NAGB Conference on Increasing the Participation of SD and LEP Students in NAEP

Commissioned Paper Synopsis

The attached paper is one of a set of research-oriented papers commissioned by NAGB to serve as background information for the conference attendees. The authors bear sole responsibility for the factual accuracy of the information and for any opinions or conclusions expressed in the paper.

Research on Mathematics Test Accommodations Relevant to NAEP Testing

Gerald Tindal, Ph.D.

Leanne R. Ketterlin-Geller, Ph.D.

University of Oregon

January 5, 2004

- Summary of the primary issues relevant to including students with disabilities in large-scale testing: Accommodations include changes that are allowable by state policies and which do not change the construct being measured by the test. Modifications are changes that substantially alter the construct being measured. Students needing modifications are either excluded from the assessment or if allowed to participate, the outcomes are not aggregated with scores generated under standard administration conditions. This paper focuses on the use of accommodations by students with disabilities in large-scale assessment programs.

- A summary of accommodations decisions by teachers indicates teachers lack the knowledge and skills to formulate sound decisions about appropriate accommodation placements. With fluctuating decision-making practices at the local level and inconsistent research findings, most states have resorted to a logic-based approach when deciding whether or not an accommodation is allowable on large-scale tests. Although seemingly reasonable, this decision-making approach leads to incongruent practices across domains and further differentiates states participation policies.

- The most common accommodation on NAEP is small group setting, followed by extended time and read aloud.

- Research on participation in and achievement on NAEP assessments under accommodated and nonaccommodated conditions indicated that more students with disabilities participated when accommodations were permitted. Another finding was that on the fourth- grade assessment a significantly larger number of students who were permitted accommodations did not reach basic proficiency. However this last finding is confounded by the lack of comparability between the two groups (students that received accommodations and students that did not receive accommodations).

- Research on testing accommodations for mathematics tests has specifically focused on (a) using calculators, (b) reading mathematics problems to students, and (c) having the test timed versus having extended time. The read-aloud accommodation is more likely to benefit younger students or those with lower reading skills. Calculator use appears to act as a facilitator on some items and a detractor on others. Three studies on extra time with relevance for NAEP indicate no changes when students with disabilities take the test under timed versus untimed conditions.

- Three strategies for practically addressing accommodation decisions are proposed. These recommendations include (a) applying the principles of Universal Design to the development of assessments to reduce bias based on gender, language, culture, and disability; (b) organizing tests into sections based on the skills assessed so that accommodation decisions can be made at the skill level; and (c) using computer adaptive testing to incorporate the item's target skill relative (e.g., addition) to an access skill (e.g., reading a math problem).

Running Head: MATH TEST ACCOMMODATIONS

Research on Mathematics Test Accommodations Relevant to NAEP Testing

Gerald Tindal, Ph.D.

Leanne R. Ketterlin-Geller, Ph.D.

University of Oregon

January 5, 2004

Address correspondence to: Gerald Tindal, Behavioral Research and Teaching, College of Education, 232 Education, University of Oregon, Eugene, Oregon, 97403-5262. geraldt@darkwing.uoregon.edu

The paper is divided into three sections. In the first section, we summarize the primary issues relevant to including students with disabilities in large-scale testing. In the second section, we summarize the research on such participation in the National Assessment of Educational Progress (NAEP). In the third section, we synthesize the research on mathematics test accommodations conducted over the past 20 years. Finally, we abstract eight major reviews on accommodation research and describe 37 primary research studies by the type of test (demands), the use of accommodations, the student populations, and the general outcomes (of both accommodations and participation).

Section I: Including Students with Disabilities in Large Scale Testing

Since the early 1990s, the National Center on Educational Outcomes (NCEO) has helped develop a systematic database on the participation of students with disabilities in large-scale testing programs. In part, the philosophy then (and now) is that "students who are not measured in educational accountability systems tend to be ignored when educational reforms are enacted" (Elliott, 2001, p. 4). Not only are programs likely affected, but also interpretations of student abilities are misplaced. Particularly with high stakes testing programs where all students count, the first order issue is participation the outcomes are directly influenced by the characteristics of who is taking the test. To the degree that populations with certain characteristics are excluded, the outcomes are constrained and not generalizable to the entire population. Therefore, it is critical to examine the issues that preclude appropriate participation and identify mechanisms that can increase the meaningful participation of students with disabilities in large-scale assessment systems.

Decisions about participation of students with disabilities in large-scale assessment systems are directly a function of state policies on accommodations, modifications, and alternate assessments. For purposes of this paper, accommodations include changes that are allowable by state policies and which do not change the construct being measured by the test. Modifications are changes that substantially alter the construct being measured. Students needing modifications are either excluded from the assessment or if allowed to participate, the outcomes are not aggregated with scores generated under standard administration conditions. Typically, accommodations and modifications are applied to students with high incidence disabilities while alternate assessments (entirely different measurement systems than the large-scale tests) are reserved for students with significant disabilities. The distinction between accommodations, modifications, and alternate assessments is critical when reporting participation rates due to the significantly different interpretations that result from the observed score. In this paper, we will focus on the use of accommodations by students with disabilities in large-scale assessment programs.

To document the level of participation and use of accommodations by students with disabilities in state assessment systems, NCEO conducted a national survey in 1991 with state department directors of special education focusing on several aspects of state-level data collection. This same survey was readministered after five years in 1995 with fairly impressive results. Over this time span, state departments began to collect specific types of participation and exit data (now required with federal legislation) and documented the emergence of systematic and written procedures for making decisions about participation and accommodations for students with

special needs. Though Individual Education Program (IEP) teams continued to assume primary responsibility for making these decisions, the number of states with written policies had nearly doubled.

In a similar manner, over the past 10 years, the amount of research on participation and accommodations for individuals with disabilities also has increased. Much of the nascent research in this area was conducted by Willingham, Ragosta, Bennett, Braun, Rock, and Powers, (1988) at Educational Testing Service (ETS). Despite technical rigor and analytic precision, Thurlow, Ysseldyke, and Silverstein (1998) stated that "this research, unfortunately is of only limited value because of its focus on assessments used to make admissions decisions (ACT, SAT, and GRE), rather than assessments used for accountability purposes" (p. 5). Furthermore, the populations they studied may not have been generalizable to those participating in state testing programs. Given ETS's emphasis on admission to college, the sample population represented a fairly selective group, unlike those who participate in compulsory K–12 public schools. Nevertheless, this initial research was invaluable in establishing a foothold on the development of an empirical basis for making participation and accommodations decisions; it also fostered better understanding of the affect of accommodations on tests and performance.

The early work of the NCEO expanded the research on accommodations from the perspective of score comparability (reflecting comparable meaning and interpretations of test performance) and task comparability (equivalence of cognitive demands made on subgroups of individuals) to examining the prevalence of accommodations in naturalistic settings (see Thurlow, Ysseldyke, & Silverstein 1995). Several literature syntheses have been conducted to track the development of the field and provide a cogent summary of the relevant issues (see Appendix A for a summary of recent syntheses). Through these documents, it has become more obvious that the field of research on accommodations has become diverse with few unified findings. "One thing that is clear from our review is that there are no unequivocal conclusions that can be drawn regarding the effects, in general, of accommodations on students' test performance. The literature is clear that accommodations and students are both heterogeneous" (Sireci, Li, & Scarpati, p. 48).

