubert & Hauser, 1999). Today, students with disabilities are not only required to participate in state- and district-wide assessments, but their scores are to be used in determining whether a school or district has made adequate yearly progress (AYP). This information is subsequently used to determine certain consequences for schools and districts. To avoid negative consequences, districts need to show that their students with disabilities are making progress toward the goal of 100 percent proficient by the year 2013-14, as measured by state- and district-wide assessments. It is clear that the performance of students with disabilities on state- and district-wide assessments is now of great concern across the nation.

Although it is now clear that students with disabilities must participate in state- and district-wide assessments, questions remain about how they can most appropriately and effectively participate. Many of the assessments that are currently used for accountability purposes were developed without careful consideration of the needs of many students with disabilities. As a result, some students with disabilities have difficulty demonstrating what they know and can do under standard testing conditions. Testing accommodations have been proposed as a way to allow many students with disabilities better access to test content. IDEA 1997 requires the use of testing accommodations, as necessary, to allow students with disabilities to demonstrate what they know and can do on tests.

The provision of testing accommodations to students with disabilities has corresponded to an increase in the percent of students with disabilities participating in state- and district-wide assessments and accountability systems (Koretz, 1997; Olson & Goldstein, 1996). However, empirical support for specific testing accommodations remains somewhat limited (Tindal & Fuchs, 1999). Given that testing accommodations represent changes in standardized testing conditions, there is concern that they may alter what the test measures among those who receive them. Testing accommodations are intended to improve measurement of intended skills among students with disabilities; however, empirical support is necessary to show that they serve this intended purpose. Denying appropriate accommodations to students with disabilities can limit them from demonstrating true knowledge, which may result in schools being inappropriately penalized for low scores. At the same time, the provision of inappropriate accommodations to students with disabilities may reduce accountability for them to learn the skills deemed necessary for success.

Investigating the effects of testing accommodations for students with disabilities is an overwhelming task, given the many different variables that may influence results. The appropriateness of a testing accommodation may depend on characteristics of the student, the skills intended to be measured, the specific accommodation, and the student's prior exposure to the accommodation. A variety of approaches have been developed to study whether an accommodation is appropriate (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000), and support for an accommodation may vary depending on the approach that is used. For instance, some researchers have suggested that an accommodation is appropriate if it shows "differential boost" for students with disabilities, such that it

boosts the scores of students with disabilities significantly more than the scores of students without disabilities (Fuchs et al., 2000). Other researchers have studied the appropriateness of testing accommodations by looking at item-level characteristics to better understand how the accommodation does or does not influence measurement comparability (Koretz, 1997; Pomplun & Omar, 2000). Although research is beginning to accumulate, it has not provided clear direction as to the appropriateness or inappropriateness of specific test accommodations.

Without sufficient empirical evidence, policymakers and educators have been forced to make decisions about the appropriateness of testing accommodations on the basis of opinions and beliefs. Some accommodations are widely agreed to maintain the validity of a test; others are widely agreed to substantially alter the validity of a test. The appropriateness of several other accommodations is hotly debated, and accommodation policies vary considerably in terms of whether many individual accommodations are considered “appropriate” and “standard” (Thurlow, Lazarus, Thompson, & Robey, 2002). Decisions about which accommodations an individual student will receive are ultimately made by individualized education program (iep) teams, which are composed of parents, teachers, the individual student, and various other school professionals (administrators, school psychologists, consultants, etc.). Although accommodation decisions are often guided by state policies, iep teams are typically encouraged to make decisions based on the specific needs of the individual. Research has indicated that educators do not accurately identify which accommodations will be effective for an individual student (Fuchs et al., 2000; Helwig & Tindal, 2003).

Given that opinions and beliefs can be wrong, it is important to examine the extent to which they are supported by data. The purpose of the current study was to identify through an analysis of actual test data the extent to which several commonly-held beliefs about testing accommodations are supported. The commonly-held beliefs targeted in this study are the following: 1) the read-aloud accommodation is more appropriate on a math test than a reading test; 2) the appropriateness of accommodations varies by disability type (e.g., sensory/physical vs. cognitive disability), and 3) accommodations fall on a continuum in terms of their appropriateness (i.e., some are highly appropriate, some are highly inappropriate, and some are moderately appropriate). Each of these beliefs is described in the following sections, along with currently available empirical evidence.

Belief One: The Read-Aloud Accommodation is More Appropriate on a Math Test than on a Reading Test

Although 46 states allow the read-aloud accommodation (i.e., having test directions, items, responses, and other stimulus material read aloud to the student) on some tests, over 30 states prohibit its use or consider it an inappropriate accommodation on reading tests (Thurlow et al., 2002). In addition, when researchers and policymakers were asked to place accommodations into one of three categories based on the degree to which they were perceived to influence the validity of a test, many respondents placed the read-aloud accommodation in separate categories depending on the content of the test.

Research on the effectiveness of the read-aloud accommodation for math tests has had mixed results. Some studies have demonstrated differential boost for students with

disabilities (Weston, 1999; Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998), others have not (Meloy, Deville, & Frisbie, 2002; Schulte, Elliott, & Kratochwill, 2001). Another study found that differential boost was evident on some test items, but not others (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000). Some studies have shown measurement comparability for those receiving and not receiving the accommodation (Lewis, Green, & Miller, 1999; Pomplun & Omar, 2000), although another study has questioned whether the accommodation was necessary to improve measurement on a math test for many students with reading disabilities (Bolt & Bielinski, 2002).

Empirical evidence for the appropriateness of the read-aloud accommodation on a reading test has not substantially supported this accommodation. Three studies were identified that examined differential boost for students with disabilities; none of these found that there was significant differential boost for students with disabilities (Kosciolek & Ysseldyke, 2000; McKevitt & Elliott, 2003; Meloy, Deville, & Frisbie, 2002). Despite the fact that Tippets & Michaels (1999) found similar factor structures for a reading test among students receiving and not receiving this accommodation, other studies have identified large differences in measurement characteristics for students receiving the read-aloud accommodation on a reading test compared to non-accommodated students without disabilities (Bielinski, Thurlow, Ysseldyke, Friedebach, & Friedebach, 2001; Lewis et al., 1999).

Belief Two: The Appropriateness of Accommodations Varies by Disability Type (e.g., Sensory/Physical vs. Cognitive Disability)

Because academic achievement tests are not typically intended to measure physical or sensory skills, it is commonly considered appropriate to provide

accommodations to students with impairments in these areas. However, academic achievement tests are typically intended to measure cognitively-related skills among students. There consequently tends to be more concern with accommodating students with cognitive disabilities. Phillips (1994) suggests that accommodations are often considered inappropriate for students with cognitive disabilities because 1) these students are less distinguishable from students without disabilities than students with sensory or physical disabilities, 2) it is believed that both students with and without these disabilities would likely benefit from the accommodations, and 3) skill deficits associated with these disabilities are often intertwined with the skills the test is intended to measure.

