The Effects of Test Accommodation on Test Performance: A Review of the Literature[1]

Stephen G. Sireci, Shuhong Li, and Stanley Scarpati

University of Massachusetts Amherst

[1] Center for Educational Assessment Research Report no. 485. Amherst, MA: School of Education, University of Massachusetts Amherst. This paper was commissioned by the Board on Testing and Assessment of the National Research Council of the National Academy of Sciences.

Abstract

Over 150 studies pertaining to test accommodations were identified in the literature and 40 studies that empirically studied the effects of test accommodations on the performance of students with disabilities or English language learners were reviewed. The results of these studies are discussed as are the internal and external validity of the authors' conclusions. All studies were critiqued with respect to the interaction hypothesis that test accommodations should improve the test scores for targeted groups, but should not improve the scores of examinees for whom the accommodations are not intended. Data are provided regarding the types of accommodations studied. In general, consistent conclusions were not found across studies due to the wide variety of accommodations, the ways in which they were implemented, and the heterogeneity of students to whom they were given. However, a fairly consistent finding was that the accommodation of extended time improved the performance of students with disabilities more than it improved the performance of students without disabilities. In light of this finding and similar results in some studies focusing on other accommodations, a revision of the interaction hypothesis is proposed. Directions for future research and for improved test development and administration practices are also proposed.

The Effects of Test Accommodation on Test Performance: A Review of the Literature

Standardized tests are a common part of educational systems throughout the United States. However, some aspects of standardized testing make the administration of these tests infeasible or unfair to certain students, particularly students with disabilities or students who are not native speakers of English. To address this problem, many tests are altered, or the test administration conditions are adjusted, to "accommodate" the special needs of these students. This practice is designed to level the playing field so that the format of the test or the test administration conditions do not unduly prevent such students from demonstrating their "true" knowledge, skills, and abilities.

The practice of accommodating standardized tests for certain groups of students is often heralded as promoting equity in assessment. However, the resulting oxymoron—an accommodated standardized test—is not without controversy. At least two questions fuel the debate on the value of test accommodations. One question is "Do the test scores that come from nonstandard test administrations have the same meaning as test scores resulting from standard administrations?" A related question is "Do current test accommodations lead to more valid test score interpretations for certain groups of students?" These questions, and many related ones, have presented significant challenges for psychometricians, educational researchers, and educational policy makers for decades.

The professional literature contains numerous published and unpublished empirical and non-empirical studies in the area of test accommodations. This literature is vast and passionate. In many cases, researchers argue against test accommodations in the name of fairness to the majority of examinees who must take the tests under perceivably stricter, standardized

conditions. In many other cases, researchers argue that test accommodations are the only way to validly measure the knowledge, skills, and abilities of significant numbers of students.

In this paper, we critically evaluate the literature on test accommodations focusing on those studies that evaluated the *effects* of test accommodations on students' test *performance*. Several fundamental questions guided our review including

(a) Do test accommodations affect the test performance of students with disabilities?

(b) Do test accommodations affect the test performance of students who are non-native speakers of English?

(c) What specific types of test accommodations best facilitate valid score interpretations for specific types of students?

Many other important questions are addressed in this review; however, our major purpose is to critically evaluate the *empirical* research in this area to inform the educational measurement and policy communities about the pros and cons of test accommodations.

Willingham et al. (1988) classified students with disabilities (SWD) into one of four categories: visually impaired, hearing impaired, physically disabled, or learning disabled. Currently, test accommodations are used for students classified into one or more of these categories as well as for students who are non-native speakers of English (hereafter referred to as English language learners or ELL). Our review looks at the literature associated with the first four categories (i.e., SWD) as well as with ELL.

Understanding the "Interaction Hypothesis"

A fundamental tenet underlying the provision of test accommodations to certain examinees is that some examinees need them, while most do not. Put in the perspective of test validity theory, some features of a standardized test administration introduce *construct-irrelevant*

variance for some students. For example, a student who learned Spanish as her first language may do worse on a math test administered in English than on a parallel math test administered in Spanish. In this case, English proficiency may be considered extraneous to the math construct targeted by the test, but would certainly affect her test performance on the English language version of the test. Similarly, the ability to maneuver test materials may introduce construct-irrelevant variance for examinees with motor disabilities and the ability to see would obviously present construct-irrelevant difficulties for a blind student taking a standard math exam.