To bridge the research on technical to practical issues and to put order to the quickly changing policy landscape, a taxonomy of accommodations was developed with four major categories that were used to cluster the manner in which test changes were made: (a) setting, (b) presentation, (c) timing, (d) response; later this taxonomy was expanded to include (e) scheduling, and (f) other changes as summarized by Thurlow, Ysseldyke, and Silverstein (1998). This organizational structure was the first step in standardizing the available accommodations and allowed states to develop participation and accommodation policies for students with disabilities. This taxonomy has been useful in organizing research in the K–12 world of public schools and the use of large-scale testing for ALL students. It has, however, also fostered a line of research focused on more than simple outcomes from the use of accommodations to issues in the process of making decisions (by who and for whom).

Accommodation Decision Making at the Teacher Level

Decision making at the state and local level has been aided by this articulation of various accommodations and their effects on student performance. However, no widely accepted procedures are in place to help teachers make systematic decisions about participation and

accommodations. Although these decisions are linked to an individual's IEP team recommendations and/or prior experience with an accommodation, the practices of identifying and applying relevant and adequate data varies widely and result in an idiosyncratic decision-making process.

The instability of accommodation recommendations has been documented through studies of teachers' decision-making processes. For example, Fuchs, Fuchs, Eaton, Hamlett, and Karns (2000) administered mathematics tests to 200 students with learning disabilities (LD) and 200 students without LD. Each student took tests under both standard and accommodated conditions that included extended time, use of calculators, teacher reading text aloud, and scribing answers on problem solving tasks. Results showed that students whose teachers had not recommended accommodations failed to benefit from them at the same rate as those students whose teachers had recommended accommodations. And Hollenbeck, Tindal, and Almond (1998) found teachers (from both general and special education) to be nearly as incorrect as correct in their knowledge of allowable state-level accommodations. Thus, the findings by DeStephano, Shriner, and Loyd (2001) are not surprising: In an assessment accommodation survey completed by over 100 special education teachers in a large urban school district, they reported that "accommodation patterns did not appear to be based on access to the general education curriculum or the nature of instructional accommodations" (p. 18). They concluded that, in general, all students with disabilities received the same set of accommodations. Unfortunately, ease of administration appeared to be a significant factor in teachers' accommodation decisions (Gajria, 1994). It follows that teachers lack the knowledge and skills to formulate sound decisions about appropriate accommodation placements.

In the end, this state of affairs is untenable. Providing inappropriate placements can jeopardize student success by providing incompatible accommodations or withholding accommodations that are necessary. Additionally, inappropriate placements into accommodations may lead to misunderstanding of student ability levels. As proffered by Koretz (1997), accommodations may lead to outcomes that reflect greater than expected performance by students with disabilities. On the other hand, it is entirely possible to find lower performance for this group than students with disabilities who do not receive accommodations (in essence, teachers recommend accommodations to lower performing students and the two populations are not comparable). Therefore, teachers need systematic measures (such as curriculum-based measures) that are independent of state testing programs to determine whether or not an accommodation should be used or withheld.

With fluctuating decision-making practices at the local level and inconsistent research findings, most states have resorted to a logic-based approach when deciding whether or not an accommodation is allowable on large-scale tests. Although seemingly reasonable, this decision-making approach may lead to incongruent practices across domains and further differentiates states participation policies. For example, reading test directions aloud has been viewed as acceptable but reading a reading test has generally been viewed as unacceptable even though few studies have been conducted on this change. See Crawford and Tindal (in press) for a study on this change in testing. For some states, this change is a modification because the construct also is changed. These states argue that, for schools to be accountable to the public in stating that all students can read, reading the reading tests cannot be allowed. Yet in other states, it is allowed

because it is argued that the standards for advanced grades have been written in reference to high inference interpretations from reading rather than text-based decoding. Furthermore, these states argue that reading IS the disability and therefore precludes the assessment of making interpretations. Defense of such logic-based practice is difficult without resort to regulatory fiat.

Accommodation Decision-Making at the State Policy Level

States vary considerably in the participation rates of students with disabilities on large-scale assessment systems. These differences have significant consequences when comparing student performance across states and on independent measures of student achievement, such as NAEP. In part, participation is a function of the type of test being used and the decisions being made (e.g., high school exit). Such variations result in great differences in percentages of students making adequate yearly progress. For example, Minnesota has 82% of its students making AYP while Florida reports 13% (Robelen, September 3, 2003). These differences cannot be interpreted: They may reflect differences in the test, population, or educational effects individually or in combination. Therefore, the state-level decision-making process for including students with disabilities in large-scale assessments must be considered in order to equalize the participation rates across states. As states participate in NCLB and NAEP is used as a common measure to understand the variance among states, it becomes critical to understand the differences between state tests and policies on participation and those required by NAEP. These differences may influence the outcomes in either of three ways: (a) the results on NAEP are a function of state policies on participation (who takes the test), (b) state NAEP outcomes directly reflect level of achievement differences (which may still be attributable to test differences), or (c) both influences interact with each other.

Section II: Description of the NAEP Mathematics Testing Program

The National Assessment of Educational Progress assesses' students mathematical knowledge and skills in grades 4, 8, and 12 in five domains: (a) number sense, properties, and operations, (b) measurement, (c) geometry and spatial sense, (d) data analysis, statistics, and probability, and (e) algebra and functions. *Number sense, properties and operations* is defined as the conceptual understanding, manipulation, and application to real-world situations of all levels of numbers and numerical relationships. *Measurement* involves the application of numbers and measurement tools and units to analyze and communicate relationships between objects, attributes of objects, and measurement concepts. *Geometry and spatial sense* measures the ability to understand and manipulate geometric shapes to combine, transform, and reconstruct proportional figures. *Data analysis, statistics, and probability* integrates data collection and organization skills with the ability to analyze and interpret information using statistical procedures. An understanding of the principles of probability is required to solve problems and make decisions. *Algebra and functions* is defined as the ability to solve mathematical and real-world problems as open sentences and equations using increasingly complex algebraic concepts and functions and communicate the findings using algebraic notation. Students are required to demonstrate conceptual understanding, procedural knowledge, and problem solving skills in each strand using reasoning and communication skills to link information across domains.

Inclusion Policies for Students with Disabilities on the NAEP

The goal of NAEP is to assess all students within designated samples of students. Only recently (beginning in 2002) have accommodations been allowed on NAEP, thereby permitting assessment results from a representative sample of the population. School staff use the IEP and NAEP guidelines to determine whether a student can meaningfully participate in the assessment. Students are excluded from the NAEP sample if the IEP does not allow the student to participate in such assessments, the student's cognitive functioning is severely impaired as to prohibit meaningful participation, or the IEP requires accommodations not allowed by NAEP.