Research on the effects of accommodations for students with both of these types of disabilities has been conducted; however, this research has rarely examined the effects for these two groups of students on the same test. Studies have found that some items tend to be particularly difficult for students with sensory/physical disabilities even with accommodations (Bennett, Rock, & Novatkowski, 1989). However, in general, measurement tends to be comparable for students with these disabilities who receive accommodations (Bennett, Rock, & Jirele, 1987; Bennett, Rock, & Kaplan, 1987). Much more research has recently been conducted on the effects of accommodations for students with cognitive disabilities, with some studies supporting the provision of accommodations to these students (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000; Tindal et al., 1998; Weston, 1999), and others questioning whether there are substantially greater effects for this group of students than students without disabilities (Helwig, Rozek-Tedesco, & Tindal, 2002; Meloy et al., 2002).

Belief Three: Accommodations Fall on a Continuum in Terms of their Appropriateness (i.e., Some are Highly Appropriate, Some are Highly Inappropriate, and Some are Moderately Appropriate)

Elliott (1999) argues that validity is a matter of degree, rather than a discrete test characteristic. Tindal (1998) and Phillips (2002) indicate that accommodations could be placed on a continuum in terms of the degree to which they are expected to impact score comparability. An analysis of state accommodation policies shows that there are some accommodations that are very frequently allowed, some that are allowed in a moderate number of states, and some that are allowed in only a few states (Thurlow et al., 2002). This seems to indicate that there are some accommodations that are considered highly appropriate (e.g., individual or small group setting), some that are considered moderately appropriate (e.g., dictated response, extended time, read-aloud on a math test), and others that are considered highly inappropriate (e.g., read-aloud on a reading test, calculator for computation items on a math test). Although research has examined the appropriateness of these accommodations individually, rarely has their relative impact on test measurement been compared for different accommodations on the same test.

According to Fuchs et al. (2000), accommodations are intended to allow for the same attributes to be measured among accommodated students with disabilities and students without disabilities taking a standard administration of the test. It follows that if accommodations are serving their intended purpose, the measurement characteristics of accommodated test administrations for students with disabilities should be similar to those for non-accommodated students without disabilities. Differential item functioning (DIF) analysis is an application of item response theory (IRT) that involves comparing

item parameter estimates for a focal group and a reference group. The detection of DIF indicates that an item measures differently for one group of students than another group of students, and therefore must be less valid for one of the groups of students (Thissen, 2001). The specific research questions addressed in the current study were therefore:

1. When compared to non-accommodated students without disabilities, are there greater differences in item parameters for students with disabilities receiving the read-aloud accommodation on a reading test than a math test?

2. When compared to non-accommodated students without disabilities, are there greater differences in item parameters for accommodated students with cognitive disabilities than accommodated students with sensory or physical disabilities?

3. Do differences in item parameters for accommodation groups follow the anticipated patterns (i.e., large differences for those accommodations believed to be highly inappropriate, moderate differences for those accommodations believed to be somewhat appropriate, small differences for those accommodations believed to be highly appropriate)?

Methods

Datasets

In order to examine differences in IRT item parameters between accommodated and non-accommodated conditions, three datasets were obtained that included item-level performance on statewide assessments, as well as information on disability status and the accommodations individual students received. The datasets were as follows:

- 2001-2003 reading and math data from a statewide assessment program (State One; S1)

- 1998-2001 math and communication arts data from a statewide assessment program (State Two; S2)
- 2002 data from a statewide assessment program, math and reading sections (State Three; S3)

The State One and State Two datasets represented elementary student performance; the State Three dataset represented high school student performance. In State One and State Two, overall scores on the tests that were analyzed were used to determine accountability ratings for schools. Student performance on the test represented in the State One dataset was additionally used in making grade promotion decisions for individual students. The State Three test was used for instructional planning and student career planning. Only selected response items from each of these assessments were chosen for inclusion in the current analysis.

Procedures

Group selection. In order to investigate differences in IRT item parameters for various accommodation conditions, groups were created separately by dataset and content area to address the comparisons listed in Table 1. Analyses of differences in item parameters (i.e., DIF analyses) were only conducted for groups with $N > 150$. For groups that were intended to allow for the examination of specific accommodations, an effort was made to include only those cases receiving the specific accommodation in isolation; however, given the nature of some accommodations, in many cases it was considered important to include students who had received the target accommodation in addition to other accommodations. For instance, in order for the read-aloud accommodation to be most effective, an additional extended time accommodation is often necessary. Students

receiving this accommodation in addition to the read-aloud accommodation were therefore included in the "read-aloud" accommodation group. The results tables indicate which accommodations were included in addition to the target accommodation (if any) for each accommodation group. The number of students receiving only a calculator accommodation was too small to allow for analysis; however, the number of students receiving either the calculator accommodation in isolation or in addition to the read-aloud accommodation was large enough to conduct the DIF analysis, and so this group was analyzed. A group of 1,000 non-accommodated students without disabilities was randomly selected to serve as the reference group in the analyses, and a second group of 1,000 non-accommodated students without disabilities (DIF ref.) was randomly selected for comparison purposes. Non-accommodated students with disabilities were also included in the analysis for comparison purposes. Descriptive statistics, including the number of students, average performance, and standard deviation were calculated for each group.

DIF analyses. In this study, the three parameter logistic model (3-PL) was applied to examine DIF for the various accommodation groups. Using the 3-PL model, item-characteristic curves (ICCs) can be plotted for a focal group (i.e., accommodation group) and a reference group according to the following equation:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - bi)}}{1 + e^{Da_i(\theta - b_i)}} \quad (1)$$

where P represents the probability that a randomly selected examinee with ability θ will respond to a given item (i) correctly, c_i represents the lower asymptote parameter (typically the correction for items that can be answered correctly simply by guessing), D

is a constant (1.7 for this analysis), b_i represents the item difficulty, and a_i represents the discrimination value. The extent to which the estimated item parameters (i.e., item difficulty, item discrimination, and lower asymptote parameter) differ, and therefore the extent to which the ICCs differ across the focal and reference groups, represents the degree of measurement incomparability of the item across the groups. When item parameters are significantly different, DIF is present, and it is possible that the item is not measuring similarly across groups.