From these examples it is clear that test accommodations are designed to promote fairness in testing and to lead to more accurate interpretations of students' test scores. However, if the accommodation leads to an unfair advantage for the students who get them, for example, if *everyone* would benefit from the accommodation, then the scores from accommodated exams may be invalidly inflated, which would be unfair to students who do not receive accommodations. For this reason, an *interaction hypothesis* has been proposed to justify the use of test accommodations. This hypothesis states that test accommodations will lead to improved test scores for students who need the accommodation, but not for students who do not need the accommodation (Malouf, 2001, cited in Koenig, 2002; Shepard, Taylor, & Betebenner, 1998; Zuriff, 2000). That is, it is hypothesized that there is an interaction between accommodation condition (accommodated versus standard test administration) and type of student (e.g., students with disabilities versus students without disabilities) with respect to test performance. This hypothesis has also been called the *maximum potential thesis* (Zuriff, 2000). In our review, we look for evidence in support of this hypothesis.

Psychometric Issues in Test Accommodations

Before discussing our review, a brief discussion of the psychometric issues associated with test accommodations is necessary. Many of these issues have been clearly elucidated in the literature (e.g., Geisinger, 1994; Green & Sireci, 2000; Koretz & Hamilton, 2000; Phillips, 1994; Pitoniak & Royer, 2001; Scarpati, 1991, 2003; Sireci & Geisinger, 1998; Willingham et al., 1988), including guidelines presented in the *Standards for Educational and Psychological Testing* (American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education, 1999).

Psychometric issues in test accommodations stress the need to remove construct-irrelevant barriers to test performance while maintaining integrity to the construct being measured by the test. In situations where individuals who take accommodated versions of tests may be compared to those who take the standard version, an additional validity issue is the *comparability* of scores across the different test formats. This issue was addressed in some depth in the *Standards*, for example

> Score equivalence is easiest to establish when different forms are constructed following identical procedures and then equated statistically. When that is not possible…additional evidence may be required to establish the requisite degree of score equivalence for the intended context and purpose…Some testing accommodations may only affect the dependence of test scores on capabilities irrelevant to the construct the test is intended to measure. Use of a large-print edition, for example, assures that performance does not depend on the ability to perceive standard-size print. In such cases, relatively modest studies of professional judgment may be sufficient to support claims of score equivalence. (AERA, et. al., p. 57)

Although we do not address studies of score equivalence or comparability in this review, it is important to bear in mind that test accommodations are made in the pursuit of more valid test score interpretations.

Review Process

Several sources were used to locate research on the effects of test accommodations on test performance. First, we received a set of papers from the Board on Testing and Assessment of the National Research Council, who commissioned this review. Second we searched two electronic databases: ERIC and PsychInfo. We also contacted several researchers whose work was frequently cited or appeared numerous times in our electronic searches. We sent these researchers a list of the citations of their work that we had and we asked them to send us anything else they may have done in this area. Finally, we searched the web sites of the Center for Research on Evaluation, Standards, and Student Testing (CRESST) and the National Center on Educational Outcomes. These activities yielded over 150 papers, many of which were unpublished. This extensive list was closely reviewed to identify papers that explicitly addressed the effects of test accommodations on test performance. We initially identified 94 documents that fit this description, but there were a small number of papers that we were unable to locate. After reviewing the papers we did locate, we concluded that only 46 studies actually focused on the effects of test accommodations, and only 38 involved empirical analysis. Appendix A is an annotated bibliography of many of the studies we initially found in our scan of the literature. This appendix provides annotations for all of the studies included in our review, as well as for those studies that were relevant, but deemed less important due to methodological concerns or extreme lack of generalizability. A listing of all studies located in our initial scan of the literature, including those that were not annotated, is presented in Appendix B.