In mathematics, NAEP allows explanation of directions*, oral reading in English*, presence of a familiar test administrator, use of Bilingual booklet* or Bilingual dictionary, having the directions repeated, use of large print*, being alone in study carrels, having the test administered in a separate room or with a small group*, use of preferential seating, special lighting, or special furniture, extended time on the same day*, use of Braille writers or word processors*, writing in test booklet, use of scribes*, having the student answer orally/point/sign, and one-on-one administration*. The most frequently used accommodations are identified with an asterisk (*). Table 1 highlights the uses of accommodations on the 2000 NAEP assessment in mathematics. The most common accommodation across all grades was small group setting, followed by extended time and read aloud. Accommodation use changed less than 1% from 1996 to 2000.

Table 1
Summary of Accommodation Use: Students with Disabilities on the 2000 NAEP Administration

		4th Grade			8th Grade			12th Grade	
Accommodation		Number of students	Weighted percent of students sampled		Number of students	Weighted percent of students sampled		Number of students	Weighted percent of students sampled
Bilingual book		0	0		0	0		NA	NA
Large-print book		1	0.04		0	0		1	0.05
Extended Time		55	0.61		68	0.44		51	0.42
Read aloud		20	0.31		28	0.23		7	0.10
Small group		118	2.34		164	1.59		53	0.83
One-on-one		20	0.45		12	0.11		2	0.00
Scribe/computer		2	0.03		1	0.00		0	0
Other		0	0		8	0.07		1	0.01

Impact of Participation on Student Scores and Effect of Accommodation on Proficiency

To examine the effects of accommodations on student scores, NAEP conducted a split-sample design during the 1996 and 2000 administrations of the mathematics test. The sample was divided into two groups based on the availability of accommodations. In the *accommodations permitted* condition, students with disabilities were allowed to take the NAEP with accommodations as permitted by the guidelines. In the *accommodations not-permitted* group, students with disabilities were not able to take the NAEP with accommodations. Table 2 summarizes the participation and achievement results from this study for 2000.

Table 2
Participation and Achievement by Accommodated Condition for the NAEP 2000 Administration

	Accommodations Permitted	Accommodations Not Permitted
4th Grade		
Identified Students with Disabilities	706 (12%)	672 (11%)
Assessed	526 (9%)	292 (5%)
With accommodations	143 (4%)	
Without accommodations	172 (4%)	
Average Score	226	228
Students Scoring Below Basic Proficiency (%)	33	31
Students Scoring at Basic Proficiency (%)	42	43
Students Scoring above Basic Proficiency (%)	25	26
8th Grade		
Identified Students with Disabilities	1206 (10%)	1316 (11%)
Assessed	804 (7%)	597 (5%)
With accommodations	281 (2%)	
Without accommodations	523 (5%)	
Average Score	275	274
Students Scoring Below Basic Proficiency (%)	35	34
Students Scoring at Basic Proficiency (%)	38	38
Students Scoring above Basic Proficiency (%)	27	27
12th Grade		
Identified Students with Disabilities	681 (7%)	680 (7%)
Assessed	453 (5%)	301 (3%)
With accommodations	115 (1%)	
Without accommodations	338 (4%)	
Average Score	300	301
Students Scoring Below Basic Proficiency (%)	36	35
Students Scoring at Basic Proficiency (%)	48	48
Students Scoring above Basic Proficiency (%)	16	16

In general, more students participated in the NAEP when accommodations were permitted. In 4th grade, students taking the NAEP with accommodations scored significantly lower than students not using accommodations. Additionally, a significantly larger number of students in the *accommodations permitted* group did not reach basic proficiency than did students in the *accommodations not-permitted* group. This conclusion, like all descriptive findings conducted on task comparability and, as noted by Tindal (1998) must be viewed with caution as the populations may have been different in critical skills. However, most of this research contains little information of student skills to ensure that the two groups are comparable.

Section III: Review and Synthesis of Research on Math Test Accommodations

In an effort to put some order to the state-level decision-making process, researchers have increasingly pursued a line of research on the effects of accommodations on test performance. Typically, experimental or quasi-experimental designs have been used to sort out various confounds in populations or treatments and ascertain differential effects (e.g., between treatment and control groups). Generally an interaction has been embraced as supporting the use of accommodation: Higher performance levels are predicted for students with disabilities who receive an accommodation and no such improvements for students without disabilities also receiving an accommodation. This outcome, referred to as the "interaction hypothesis" by Sireci, Li, and Scarpati (2003), however, has been questioned: Often accommodation treatments reflect a main effect with no interaction, thereby either justifying its use for everyone or nullifying it for anyone. In this paper, analyze the variables researched as part of the accommodation and then consider the overall implications of the accommodation research on participation rates of students with disabilities on large-scale tests.

Effects of Accommodations on Performance on Large-Scale Mathematics Tests

In this paper, we review the published research literature on the use of accommodations on large-scale mathematics tests, taking no particular stand on the issue of hypothesized overall test outcomes and whether or not accommodations result in improved performance or interactively are effective for specific groups. We do, however, make the following assertion based on the research reviewed in this paper. Research in mathematics testing accommodations highlights specific accommodations that function interactively by the characteristics of individual items and in reference to specific skills of individuals (not their disability). Therefore, the issue of accommodation use cannot be made (or effect cannot be interpreted) at the test level; rather, it needs to be made at the item level. Construct-irrelevant variance (unintended influence of skills and knowledge that are not part of the construct being measured) is item specific.

In our review of the mathematics accommodations research, we focus on four major classes of accommodations: (a) using calculators, (b) reading mathematics problems to students, (c) having the test timed versus having extended time, and (d) using multiple accommodation packages. In general, the findings from using calculators and reading mathematics problems to students clearly document the effect of accommodations to be dependent on the type of items and populations. For some items, calculators are facilitative (e.g., solving fractions problems) and for others detractive (e.g., on complex calculations as part of mathematical reasoning). Similarly, item specific findings are beginning to appear in reading mathematics problems: when the

problems are wordy (both in count and difficulty) and contain several verb phrases, the accommodations appear effective. Likewise, student characteristic is an important variable. The effects of the read-aloud accommodation are more likely with younger students or those with lower reading skills. Finally, the use of extended time appears relatively inert though often it appears as part of other accommodations. For example, calculators and reading mathematics problems often take more time. The findings from these experimental studies reflect similar implications as the naturalistic study of accommodation packages.

Methodology of the Research Review

The first step in this review was to identify all the published literature on accommodations using two steps. Previous reviews were re-reviewed and studies identified that pertained to mathematics testing accommodations; we identified a total of 28 studies published prior to 2000. Then a new search was conducted for all primary investigations published after 2000; we located 14 new studies published in 2000 and beyond. The second step was to sort the research by the populations being studied: studies addressing K-12 students (n=37) and studies addressing college students and adults (n=5). Third, the research was sorted into the general type of accommodations: (a) use of calculators, (b) reading the mathematics test, (c) extended time, and finally, (d) multiple accommodations as part of a package. These categories were derived inductively from the research identified in step one. Step four involved abstracting the research into brief summaries that identified the author and publication date, the type of accommodation used, the population studied, and the main findings; these abstracted summaries have been listed in the third section. The fifth and final step in this review involved re-reading the primary studies to better understand the methodology of the research and consider its implications for the use of this accommodation on the National Assessment of Educational Progress (NAEP), particularly in terms of participation rates: Would the use of the accommodation influence who takes the NAEP mathematics test in grades 4, 8, and 12.