In this study, IRTLRDIF (Thissen, 2001) was used to estimate item parameters and detect DIF for each accommodation group based on the available data, with the randomly selected group of 1,000 non-accommodated students without disabilities serving as the reference group for all comparisons. Additional focal groups were used to determine the extent of DIF that might be identified simply due to sampling error and disability status.

When DIF is identified, it can be large or small, and it is important to examine the magnitude of DIF in order to know whether the difference is important. In this study, the magnitude of DIF present for each DIF item identified within a comparison was determined by calculating a weighted average of the vertical distance between focal (i.e., accommodation group) and reference group ICCs. This was done in a manner similar to that described by Wainer (1993) for the standardized index of differential item functioning. Wainer's (1993) equation is as follows:

$$T(3) = \int_{-\infty}^{\infty} [P_F(\theta) - P_R(\theta)]^2 \, dG_F(\theta) \qquad (2)$$

where $P_F(\theta)$ is the probability of an examinee answering correctly if in the focal (accommodation) group, $P_R(\theta)$ is the probability of an examinee answering correctly if in the reference group, and $dG_F(\theta)$ represents the density of the proficiency distribution of the focal group.

In order to approximate Wainer's (1993) index, vertical distances between accommodation and reference group ICCs were calculated for 100 equidistant points between -4 and +4 on the latent trait distribution for each DIF item, and weighted according to the density of the proficiency distribution for the focal group at the given point. These weighted distances were summed across the 100 points to provide an overall indication of the magnitude of DIF for the given DIF item. Weighting was necessary to counter issues related to poor parameter estimation at the extremes of the latent trait distribution.

The magnitude of DIF present in each DIF item was then evaluated according to the STD P-DIF index described by Dorans and Holland (1993), which considers values greater than or equal to .05 mildly problematic and worthy of inspection (moderate DIF), and values greater than or equal to .10 particularly worthy of investigation (large DIF). The number of DIF items with a magnitude of DIF within these ranges was determined for each accommodation group. In order to compare results for various accommodation groups across content areas, the proportion of items with moderate ($.05 \leq DIF < .10$) and large ($DIF \geq .10$) DIF was determined. Greater support for the appropriateness of an accommodation was considered present when a lower proportion of items showed moderate or large DIF, and when these results were consistent across datasets.

Results

Tables 2 to 7 provide descriptive information for the groups examined. The information is presented separately by dataset and content area. The proportion of students with disabilities represented in the test administrations ranged from nine percent (State Three) to sixteen percent (State One). Between 63 percent and 77 percent of students with disabilities received accommodations across datasets and content areas, with State One typically having the highest proportion of students with disabilities receiving accommodations, and State Two typically having the lowest proportion of students with disabilities receiving accommodations.

Tables 8 to 13 provide information on the number of items and proportion of items displaying DIF, as well as the number of items and proportion of items displaying moderate and large DIF across the examined accommodation groups. This information is presented in a figure for comparison purposes on the pages that follow the tables (Figure 1). In every dataset section, the DIF reference group (DIF ref.) had the least amount of DIF, and in all but one of the dataset sections, the group of non-accommodated students with disabilities had the second least amount of DIF. This seems to suggest that accommodated test administrations were associated with greater measurement incomparability than that which was present due to random sampling error or that which was due to disability status.[*]

Read-aloud Accommodation on Math Test vs. Read-aloud Accommodation on Reading Test

[*]The non-accommodated students with disabilities group differed in composition, however, from those who received accommodations. It is impossible to know the extent of DIF that would have been present among those students with disabilities who received accommodations had they not received accommodations.

In two of the three datasets, a similar proportion of items was identified as displaying moderate to large DIF for this accommodation group across math and reading sections. For the remaining dataset, more moderate and large DIF was present for the read-aloud accommodation group on the reading test than on the math test. However, in all three datasets, a greater proportion of items were identified as displaying large DIF for the read-aloud accommodation group on the reading sections than on the math sections.

Accommodations for Sensory/Physical vs. Cognitive Disabilities

In State One, a greater proportion of items displayed moderate to large DIF for accommodated students with sensory/physical disabilities than accommodated students with cognitive disabilities; this was more pronounced in the math section than in the reading section. However, for both sections, slightly more items displayed large DIF among accommodated students with cognitive disabilities than accommodated students with sensory/physical disabilities.

On the State Two math section, a similar proportion of items displayed moderate to large DIF for both disability groups; however, more items were identified as displaying large DIF for the accommodated students with cognitive disabilities than the accommodated students with sensory/physical disabilities. On the State Two communication arts section, more moderate and large DIF was present for accommodated students with cognitive disabilities than accommodated students with sensory/physical disabilities.

In summary, in only one of the four test sections available for this comparison was measurement comparability substantially poorer for accommodated students with cognitive disabilities than accommodated students with sensory/physical disabilities,

when examining the proportion of items with moderate to large DIF. However, when looking only at the proportion of items with large DIF across these accommodation groups, results more frequently supported the commonly-held belief that accommodations for students with cognitive disabilities corresponded to poorer measurement.

Continuum of Appropriateness: How Well Did DIF Results Match Anticipated Patterns?

In Figure 1, the accommodation groups are listed beginning with those groups that were not expected to have substantial DIF to those that were expected to have substantial DIF, based on commonly-held beliefs about the extent to which the accommodations influence test measurement. They are ordered according to the number of states that allow the given accommodation without limitations, as provided in Thurlow and Bolt (2001). The figure displays the proportion of items with moderate and large DIF for each accommodation group by dataset section.

For State One, there were not enough groups with a large enough size to compare results for several accommodation groups. For the State Two math section, several of the accommodation groups followed the expected pattern. The extended time and dictated response accommodation groups had a moderate proportion of items displaying moderate to large DIF, as was anticipated. The setting accommodation group and read-aloud accommodation groups had a somewhat greater proportion of items displaying moderate DIF than was anticipated. Although the calculator accommodation did not have the greatest proportion of items with moderate to large DIF, it did have the greatest proportion of items with large DIF. In the State Two reading section, there was a greater proportion of items identified with moderate to large DIF than in the math section. The

results followed the expected pattern, with the setting accommodation group having the least amount of DIF, and the read-aloud accommodation group having the greatest amount of DIF. However, the read-aloud accommodation group had only a slightly greater proportion of items with large DIF than the dictated response accommodation. If the read-aloud accommodation truly was much more inappropriate than the dictated response accommodation for this test section, one might expect to identify substantially more DIF corresponding to the read-aloud accommodation on the communication arts test.

In the State Three dataset, the read-aloud accommodation group displayed a greater proportion of items with moderate to large DIF than the extended time group across both test sections, but particularly on the reading section, as was expected. The reading section of the State Three dataset was the only section in which a greater proportion of items were identified as displaying moderate to large DIF for the non-accommodated students with disabilities than for the other accommodation groups.