Organization of the Review

Our review is structured using three primary criteria: group intended to be helped by the accommodation (SWD or ELL), type of accommodation, and research design. However, this organization is not perfect, since many studies involved multiple accommodations and some looked at both SWD and ELL. Moreover, the term "students with disabilities" includes students with learning disabilities, motor disabilities, and sensory disabilities. The term "English language learners" (ELL) includes examinees who come from a variety of linguistic backgrounds and are in various stages of becoming proficient in English. Although we use these labels to describe general classifications of students for the purpose of accommodation, we do not consider individuals within these groups to be homogeneous, nor do we consider these general labels appropriate for all individuals with a group.

As for type of accommodation, a wide variety of assessment accommodations exist. In their review of the literature, Thompson, Blount, and Thurlow (2002) categorized eleven types of accommodation into one of four general categories: presentation, response, setting, or timing/scheduling. Borrowing this categorization scheme, we defined test presentation accommodations to include oral administration (i.e., "read aloud" protocols where the test directions and/or items are read to the test taker), changes in test content (e.g., simplified language), and changes in test format (e.g., Braille, large print). Timing/scheduling accommodations included allowing extended or unlimited time and breaking up test administration into separate sessions (in a single day or over multiple days). Response accommodations included allowing students to write directly into test booklet or dictate their answers. Setting accommodations typically involved administering the tests individually or in a separate room.

Our discussion of the literature first focuses on accommodations for SWD. Subsequently, we focus on accommodations designed for ELL. Within each of these student groups, we distinguish between literature reviews, experimental studies, and non-experimental studies, stratified by type of accommodation, where possible.

Descriptive Characteristics of Studies

Due to the fact that the literature on the effects of test accommodations is vast, some general descriptions of the number and types of studies that we found are helpful for understanding this body of research. The number of studies that are relevant to specific types of accommodation are presented in Table 1, stratified by student group (SWD or ELL). As illustrated in Table 1, the majority of research in this area has focused on SWD (75%) and the most common accommodations studied were oral administration (31%) and the provision of extra time (20%). These findings are similar to a recent review of the literature conducted by Thompson, Blount, and Thurlow (2002) who found that studies investigating oral administration were the most common, followed closely by studies investigating extended time. In another recent review of the literature, Chiu and Pearson (1999) found that extended time was the most frequently investigated accommodation and setting and response format were least frequently investigated. It should be noted that oral presentation is often given with extended time and so separation of the effects of these two variables is not always possible. Another point to bear in mind about Table 1 is that many studies, such as those that looked at scores from large-scale test administrations, analyzed scores from many different types of accommodations, while other studies focused on just a single accommodation. Extended time and oral presentation were the most common accommodations investigated in those studies that focused on a single accommodation.

Table 1

General Description of Studies Included in the Review

Type(s) of Accommodation	Study Focused On		
	SWD	ELL	Total
Presentation:			
Oral*	22	1	23
Dictionary/Glossary	--	9	9
Linguistic modification of test	--	3	3
Dual-language	--	2	2
Paraphrase	2	--	2
Technological	2	--	2
Braille/Large Print	1		1
Sign Language	1	--	1
Encouragement	1	--	1
Cueing	1	--	1
Spelling assistances	1	--	1
Manipulatives	1	--	1
Timing:			
Extended time	12	3	15
Multi day/sessions	1	--	1
Separate sessions	1	--	1
Response:			
Scribes	2	--	2
In booklet vs. answer sheet	1	--	1
Mark task book to maintain place	1	--	1
Transcription	1	--	1
Setting:			
Separate room	1	--	1
No specifics listed***	4	1	5
Total	56	19	75

Notes: *Includes read aloud, audiotape, or videotape, and screen-reading software. Literature reviews and issues papers are not included.

The studies we reviewed can also be distinguished with respect to the research designs used to address specific questions. A summary of the research designs used across the studies is presented in Table 2. A study was classified as using an experimental design if test administration condition (accommodated or standard) was manipulated *and* examinees were randomly assigned to conditions. Studies were classified as quasi-experimental if the test administration condition was manipulated, but examinees were *not* randomly assigned to conditions. Non-experimental studies include ex post facto studies that compared the results of students who took a test with an accommodation to those who took a standard version of the test and studies that looked at differences across standard and accommodated administrations for the same (self-selected) group of students.