Synthesis of Primary Research on Accommodations in Mathematics Testing

In the summary that follows, the research has been limited to only those studies that address mathematics test accommodations with relevance for NAEP testing. The primary studies being synthesized in this section have been abstracted in Appendix B. N.B. If the studies in the abstract were not done with populations or tests having relevance for NAEP testing, they have not been synthesized in this section.

Calculator use. Probably the earliest studies of specific accommodations in mathematics testing were conducted on the use of calculators, in part because of the stance of the National Council of Teachers of Mathematics with their unequivocal support for having them readily available throughout all teaching and testing. Although much of this research examined the effects of calculators in general rather than as accommodations for students with disabilities, the findings are very relevant.

The research on calculator use presents exemplary methodology because of the dual emphasis on items and populations. Rarely is the study of calculators simply focused on achievement gains in performance. Rather, the focus typically is on the type of items where calculators enhances versus detracts from performance and characteristics of the populations taking the test. Likewise, prior experience with and collateral effects from calculators is considered. For example, (Loyd,

1991) specifically created different items that were thought to favor calculator use, to be neutral (or transparent), or detract from performance when calculators were used. Results confirmed this grouping of items: "the effect of calculator use differs by the item types included in the tests" (p. 21). Using a different methodology but substantially similar focus, Cohen and Kim (1992) used differential item functioning (DIF) to ascertain the effects of calculators on performance on problems with differing computational demands. Using two different analytic procedures, they detected DIF for 5 and 12 items and then analyzed the problems for their demands and concomitant advantages and disadvantages by using a calculator. As they note, "analysis of item-level functioning is an important component of any effort to detect and understand the impact of calculator use…" (p. 318). Importantly, of the 12 items, eight were more difficult when the calculators were permitted; they also note that the type of calculator may influence the results with those having more function keys providing an unfair advantage. In the research by Bridgeman, Harvey, and Braswell (1995), high school juniors planning to attend college and taking the Preliminary Scholastic Aptitude Test (PSAT) were studied using three different item types: regular mathematics, quantitative comparisons, and student-produced responses. All items were rated for "expected calculator sensitivity" and students were questioned about which items they used calculators and the degree to which it helped. They found "calculator and no calculator groups differed by an insignificant 4 SAT points" (p. 327). More importantly, they investigated effects of race and gender (no difference reported), practice or fatigue (none found), experience with calculators (a significant effect was found), and speededness (none found). In their analysis of DIF, they found consistency in DIF categories of calculator effects (large, moderate, trivial, and negative) and both the ratings of calculator sensitivity by test developers as well as students' judgments of helpfulness. Finally, as reported by Scheuneman, Camara, Cascallar, and Lawrence (2002) items identified as favoring calculator use required computations or the use of fractions; items favoring nonuse of calculators tended to be reasoning items that included "numeric values, but required manipulations for which a calculator was unlikely to be of assistance" (p. 107). Calculator use, however, was inversely related to test completion: "The more examinees used calculators, the less likely they were to finish" (p. 108) (though it was the more capable students who used them more often).

Read aloud. The most significant research to arise in the area of mathematics testing is allowing students to have mathematics tests read to them. This line of research has become prominent for four reasons: (a) most multiple-choice mathematics tests require students to have fairly extensive reading skills, (b) the correlation between multiple-choice mathematics and reading tests is relatively high (.70), and (c) recent legislation (both the Individuals with Disabilities Education Act of 1997 and No Child Left Behind) require participation of students with disabilities in large-scale testing. Therefore, a number of researchers have investigated this accommodation to eliminate reading as an access skill. Importantly, this area of research has become increasingly sophisticated in its focus on both items and populations.

Though not directly including students with disabilities, an early study by Wheeler and McNutt (1983) investigated the influence of syntax on eighth-grade students with low abilities to solve mathematics word problems. They found differences between easy or medium and hard syntactically complexities. Using Item Response Theory (IRT), Ansley and Forsyth (1990) focused on "how different abilities interact when examinees respond to test items" on the Iowa Tests of Basic Skills (p. 320). A two by two matrix was used to summarize how four categories

of reading and mathematics ability interacted (very low, low, high, and very high) to create unidimensional, compensatory, and noncompensatory skill relationships. They found the "majority of the items were compensatory" (p. 323) requiring a relatively heavy reading load and a light computation load. In a series of studies by Tindal and colleagues, the influence of reading on mathematics multiple-choice tests have been studied.

An early study by Tindal, Heath, Hollenbeck, Almond, and Harniss (1998) found that reading a mathematics multiple-choice test to fourth-grade students (with Individualized Educational Programs in reading) resulted in no significant differences from general education students ranked low in achievement although when students read the test themselves these two groups had been different. In a follow-up study, Helwig, Tedesco, Heath, Tindal, and Almond (1999) used video-taped reading to standardize the process and found no differences for students in sixth grade; however, when items and students were analyzed more specifically (with reference to problem type: number of words, verb phrases, and difficult words), significant differences were found for students low in reading and for items complex in language. For them, the number of verbs appeared to be an important dimension that influenced the challenge of the language in mathematics problems: "Items that contained large numbers of words, verbs, and unfamiliar vocabulary resulted in more impressive differential performance" (p. 120).

When this line of research was extended comparing a video presentation of the read aloud versus a computer presentation of the read aloud, Hollenbeck, Rozek-Tedesco, Tindal, and Glasgow (2000) framed the issues as pacing: Students were teacher paced with the video but could move at their own pace with the computer. Significant differences were found in favor of self-pacing, particularly for students with disabilities. Finally, in a study that was presented in its entirety by Tindal (2002) and summarized by Helwig, Rozek-Tedesco, and Tindal (2002), students in grades 4–5 and 7–8 participated in both a video and a standard administration of a mathematics test. Tindal (2002) found "only limited evidence, and only at the elementary level, that reading test items aloud was an effective accommodation" (p. 46) though a form effect also was present. No effect (of any type) was found for middle-school students. When these data were analyzed with a subset of the population, Helwig, Rozek-Tedesco, and Tindal (2002) reported significant main effects for all students (both those with and those without disabilities) when presented a video read aloud. When looking at the entire distribution of this subset, they reported that "for elementary students, a slightly greater percentage of general education students performed worse with the video (48%) than the standard version (46%), while for special education students, the opposite was true: More students performed better with the video (53%) than the standard version (40%)" (p. 17).