Discussion

The purpose of this study was to examine the extent of support for several commonly-held beliefs about testing accommodations for students with disabilities using existing data from statewide assessments. A discussion of the support identified for each belief is provided below.

Read-aloud Accommodation for Reading vs. Math Test

Only one of the three datasets fully supported the belief that the read-aloud accommodation results in poorer measurement on reading than math tests for students with disabilities. However, across all three datasets a greater proportion of items with

large DIF was identified for this accommodation group on reading than math tests. It may be the case that the read-aloud accommodation has a substantial impact on the measurement characteristics of a few items, but allows for appropriate measurement on other items.

Accommodations for Sensory/Physical vs. Cognitive Disabilities

Based on the current analysis, accommodations for students with cognitive disabilities were not associated with substantially greater measurement incomparability than accommodations for students with sensory/physical disabilities, when comparing IRT item parameters for each group to the reference group of non-accommodated students without disabilities. In only one of the four dataset sections available for the analysis did results fully support the hypothesis that accommodating students with cognitive disabilities is associated with poorer measurement than accommodating students with sensory/physical disabilities. It is important to note, however, that there was a greater proportion of items with large DIF identified across all four dataset sections for accommodated students with cognitive disabilities than for accommodated students with sensory/physical disabilities. Therefore, although the belief was not fully supported, it may be the case that accommodations for students with cognitive disabilities result in poorer measurement on certain items. At the same time, it may be the case that the results identified in this disability comparison are related to the influence of the read-aloud accommodation on reading tests, given that a greater proportion of items with large DIF were identified for the accommodated cognitive disability group on reading sections than math sections.

Continuum of Accommodation Appropriateness

As anticipated, accommodations were associated with varying levels of measurement comparability in the datasets examined. This provides support for the belief that accommodations influence test measurement to varying degrees, and should not necessarily be considered either "appropriate" or "inappropriate." For the most part, the results followed the expected patterns for various accommodations. For instance, the setting accommodation typically had the least amount of moderate to large DIF, the extended time accommodation had a medium amount of moderate to large DIF, and the read-aloud accommodation for a reading test had the largest amount of moderate to large DIF. Although the calculator accommodation group had the greatest proportion of items with large DIF within the State Two math section, it did not have the most items with moderate to large DIF. In addition, the read-aloud accommodation group on math tests tended to have more items displaying moderate to large DIF than one might anticipate. It appears that the read-aloud accommodation is associated with poorer measurement comparability than might be expected.

Differences Between Datasets

Despite the fact that some patterns tended to be consistent across datasets, other patterns were different across datasets. This may be due to different skills and knowledge being tested across the various assessments, which in turn may be related to different effects of the accommodations on these tests. It will be important to clearly identify the particular skills and knowledge to be tested to determine whether certain accommodations allow for appropriate measurement of the intended skills and knowledge among students.

Across all but one dataset section (of six examined), accommodation groups were associated with more items displaying DIF than reference groups of non-accommodated students with and without disabilities. This seems to indicate that accommodations are not effective in fully removing construct-irrelevant variance associated with disabilities. Whether or not the accommodations improved measurement for students with disabilities remains questionable, given that information on the measurement comparability of non-accommodated testing for those students who did receive accommodations is not available.

Limitations of the Study

Although this study allowed for an examination of accommodated test administrations in real testing situations with large groups of students, it was not an experimental study, and there are subsequently limits to the inferences that can be made. As mentioned earlier, it is not possible to know whether the accommodations improved measurement for students who received them because these students were never tested without the accommodations. In addition, problems associated with inappropriate accommodation administration and coding of the data may have impacted the results obtained. It may be the case that test proctors who write down student responses and read-aloud test items have an inappropriate influence on student responses that could be minimized with additional training. It also may be the case that accommodations were not targeted to the right students in this study; previous research has indicated that decisions about which accommodations to provide to individual students are often flawed (Fuchs et al., 2000; Helwig & Tindal, 2003). Experts have suggested that testing accommodations decisions should be based in part on the types of instructional accommodations students

receive (Quenemoen, Thompson, Thurlow, & Lehr, 2001); it is not clear whether students in this study had been receiving the accommodations during instruction.

Direction for Future Research

The current study raises questions about whether accommodations are successful in allowing for comparable measurement of skills among students with disabilities on statewide assessments. Research should continue to investigate accommodation decision-making and administration practices. At the same time, the current research seems to indicate that even with accommodations, appropriate measurement of skills and knowledge among students with disabilities can be difficult to attain on currently available large-scale assessments.

A promising direction for future research and practice in the area of inclusive assessment is provided by the concept of universal design, as described in Thompson, Johnstone, and Thurlow (2002). Universal design for assessment involves the consideration of the needs of diverse students during the initial stages of test development. When applying principles of universal design to assessment, appropriate accommodations may be built-in to the test from the very beginning, and might be made available to all students, making questions about the comparability of accommodated test administrations for students with disabilities irrelevant. Applying the concept of universal design during stages of test development may lead to better measurement of the intended skills and knowledge for all students.

References

Bennett, R. E., Rock, D. A., & Jirele, T. (1987). GRE score level, test completion, and reliability for visually impaired, physically handicapped, and nonhandicapped groups. *The Journal of Special Education, 21* (3), 9-21.

Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1987). SAT differential item performance for nine handicapped groups. *Journal of Educational Measurement, 24*(1), 41-55.

Bennett, R. E., Rock, D. A., & Novatkoski, I. (1989). Differential item functioning on the SAT-M Braille edition. *Journal of Educational Measurement, 26*(1), 67-79.

Bielinski, J., Thurlow, M., Ysseldyke, J. E., Freidebach, J., & Freidebach, M. (2001). *Read-aloud accommodation: Effects on multiple-choice reading and math items* (Technical Report No. 31). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Bolt, S. E. & Bielinski, J. (2002, April). The effects of the read-aloud accommodation on math test items. *Paper presented at the annual conference of the National Council on Measurement in Education (NCME)*, New Orleans, LA.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning,* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Elliott, S. N. (1999, June). Valid testing accommodations: Fundamental assumptions and methods for collecting validity evidence. *Paper presented at the annual large-scale assessment conference,* Snowbird, UT.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. L., & Karns, K. M. (2000). Supplementing teacher judgments of mathematics test accommodations with objective data sources. *School Psychology Review, 29*(1), 65-85.

Helwig, R., Rozek-Tedesco, M. A., & Tindal, G. (2002). An oral versus a standard administration of a large-scale mathematics test. *The Journal of Special Education, 36*(1), 39-47.