In two studies recently completed on reading mathematics research, the findings have been consistent in the interaction of item type with individual skill (in both reading and mathematics). For Johnson (2000), a state test was used in successive years to examine both a test and a read aloud treatment effect. Indeed, she found a test effect, which she explained as likely due to fatigue. A small treatment (and nearly significant interaction) also was found with one of the groups containing low reading students with disabilities versus a group of general education students, leading her to conclude both that "students with learning disabilities benefited from having the math test read to them" and that "reading the math items may benefit only students who are at lower levels of performance in the mathematics assessment" (p. 265). The study by

Meloy, Deville, and Frisbie (2002) focused on the effect using the Iowa Test of Basic Skills (including a mathematics subtest) for middle school students. They found a main effect for both students with and without disabilities, though the read aloud may have been more effective for students with disabilities. They also found the read aloud condition resulted in more time being taken than in the standard administration.

Extended time. This area has not been extensively researched and when studies have been conducted, the findings have been less than clear and certainly not supportive. The three studies by Alster (1997), Johnson (2000), and Munger and Loyd (1991) with relevance for NAEP indicate no changes when students with disabilities take the test under timed versus untimed conditions. The most significant problem with this area of research is the lack of specificity in the populations studied and the confounded treatment with other accommodations (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000).

Administration accommodation packages. When viewed in general, accommodation packages have been found to have positive effects for students with disabilities. In a study examining the effects of various accommodations on the performance of 4th grade students on the Wisconsin Student Assessment System, McKevitt, Marquart, Mroch, Schult, Elliott, and Kratochwill (2000) found positive effects for students with disabilities (78.1-81%) and students without disabilities (51-54.5%). Effect sizes for the accommodated versus non-accommodated items ranged from .88 - .94 for students with disabilities, as compared to.45 - .55 for students without disabilities who took the tests with accommodations and .44 for students without disabilities taking the standard format. Using the same measure and target population, Elliott, Kratochwill, and McKevitt (2001) documented positive effects for 75% of students with disabilities. Accommodations were also found to be beneficial for students without disabilities. Schulte, Elliott, and Kratochwill (2001) also found gains for both students with and without disabilities when provided with various accommodations on the TerraNova Multiple assessment Battery.

In an analysis of the Kentucky statewide assessment, Koretz (1997) focused on item functioning with the use of several accommodations. Level of difficulty was affected by accommodations: Fewer students with disabilities and receiving accommodations responded with a blank or zero score than those with disabilities and no accommodations. Most students with disabilities received two or more accommodations (over 80% in 4th grade and 67% in 8th and 11th grades). The effects of accommodations on scores was significant: “Students with disabilities and who received accommodations often scored higher than grade level general education peers who received no such accommodation, though considerable variation in effects also was present. Finally, item-total correlations appeared comparable: items…differentiate between high and low achievers as well as for students with disabilities as for other students” (p. 65), although they also found “frequent and often large DIF for students with disabilities who were provided with assessment accommodations” (p. 66); most of them were in mathematics. Trimble’s (1998) reanalysis of these data did not confirm Koretz’s conclusions.

Recent studies examining item comparability on large-scale state assessments in mathematics across accommodated and non-accommodated formats found differential item functioning (DIF) in some items (Belinski, Thurlow, Ysseldyke, Freidebach, and Freidebach, (2001); Koretz and

Hamilton (2001). These findings suggest that features of the accommodated and non-accommodated items differ, causing construct-irrelevant sources of variance.

Conclusion

In summary, the use of accommodations in large-testing is a function of state policies, teacher decision-making practices, and a growing emergence of research. Rarely do the three coalesce to form a coherent decision-making model. Some of this research is based on naturalistic evaluations, creating problems in interpretation, while other research is more experimental, leading to better inferences of causation and explanations. Much of this research is tentative with conflicting overall test results: some findings show positive effects for all students, other findings reflect interactions between an accommodation and a population. One consistent finding that is beginning to emerge, however, is the interaction of the item with specific skills of individuals. Therefore, it is not so much the test or the person as it is the specific item and the skill that needs to be considered in understanding the recommendations for using accommodations.

Given the item and skill specificity of (accommodation) outcomes, three strategies are currently available for practically addressing the issue of decision-making in accommodations with large-scale tests. First, principles of universal design can be applied to item development so that as much of the construct irrelevant variance as possible can be eliminated. This strategy requires careful attention to the development and review of items so they are free of bias due to gender, language, culture, and disability. Currently, such principles are being articulated by the Center for Assistive Special Technology (CAST) and the National Center on Educational Outcomes (NCEO). Second, tests can be organized into sections so that construct-irrelevant variance is essentially quarantined and appropriate accommodations are used where needed and not permitted where they threaten the construct. For example, if the items reflect statistical-probabilistic algorithms (and not computation facility), calculators could be used; in contrast, where computation facility is the construct, calculators cannot be permitted. Another example is the influence of reading skill on the ability to interpret mathematics story problems. If language and reading is a significant barrier and an (irrelevant) access skill, these items should be read to students; however, for items tapping algorithms and operations embedded in the story problem, no such reading should be allowed. A third and final approach is to use computer-adapted testing in this process, basing the presentation of items not only as a function of item characteristic curves (ICC) and distribution on an ability scale, but also as a function of the item's target construct relative to an access skill. This kind of system is not presently available but may eventually become part of the testing landscape. See the computer-based research being conducted by Ketterlin-Geller (2003). In this approach, students take specific skill tests with a computer and then are assigned particular accommodations for items that would otherwise preclude their successful performance on target skills. In the end, the decision-making reflects a smart system sensitive to items and individual skills in an interactive manner.

We end where we began: Who participates in large-scale testing dictates who counts. Even more important, it dictates what counts. From the policy and practice perspective, we believe that what counts the most is the decision-making of teachers and IEP teams. From the research perspective, what counts is the interaction of specific items (which define the construct, not overall tests) and specific skills of individuals (not type of disability). In both practice and research, clarity is being achieved but not consistency and systematicity, which is critical connecting them together.

Appendix A: Abstracted Primary Summaries on Mathematics Test Accommodations in K–12

The first summary of research on test accommodations was published by Willingham, Bennet, Braun, Powers, Ragosta, and Rock (1988) addressing a number of issues in assessment of individuals with disabilities. Summarizing a rather extensive analysis of accommodations in 'handicapped' individuals taking various college entrance examinations, these authors focus on score and task comparability and then addressed a number of technical analyses to outcomes including reliability, factor structure, differential item functioning, prediction of performance, admission decisions, and test content.

Some of the first summaries of ETS research (and others) on test accommodations of students with disabilities were reported by Thurlow, Ysseldyke, and Silverstein (1993) and Thurlow, Hurley, Spicuzza, and Sawaf (1996) from the National Center on Educational Outcomes. The first publication was confined to outcomes from two ETS studies and the second publication expanded to six studies. After addressing a host of issues surrounding the use of accommodations (policy and legal considerations, type of test, and type of decision), they conclude that, "we will continue to have confusion over policies, scores, and interpretations of data. This confusion will not end until practices are more consistent" (1993, p. 20).

Olson and Goldstein (1997) published a report through the National Research Council on the participation of students with disabilities or limited English proficiency in large-scale assessments. The focus of this document is to summarize the current (at that time) research activities on accommodations from the National Center on Educational Statistics, the National Academy of Science, the National Center on Educational Outcomes, the Office of Educational Research and Improvement, the Office of Special Education and Rehabilitation Services, the Council of Chief State School Officers, and Educational Testing Service.

Chiu and Pearson (1999) summarized the results from a meta-analysis of 30 studies on the effects of test accommodations for students with disabilities. They based their summary on an interaction analysis (e.g., the accommodation is effective in changing the performance levels of a target population and fails to change performance for general education) and concluded that "overall, using accommodation effect ($g_{relative\_effect} = g_{target} - g_{regular\ ed}$) accommodations had a positive effect on the target population and an almost zero effect on general education students" (p. 15). Specifically, they reported an effect size of .16 for the accommodations with the target group and .02 with general education students.

Tindal and Fuchs (1999) published a summary of 115 studies completed over two decades on the assessment of students with disabilities in large-scale testing programs. The summary of the studies was organized into seven major categories and 21 specific practices, with the most frequent researched accommodations being extended time and use of computers in administration. This summary used a narrative review of all studies by abstracting the type of accommodation, the type of subjects, the dependent variable, and the findings.

A 2000 Special Interest Group Yearbook (Research on Inclusion of Students with Disabilities and Limited English Proficient Students in Large Scale Assessments) was published,

summarizing research on the inclusion of students with disabilities in large-scale assessment programs. A wide range of papers, presentations, and policy briefs from funded projects are abstracted in this document, with very few publications referencing primary findings (Abedi, 2000).

Thompson, Blount, and Thurlow (2002) extended the literature review by Tindal and Fuchs (1999) with an analysis of 46 articles examining the effects of accommodations on the test scores of students with disabilities. Summaries of the research were organized by accommodation, research questions, dependent variables, participant characteristics, research designs, findings, limitations, and recommendations for future research. Oral presentation and extended time were the most commonly studied accommodations in the areas of mathematics and reading/language arts. Most studies focused on the effects of accommodations on student scores on criterion-referenced tests. In general, positive results were observed for computer-based delivery, oral presentation, and extended time. The authors highlight the need for replication studies to verify the tentative results.

The most recent summary of research on test accommodations was published by Sireci, Li, and Scarpati (2003) from the Center for Educational Assessment at the University of Massachusetts – Amherst (Research Report 485). Unlike prior publications, this group specifically analyzed the interaction hypothesis embedded in 150 papers with a total of “38 studies from accommodated exams with 21 using an experimental design” (p. 11) and about half of the studies being published in peer-reviewed journals and the other half being technical reports. Their final conclusion was that the vast majority of studies showed improvements with accommodations for all students.

Appendix B: Abstracted Primary Studies on Mathematics Test Accommodations in Grades K–12

The research has been organized into four major categories: (a) calculator use, (b) read aloud, (c) extended time, and (d) accommodation packages. With each of these categories, the studies have been arranged chronologically.

Calculator Use

Lloyd (1991) analyzed the effect of a calculator for high school students on four types of problems and found that, although about half the students did not use a calculator, its use was predicted by the problem type with positive effects accrued for only one type of problem.

Cohen and Kim (1992) found that, for students enrolled in precalculus and calculus courses, the use of calculators was beneficial for some items and impeded performance on other items (due to its inappropriate use).

Bridgeman, Harvey, and Braswell (1995) compared the performance of college-bound juniors on the Scholastic Aptitude Test when they completed it with a calculator versus without a calculator. They reported an overall modest score increase from use of calculator (across all levels of item difficulty and student ability) with variation in type of effects for individual problems.

Hanson, Brown, Levine, and Garcia (2001) investigated the effect of calculator types on the performance of eighth-grade students with disabilities on the National Assessment of Educational Progress (NAEP). Students completed problems using a standard calculator or using a familiar calculator. No differences were observed for problem accuracy, time needed to complete the questions, or amount of calculator use based on calculator type.

Scheunemann, Camera, Cascallar, Wendler, and Lawrence (2002) reported on the use of calculators in the SAT was associated with higher performance, though students with higher abilities were more likely to bring calculators with them to the testing situation; students who used them more and used more scientific calculators also performed higher.

Read Aloud

Tindal, Heath, Hollenbeck, Almond, and Harniss (1998) compared a read aloud administration of a mathematics test to fourth grade students to a standard administration. A differential outcome was found: Students with disabilities and IEPs in reading or mathematics improved while no such improvement occurred with low-ranked students in general education.

Helwig, Tedesco, Heath, Tindal, and Almond (1999) compared a read aloud condition to a standard administration for a mathematics test given to sixth-grade students and reported an interaction between the type of problem difficulty (those with more verbs) and the read aloud condition, particularly for students with low reading but intact mathematics skill levels.

Weston (April, 1999) had students complete a mathematics test in a standard manner and with a read aloud accommodation for students in fourth grade. He reported a significant interaction: a larger effect was found from the use of the accommodation with students having a disability.

Hollenbeck, Rozek-Tedesco, Tindal, and Glasgow (2000) examined the effects of student-paced verses teacher-paced oral presentation of mathematics items for seventh-grade students with disabilities in mathematics and reading on the state assessment system. The student-paced condition delivered items using computer-based video clips. The mean scores increased for students with disabilities and low performing general education for the student-paced oral presentation of items.

Pomplun and Omar (2000) investigated item comparability for the oral presentation accommodation on the Kansas Assessment Program for fourth-grade students with disabilities. The oral presentation of items did not change the functioning of the items.

Johnson (2000) studied the effects of oral presentation of items on the Washington Assessment of Student Learning for fourth-grade students with disabilities in reading. Students with disabilities had significantly larger gain scores with the oral presentation accommodation than students in general education.

Johnson, Kimball, and Brown (2001) studied the effects of American Sign Language (ASL) on the performance of fourth-, seventh-, and tenth-grade students with hearing impairments on the Washington Assessment of Student Learning (WASL) in mathematics. Presentation of items in ASL may result in a loss of information needed to successfully complete the problem. Therefore, ASL may alter the comparability of items.

Tindal (2002) used a video-taped read aloud of mathematics problems for fourth-and eighth-grade students; compared to the standard administration, a small but significant effect was found for some students.

Helwig, R., Rozek-Tedesco, M., and Tindal, G. (2002) examined the effects of oral presentation of items on the state assessment system for fourth-, fifth-, seventh-, and eighth- grade students with learning disabilities. Elementary-age students with disabilities differentially benefited from the oral presentation of items. Greater gains were observed when students faced high word density problems as compared to simple application items. No differential gains in performance were found for middle school students.

Meloy, Deville, and Frisbie (2002) examined the effects of oral presentation of items and extended time for middle school students with learning disabilities in reading on the Iowa Tests of Basic Skills (ITBS). Mean scores for students with disabilities and students in general education increased with the oral presentation of items. Larger gain scores were observed for students with disabilities, however the difference was not significant.