Helwig, R. & Tindal, G. (2003). An experimental analysis of accommodation decisions on large-scale mathematics tests. *Exceptional Children, 69*(2), 211-225.

Heubert, J. P., & Hauser, R. M. (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy of Sciences, National Research Council.

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky* (CSE Technical Report No. 431). Los Angeles, CA: University of California, Los Angeles, Center for Research on Evaluation, Standards, and Student Testing.

Kosciolek, S. & Ysseldyke, J. E. (2000). *Effects of a reading accommodation on the validity of a reading test* (Technical Report 28). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Lewis, D., Green, D. R., & Miller, L. (1999, June). Using differential item functioning analyses to assess the validity of testing accommodated students with disabilities. *Paper presented at the national conference on large-scale assessment,* Snowbird, UT.

McKevitt, B. C. & Elliott, S. N. (2003). Effects and perceived consequences of using read-aloud and teacher-recommended testing accommodations on a reading achievement test. *School Psychology Review, 32*(4), 583-600.

Meloy, L. L., Deville, C., & Frisbie, D. A. (2002). The effect of a read-aloud accommodation on test scores of students with and without a learning disability in reading. *Remedial and Special Education, 23* (4), 248-255.

Olson, J. F. & Goldstein, A. A. (1996). Increasing the inclusion of students with disabilities and limited English proficient students in NAEP." *Focus on NAEP, 2*(1). Washington, DC: National Center for Education Statistics.

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7*(2), 93-120.

Phillips, S. E. (2002). Legal issues affecting special populations in large-scale testing programs. In G. Tindal & T. Haladyna (Eds.), *Large scale assessment programs for all students.* (pp. 109-148). Mahwah, NJ: Lawrence Erlbaum Associates.

Pomplun, M., & Omar, M. H. (2000). Score comparability of a state mathematics assessment across students with and without reading accommodations. *Journal of Applied Psychology, 85*(1), 21-29.

Quenemoen, R. F., Thompson, S. J., Thurlow, M. L., & Lehr, C. A. (2001). A self-study guide to implementation of inclusive assessment and accountability systems: A best practice approach. Minneapolis, MN: National Center on Educational Outcomes. Retrieved March 9, 2004 from http://www.education.umn.edu/NCEO/OnlinePubs/workbook.pdf

Schulte, A. G., Elliott, S. N., & Kratochwill, T. R. (2001). Experimental analysis of

the effects of testing accommodations on students' standardized achievement test scores. *School Psychology Review, 30*(4), 527-547.

Thissen, D. (2001). *IRTLRDIF v. 2.0b: Software for the computation of the statistics involved in item-response theory likelihood-ratio tests for differential item functioning.* Chapel Hill, NC: University of North Carolina, L.L. Thurstone Psychometric Laboratory.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved November 2, 2002, http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Thurlow, M., Lazarus, S., Thompson, S., & Robey, J. (2002). *2001 State policies on assessment participation and accommodations.* (Synthesis Report No. 46). Minneapolis, MN: National Center on Educational Outcomes.

Tindal, G. (1998). *Models for understanding task comparability in accommodated testing.* Retrieved December 9, 1999, from http://www.coled.umn.edu/nceo/Accommodations/SEAEXEC.html

Tindal, G. & Fuchs, L. (1999). A summary of research on test changes: An empirical basis for defining accommodations. Lexington, KY: Mid-south Regional Resource Center.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. *Exceptional Children, 64*(4), 439-450.

Tippets, E., & Michaels, H. (1997). Factor structure invariance of accommodated and non-accommodated performance assessments. *Paper presented at the annual meeting of the National Council on Measurement in Education*. Chicago.

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P.W. Holland & H. Wainer (Eds.) *Differential item functioning.* (pp. 123-135). Hillsdale, NJ: Lawrence Erlbaum Associates.

Weston, T. (1999). *The validity of oral presentation in testing.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Table 1

Description of Comparisons Made Within Datasets

Research Question	Accommodation group	S1	S2	S3
1, 3	Students with Disabilities (SWD) Receiving Read-Aloud in Math	X	X	X
1, 3	SWD Receiving Read-Aloud in Reading	X	X	X
2	Accommodated Students with Hearing, Vision, or Orthopedic Impairments	X	X	
2	Accommodated Students with Learning or Behavioral Disability or Mental Retardation	X	X	
3	SWD Receiving Setting Accommodations		X	
3	SWD Receiving Calculator and Read-Aloud on a Math Test (including computation items)		X	
3	SWD Receiving Extended Time		X	X
3	SWD Receiving Dictated Response		X	

Table 2

Test Performance for Students Receiving Accommodation Packages Investigated for State One - Math Section

Group	N	μ	σ
Non-accommodated Students Without Disabilities (NSWOD)	21455	21.8	5.0
Non-accommodated Students With Disabilities (NSWD)	1034	19.0	6.0
Students With Disabilities Receiving Extended Time (ET)*	31	18.6	5.8
Students With Disabilities Receiving Read-Aloud (RA)**	181	14.9	5.6
Accommodated Students With Cognitive Disabilities (Cog)	2683	14.4	5.7
Accommodated Students With Sensory and Physical Disabilities (Sens/Phys)	321	15.7	6.4
DIF Reference Group	1000	21.8	4.9
DIF Reference Group for Comparison (DIF Ref.)	1000	21.6	5.1
All Students	26246	20.7	5.7

*This group included students receiving additional setting accommodations
**This group included students receiving additional scheduling and setting accommodations

Table 3

Test Performance for Students Receiving Accommodation Packages Investigated for State One - Reading Section

Group	N	μ	σ
NSWOD	21479	20.1	5.4
NSWD	978	17.5	6.2
RA*	236	13.9	5.5
Cog	2744	13.6	5.6
Sens/Phys	321	14.7	5.9
DIF Reference Group	1000	20.0	5.2
DIF Ref.	1000	20.0	5.4
All Students	26242	19.1	5.9

*This group included students receiving additional setting accommodations

Table 4

Test Performance for Students Receiving Accommodation Packages Investigated for State Two - Math Section

Group	N	μ	σ
NSWOD	233273	25.2	5.2
NSWD	12932	21.2	6.8
Dictated Response (DR)*	328	22.5	6.2
ET**	797	20.2	6.6
RA*	14444	19.5	6.5
Calculator + Read Aloud (C + RA)*	695	18.4	6.6
Read Aloud + Dictated Response (RA + DR)*	4184	19.1	6.9
Setting	972	20.0	6.4
Cog	22253	19.4	6.7
Sens/Phys	446	19.5	6.9
DIF Reference Group	1000	25.1	5.2
DIF Ref.	1000	25.2	5.4
All Students	275119	24.4	5.8