Weston (2002) investigated the effects of oral presentation of items for fourth-grade students on the National Assessment of Educational Progress. Students with disabilities and students in general education benefited from the oral presentation of items, but students with disabilities had greater gains. The effect size for students with disabilities was .64, compared to .31 for students without disabilities.

Timed versus Not Timed

Lord (1956) studied timed and untimed administration of Navy Academy students using tests of verbal and spatial ability and an arithmetic reasoning test; he found no arithmetic reasoning factor present.

Gallina (1989) compared timed and untimed administration with 54 students with two different types of disabilities and 27 students without disabilities, all of them elementary age. One group of students with disabilities benefited from an untimed administration.

Munger and Lloyd (1991) compared timed and un-timed administrations with fifth-grade students on the Math Concepts of the Iowa Test of Basic Skills. No differences were found between the two conditions, either as a main effect or an interaction.

Alster (1997) administered algebra tests under timed and un-timed administration conditions to 88 community college students. Though LD students were significantly lower in performance than non-LD students, no differential benefit was found from the accommodation.

Marquart (2000) examined the effects of providing extended time for eighth-grade students with disabilities on the TerraNova Level 18 Mathematics tests. No significant differences in performance were observed based on group membership.

Fuchs, Fuchs, Eaton, Hamlett, and Karns (2000) investigated the effects of extended time in administration, use of a calculator, and read aloud with fourth grade students. They found no benefit with computation and concept applications. On problem-solving tasks, students with a learning disability differentially benefited from the accommodations.

Administration Accommodation Packages

Abikoff, Courtney, Szeibel, and Koplewicz (1996) studied three auditory stimulation conditions for students with a variety of disabilities on the WRAT-R arithmetic subtest and the Arithmetic Screening Test (AST). Students with ADHD benefited from one of the conditions (use of music).

Swain (1997) administered both a paper-pencil and computer version of the Key Math-R and the CAMT to third-grade students. Although a main effect was found for ability of students (with and without disability), no interaction was found for students by administration.

Olson and Goldstein (1997) documented the effects of several accommodations in the National Assessment of Educational Progress Mathematics test for students with disabilities: more items were omitted and with lower percent correct statistics, resulting in a more difficult and less discriminating test.

Koretz (1997) summarized the outcomes from the most commonly used accommodations on the Kentucky statewide assessment for fourth- and eighth-grade students in special education. He reported significant effects for dictation across several academic areas, including mathematics, as well as across grade levels.

Koretz and Hamilton (2001) investigated the comparability of accommodated items with non-accommodated items for 4th, 5th, 7th, 8th, and 11th grade students with learning disabilities, speech and language impairments, mental retardation, and emotional disturbances on the Kentucky Instructional Results Information System (KIRIS). Some item pairs exhibited differential item functioning. Additionally, no differential benefit was observed for students with disabilities when using the open response format accommodation.

Trimble (1998) analyzed the outcomes from the Kentucky state test (the same data as reported by Koretz [1998]) for students in grades 4, 8, and 11. Significant use of and effects from accommodations was found across grades though none were found to eliminate differences between students with and without disabilities.

Burk (1998) used large print, increased spacing, and audio delivery of problems with several computer-administered tests, one of which was the Maryland Functional Math Test. He reported significant improvement with increased spacing and audio delivery for students with learning disabilities but not for those with developmental disabilities or in general education.

McKevitt, Marquart, Mroch, Schulte, Elliott, and Kratochwill (2000) studied the effects of various accommodations including oral presentation, simplified language, and extended time, on the performance assessments from the Wisconsin Student Assessment System for 4th grade students with disabilities. Accommodations had positive effects for 78.1 to 81% of students with disabilities and 54.5 to 51% of students without disabilities. When comparing performance on the accommodated versus non-accommodated items, effect sizes for students with disabilities ranged from .88 - .94. Effect sizes for students without disabilities who took the tests with accommodations ranged from .45 to .55, as compared to the effect size of .44 for student without disabilities who took the test without accommodations.

Schulte, Elliott, and Kratochwill (2001) examined the effects of various accommodations (including oral presentation of items, simplified language, dictated response, small group administration, extended time, and frequent breaks) on 4th graders performance on the TerraNova Multiple Assessment Battery. Accommodations were assigned to students with disabilities based on IEP recommendations. General education students were matched to students with disabilities. In general, the authors found that both student groups benefited from the accommodations. Larger but non-significant gains were observed for students with disabilities. No differential benefits for students with disabilities were observed for oral presentation of items, extended time, and constructed response. Differential benefits were observed for multiple-choice item formats and other accommodation packages.

Belinski, Thurlow, Ysseldyke, Freidebach, and Freidebach (2001) examined the effects of multiple accommodations, including oral presentation of items, small group administration, and extended time for 4th grade students with disabilities in reading on the Missouri Assessment Program. In mathematics, approximately one-fifth of the items exhibited differential item functioning when read aloud to the student. Non-standardized administration of the read-aloud accommodation may have influenced the results.

Elliott, Kratochwill, and McKevitt (2001) studied the effects of accommodations for 4th grade students with disabilities on mathematics performance assessments developed for the Wisconsin Student Assessment System. Students received either a standard package of accommodations (extended time, supported reading of directions, read aloud of selected words, and verbal encouragement) or accommodations recommended by the teacher or IEP team. The authors found that accommodations had positive effects for 75% of the students with disabilities. Students without disabilities also benefited from the accommodations.

References

Abedi, J., Kim-Boscardin, C., & Larson, H. (2000). *Sumaries of research on the inclusion of students with disabilities and limited English Proficient students in large-scale assessments*. Los Angeles, CA: National Center for Research on Evalaution , Standards, and Student Testing.

Abikoff, H., Courtney, M. E., Szeibel, P. J., & Koplewicz, H. S. (1996). The effects of auditory stimulation on the arithmetic performance of children with ADHD and non-disabled children. *Journal of Learning Disabilities, 29*, 238-246.

Alster, E. H. (1997). The effects of extended time on algebra test scores for college students with and without learning disabilties. *Journal of Learning Disabilities, 30*(2), 222-227.

Ansley, T. N., & Forsyth, R. A. (1990). An investigation of the nature of the interaction of reading and computational abilities in solving mathematics word problems. *Applied Measurement in Education, 3*(4), 319-329.

Bielinski, J., Thurlow, M., Ysseldke, J., Frieidebach, J., & Friedebach, M. (2001). *Read aloud accommodations: Effects on multiple-choice reading and math items*. Minneapolis, MN: National Center on Educational Outcomes.

Bridgeman, B., Harvey, A., & Braswell, J. (1995). Effects of calculator use on scores on a test of mathematical reasoning. *Journal of Educational Measurement, 32*(4), 323-340.

Burk, M. (1998). *Computerized test accommodations: A new approach for inclusion and success of students with disabilities.* Paper presented at the Technology and the Education of Children with Disabilities: Steppingstones to the 21st Century, Wasinbgton, D. C.

Chiu, C. W., & Pearson, D. (1999). *Synthesyzing the effects of test accommodations for special education and limited English proficient students.* Paper presented at the CCSSO Large Scale Assessment Conference, Snowbird, UT.

Cohen, A. S., & Kim, A. (1992). Detecting calculator effects on item performance. *Applied Measurement in Education, 5*(4), 303-320.

Crawford, L., & Tindal, G. (in press). Effects of a read-aloud modification on a standardized reading test. *Exceptionality*.

Destefano, L., Shriner, J. G., & Lloyd, C. A. (2001). Teacher decision making in participation of students with disabilities in large-scale assessment. *Exceptional Children, 68*(7-22).

Elliott, S., Kratochwill, T., & McKevitt, B. (2001). Experimental analysis of the effects of testing accommodations on the scores of students with and without disabilities. *Journal of School Psychology, 39*, 3-24.

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*(65-85).

Gajria, M., Salend, S. J., & Hemrick, M. A. (1994). Teacher acceptability of testing modifications for mainstreaming students. *Learning Disabilities Research & Practice, 9*, 236-243.

Gallina, N. B. (1989). *Tourette's syndrome children: Significant achievement and social behaviors.* Unpublished Dissertation Abstracts International, 50, 0046, CIty University of New York, New York.

Hanson, K., Brown, B., Levine, R., & Garcia, T. (2001). Should standard calculators be provided in testing situations? An investigation of performance and preference differences. *Applied Measurement in Education, 14*(1), 59-72.

Helwig, R., Rozek-Tedesco, M., & 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., Tedesco, M., Heath, B., Tindal, G., & Almond, P. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth grade students. *The Journal of Educational Research, 93*(2), 113-125.

Hollenbeck, K., Rozek-Tedesco, M.A., Tindal, G., & Glasgow, A. (2000). An exploratory study of student-paced versus teacher-paced accommodations for large-scale math tests. *Journal of Special Education Technology,, 15*(2), 29-38.

Hollenbeck, K., Tindal, G., & Almond, P. (1998). Teachers' knowledge of accommodations as a validity issue in high-stakes testing. *The Journal of Special Education, 32*(3), 175-183.

Johnson, E., Kimball, K., & Brown, S.O. (2001). American sign language as an accommodation during standards-based assessments. *Assessment For Effective Intervention, 26*(2), 39-47.

Johnson, E. S. (2000). The effects of accomodations in performance assessments. *Remedial and Special Education, 21*(5), 261-267.

Ketterlin-Geller, L. R. (2003). *Establishing a validity argument for universally designed assessments.* Unpublished Doctoral Dissertation, University of Oregon, Eugene, OR.

Koretz, D. (1997). *The assessment of students with disabilties in Kentucky*. Los Angeles, CA: National Center for the Study of Evaluation, Standards, and Student Testing.

Koretz, D., & Hamilton, L. (2001). Assessment of students with disabilities in Kentucky: Inclusion, student performance, and validity. *Educational Evaluation and Policy Analysis, 26*(2), 39-47.

Lord, F. M. (1956). A study of speed factors in tests and academic grades. *Psychometrika, 21*(1), 31-50.

Loyd, B. H. (1991). Mathematics test performance: The effects of item type and calculator use. *Applied Measurement in Education, 4*(1), 11-22.

Marquart, A. (2000). *The use and effects of testing accommodations on math and science performance assessments.* Paper presented at the CCSSO Large-Scale Assessment Conference, Snowbird, UT.

McKevitt, B., Marquart, A., Mroch, A., Schulte, A. G., Elliott, S. N., & Kravochwill, T. R. (2000). *Understanding the effects of testing accommodations: A single case approach.* Paper presented at the CCSSO Large-Scale Assessment Conferences, Snowbird, UT.

Meloy, L. L., Deville, C., & Frisbee, 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.

Munger, G. F., & Loyd, B. H. (1991). Effect of speededness on test performance of handicapped and nonhandicapped examinees. *Journal of Educational Reseach, 85*(1), 53-57.

Olson, J. F., & Goldstein, A. A. (1997). *The inclusion of students iwth disabilites and limtied English Proficient students in large-scale assessments* (NCES 97-482). Washington, D. C.: National Center for Education Statistics.

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*, 21-29.

Robelen, E. W. (September 3, 2003). State reports on progress vary widely. *Education Week,* pp. 1, 37.

Scheuneman, J. D., Camara, W. J., Cascallar, C. W., & Lawrence, I. (2002). Calculator access, use, and type in relation to performance on the SAT I: Reasoning test in mathematics. *Applied Measurement in Education, 15*(1), 95-112.

Schulte, A. G., Elliott, S., & Kratochwill, T. (2001). Effects of testing accommodations on standardized mathematics test scores: An experimental analysis of the performance of students with and without disabilities. *School Psychology Review, 30*(4), 527-547.

Sireci, S. G., Li, S., & Scarpati, S. (2003). *The effects of test accommodations on test performance: A review of the literature* (Research Report 485). Amherst, MA: Center for Educational Assessment.

Swain, C. R. (1997). *A comparison of computer-administered test and a paper and pencil test using normally achieving and mathematically disabled young children.* Unpublished Dissertation Abstracts International, 47, 0125, University of North Texas.

Thompson, S., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001* (Technical Report 34). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., Hurley, C., Spicuzza, R., & Sawaf, H. E. (1996). *A review of the literature on testing accommodations for students with disabilities* (State Assessment Series Report 9). Minneapolis, MN: National Center on Educational Outcomes.

Thurlow, M. L., Ysseldyke, J. & Silverstein, B. (1993). *Testing accommodations or students with disabilities: A review of the literature* (Synthesis Report No. 4). Minneapolis, MN: National Center on Educational Outcomes.

Thurlow, M. L., Ysseldyke, J. & Silverstein, B. (1995). Testing accommodations for students with disabilities. *Remedial and Special Education, 16*(5), 260-270.

Thurlow, M. L., Ysseldyke, J., & Silverstein, B. (1998). *The research basis for the need for research on accommodations.* Paper presented at the American Education Research Association, San Diego, CA.

Tindal, G. (1998). *Models for understanding task comparability in accommodated testing.* Washington, D. C.: Council of Chief State School Officers.

Tindal, G. (2002). *Accommodating mathematics testing using a videotaped read aloud administration.* Washington, D. C.: Council of Chief State School Officers.

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 empirical study of student response and test administration demands. *Exceptional Children, 64*(4), 439-450.

Trimble, S. (1998). *Performance trends and use of accommodations on a statewide assessment* (State Assessment Series). Minneapolis, MN: National Center on Educational Outcomes.

Weston, T. J. (2002). *The validity of oral accommodation in testing.* Paper presented at the NAEP Validity Studies.

Weston, T. J. (April, 1999). *The validity of oral presentation in testing.* Paper presented at the American Educational Research Association, Montreal, Canada.

Wheeler, L. J., & McNutt, G. (1983). The effect of syntax on low-achieving students' abilities to solve mathematical word problems. *The Journal of Special Education, 17*(3), 309-329.

Willingham, W. H., Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A., & Powers, D. E. (1988). *Testing handicapped people.* Boston, MA: Allyn & Bacon.