*These groups included students receiving additional scheduling and setting accommodations
**This group included students receiving additional setting accommodations

Table 5

Test Performance for Students Receiving Accommodation Packages Investigated for State Two - Communication Arts Section

Group	N	μ	σ
NSWOD	218158	31.8	6.4
NSWD	11355	28.6	6.9
DR*	360	27.5	6.3
ET**	824	26.0	6.7
RA*	9357	26.2	6.0
RA + DR*	4478	26.2	6.1
Setting	852	25.6	6.5
Sens/Phys	339	26.7	7.1
Cog	15765	26.2	6.1
DIF Reference Group	1000	31.4	6.7
DIF Ref.	1000	31.9	6.5
All Students	251082	31.2	6.6

*These groups included students receiving additional scheduling and setting accommodations
**This group included students receiving additional setting accommodations

Table 6

Test Performance for Students Receiving Accommodation Packages Investigated for State Three – Math Section

Group	N	μ	σ
NSWOD	17702	18.4	6.8
NSWD	547	11.0	4.7
ET	442	10.1	4.2
RA*	578	9.3	3.5
DIF Reference Group	1000	18.6	6.9
DIF Ref.	1000	18.3	6.9
All Students	19882	17.5	7.1

*This group included students receiving additional scheduling and setting accommodations

Table 7

Test Performance for Students Receiving Accommodation Packages Investigated for State Three – Reading Section

Group	N	μ	σ
NSWOD	17692	12.1	4.2
NSWD	547	7.8	4.7
ET	443	7.4	2.8
RA*	566	7.6	2.9
DIF Reference Group	1000	12.0	4.3
DIF Ref.	1000	12.2	4.3
All Students	19872	11.6	4.3

*This group included students receiving additional scheduling and setting accommodations

Table 8

Differential Item Functioning for Accommodation Groups in State One – Math Section

Group	N	No./Percent of Items with Sig. DIF	No./Percent of Items with Sig. DIF and $.050 \leq DIF < .100$	No./Percent of Items with Sig. DIF and $DIF \geq .100$
NSWD	1034	5 / 17%	5 / 17%	0
RA*	181	23 / 77%	17 / 57%	6 / 20%
Cog	2683	15 / 50%	13 / 43%	2 / 7%
Sens/Phys	321	22 / 73%	21 / 70%	1 / 3%
DIF Ref.	1000	3 / 10%	3 / 10%	0

*This group included students receiving additional scheduling and setting accommodations

Table 9

Differential Item Functioning for Accommodation Groups in State One – Reading Section

Group	N	No./Percent of Items with Sig. DIF	No./Percent of Items with Sig. DIF and $.050 \leq DIF < .100$	No./Percent of Items with Sig. DIF and $DIF \geq .100$
NSWD	21479	9 / 30%	9 / 30%	0
RA*	236	22 / 73%	13 / 43%	9 / 30%
Cog	2744	17 / 57%	7 / 23%	10 / 33%
Sens/Phys	321	19 / 63%	15 / 50%	4 / 13%
DIF Ref.	1000	3 / 10%	3 / 10%	0

*This group included students receiving additional scheduling and setting accommodations

Table 10

Differential Item Functioning for Accommodation Groups in State Two – Math Section

Group	N	No./Percent of Items with Sig. DIF	No./Percent of Items with Sig. DIF and $.050 \leq DIF < .100$	No./Percent of Items with Sig. DIF and $DIF \geq .100$
NSWD	12932	4 / 13%	4 / 13%	0
DR*	328	8 / 25%	8 / 25%	0
ET**	797	5 / 16%	5 / 16%	0
RA*	14444	13 / 41%	11 / 34%	2 / 6%
C + RA*	695	10 / 31%	7 / 22%	3 / 9%
RA + DR*	4184	16 / 50%	14 / 44%	2 / 6%
Setting	972	8 / 25%	8 / 25%	0
Cog	22253	14 / 44%	12 / 38%	2 / 6%
Sens/Phys	446	14 / 44%	13 / 41%	1 / 3%
DIF Ref.	1000	1 / 3%	1 / 3%	0

*These groups included students receiving additional scheduling and setting accommodations
**This group included students receiving additional setting accommodations

Table 11

Differential Item Functioning for Accommodation Groups in State Two – Communication Arts Section

Group	N	No./Percent of Items with Sig. DIF	No./Percent of Items with Sig. DIF and $.050 \leq DIF < .100$	No./Percent of Items with Sig. DIF and $DIF \geq .100$
NSWD	11355	10 / 24%	8 / 20%	2 / 5%
DR*	360	17 / 41%	11 / 27%	6 / 15%
ET**	824	18 / 44%	15 / 37%	3 / 7%
RA*	9357	24 / 59%	14 / 34%	10 / 24%
RA + DR*	4478	24 / 59%	14 / 34%	10 / 24%
Setting	852	11 / 27%	7 / 17%	4 / 10%
Cog	15765	23 / 56%	12 / 29%	11 / 27%
Sens/Phys	339	14 / 34%	8 / 20%	6 / 15%
DIF Ref.	1000	0	0	0

*These groups included students receiving additional scheduling and setting accommodations
**This group included students receiving additional setting accommodations

Table 12

Differential Item Functioning for Accommodation Groups in State Three – Math Section

Group	N	No./Percent of Items with Sig. DIF	No./Percent of Items with Sig. DIF and $.050 \leq DIF < .100$	No./Percent of Items with Sig. DIF and $DIF \geq .100$
NSWD	547	12 / 31%	11 / 28%	1 / 3%
ET	442	12 / 31%	10 / 26%	2 / 5%
RA*	578	15 / 38%	14 / 36%	1 / 3%
DIF Ref.	1000	5 / 13%	5 / 13%	0

*This group included students receiving additional scheduling and setting accommodations

Table 13

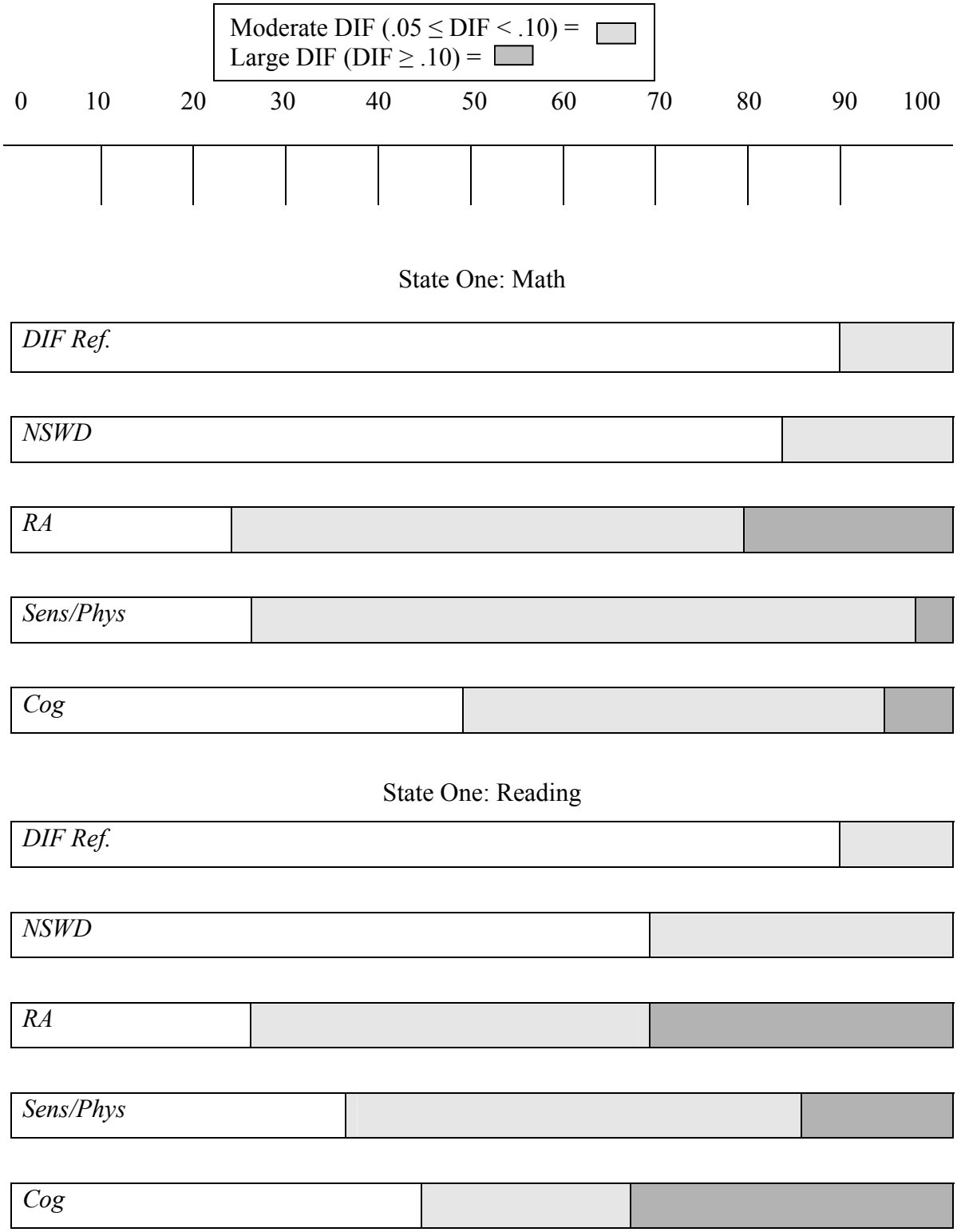
Differential Item Functioning for Accommodation Groups in State Three – Reading Section

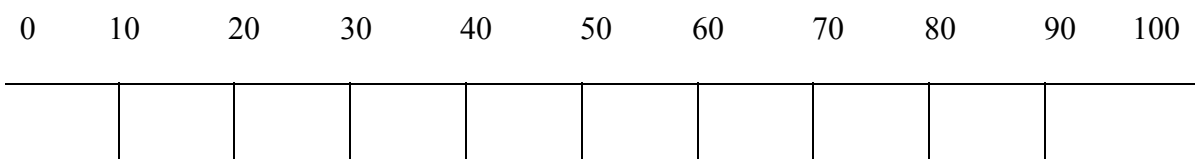
Group	N	No./Percent of Items with Sig. DIF	No./Percent of Items with Sig. DIF and $.050 \leq DIF < .100$	No./Percent of Items with Sig. DIF and $DIF \geq .100$
NSWD	547	3 / 12%	2 / 8%	5 / 20%
ET	443	6 / 24%	5 / 20%	1/ 4%
RA*	566	10 / 40%	8 / 32%	2 / 8%
DIF Ref.	1000	2 / 8%	2 / 8%	0

*This group included students receiving additional scheduling and setting accommodations

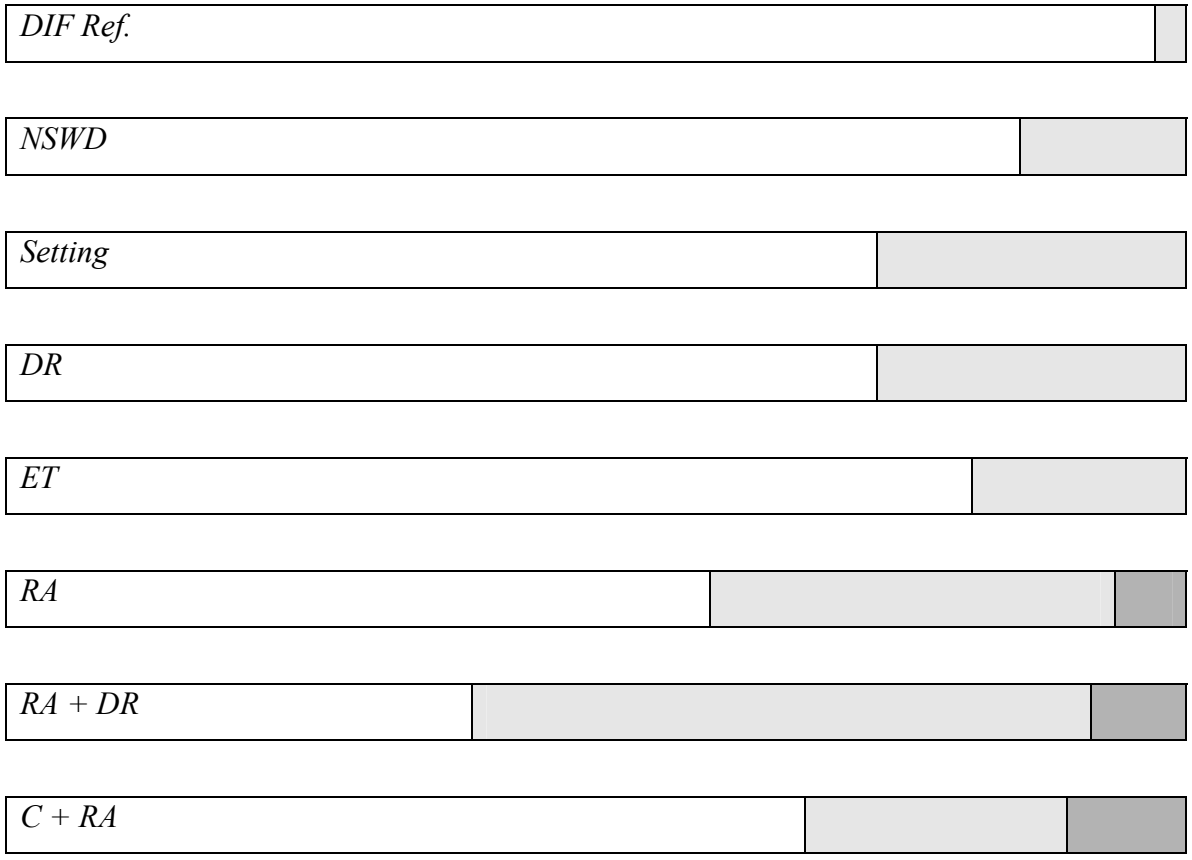
Figure 1

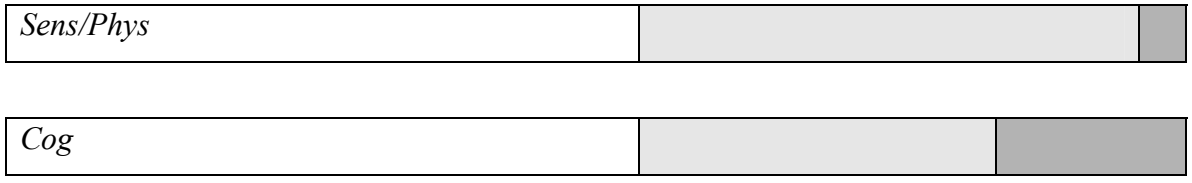
Percent of Items Displaying Moderate and Large DIF for Various Accommodation Groups

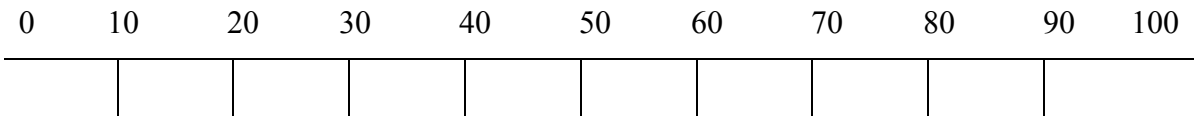




State Two: Math







State Two: Communication Arts

DIF Ref.

NSWD

Setting

DR

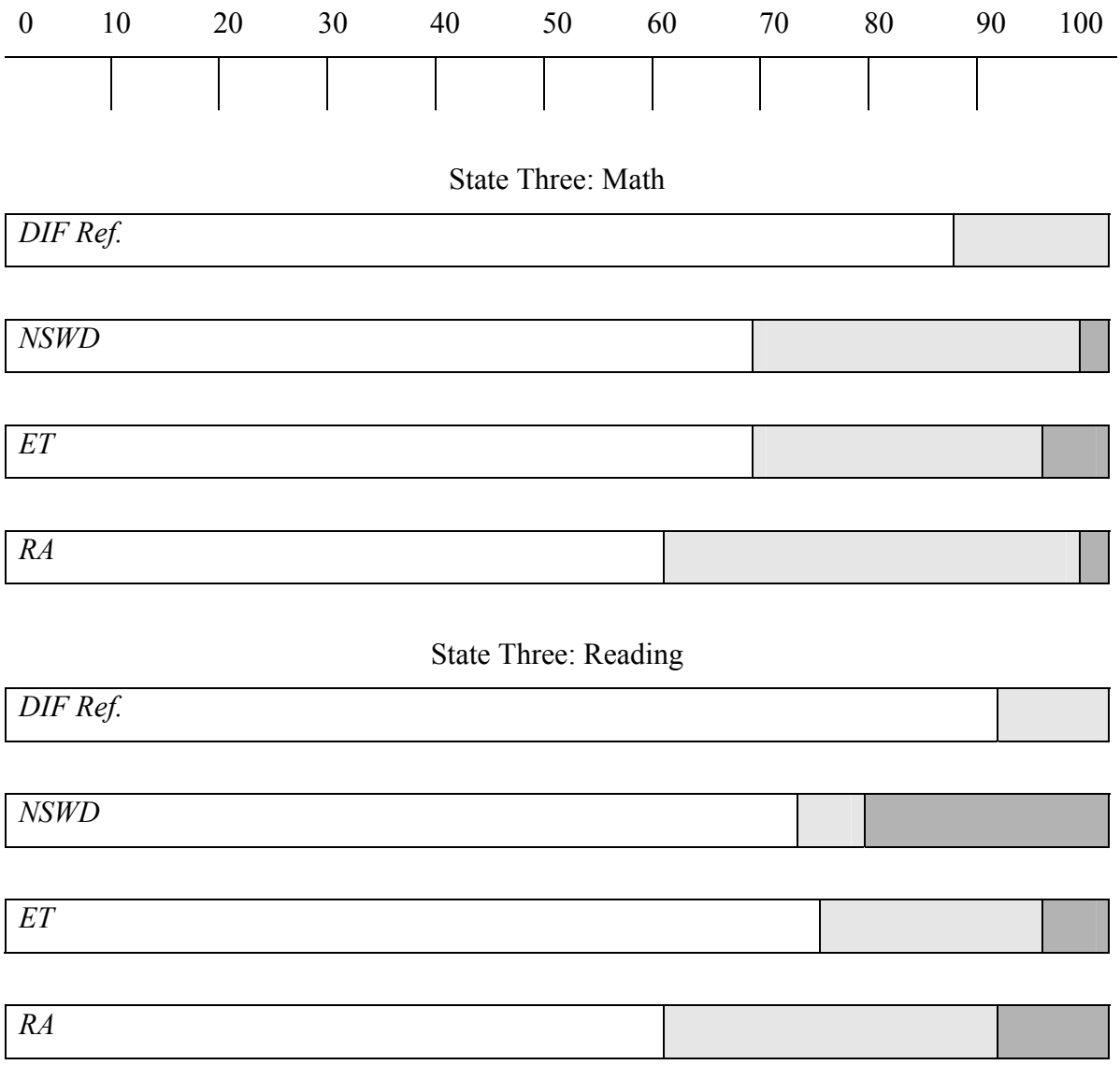
ET

RA

RA + DR

Sens/Phys

Cog


0
10
20
30
40
50
60
70
80
90
100
State Three: Math
DIF Ref.
NSWD
ET
RA
State Three: Reading
DIF Ref.
NSWD
ET
RA