As can be seen from Table 2, only 38 studies involved analysis of data from accommodated exams with 21 using an experimental design. With respect to the quasi- and non-experimental studies, 8 of the 17 studies examined scores from large-scale test administrations and made comparisons across specific groups of students who took accommodated versus standard versions of an exam.

Table 2

Research Design Classifications of Studies

Research Design	Study Focused On		Total
	SWD	ELL	
Experimental	13	8	21
Quasi-experimental	2	4	6
Non-experimental	10	1	11
Total	25	12	38

Note: Literature reviews and issues papers are not included in this table.

The studies we reviewed also spanned several subject areas and grades. Many of the studies involved multiple grades and subject areas. A cross-tabulation of grades by subject area tested is presented in Table 3. As can be seen from this table, most of the studies focused on elementary school grades; math, reading, and science were the most common subject areas investigated. It is also interesting to note that nearly two thirds of the studies focused on students in grades 3 to 8 while the remainder of the studies evaluated the effect of accommodations on test performance for students in grades 9 to 12.

Table 3

Grade by Subject Cross-tabulation

Grade	Math	Reading	Science	Listening	Writing	ELA	Social Studies	U&E	Verbal	Spelling	Study Skills	Total	Cum %
3	1	1	1	--	--	--	1	--	--	1	1	6	4.0
4	14	5	6	2	1	--	2	--	--	1	1	32	27.7
5	4	2	1	--	--	--	1	--	--	1	1	10	35.0
6	2	2	2	--	--	--	--	1	--	--	--	7	40.1
7	6	3	1	2	1	--	1	1	--	--	--	15	51.1
8	4	5	4	--	--	1	1	1	--	--	--	16	62.8
9	1	--	--	--	--	--	--			--	--	1	63.5
10	4	1	1	2	1	1	--	--	--	--	--	11	71.5
11	2	1	1	--	--	--	1	--	--	--	--	5	75.2
12	2	1	1	--	--	1	1	--	--	--	--	6	79.6
HS	--	2	1	--	--	--	1	--	--	--	--	4	82.5
C/U	--	1	--	--	--	--	--	--	--	--	--	1	83.2
PAT	10	3	--	--	--	--	--	--	10	--	--	23	100.0
Total	50	27	19	6	3	3	9	3	10	3	3	137	

Notes: Literature review and issues papers are not included. Some studies did not specify grades or subject areas. HS=high school, C/U=unspecified college or university test, PAT=Postsecondary admissions test, ELA=English language arts, Tech.=Technology, U&E= Usage & Expression

One final way we distinguished among studies was whether they were published in a peer-reviewed journal. Less than half of the studies were published in peer-reviewed journals (41%). Most of the studies were technical reports or conference papers.

Description and Interpretations of Studies

Before reviewing individual studies in this area, we first describe three literature reviews conducted recently. The first two reviews focused on studies related to test accommodations and SWD. The third review focused on studies related to both SWD and ELL.

Literature Reviews

Thompson, et al. (2002) scanned the literature on the effects of test accommodations on the test performance of SWD and found 46 studies relevant to this topic conducted between 1999 and 2001. Their review was comprehensive and included studies with sample sizes ranging from 3 to 21,000. Twenty-four of these studies focused on the effects of one or more accommodations on test scores. The accommodations they noted in the literature are similar to those listed in Table 1. After reviewing these studies they concluded

> …three accommodations showed a positive effect on student test scores across at least four studies: computer administration, oral presentation, and extended time. However, additional studies on each of these accommodations also found no significant effect on scores or alterations in item comparability. (p.2)

This conclusion illustrates a common finding among the test accommodations literature: promising effects are not always replicated across studies. Given these results, Thompson et al. also concluded

> ….additional studies are needed that investigate the effects of accommodations under much more carefully defined conditions…there is a need for clear definition of the constructs tested…At the same time greater clarity in the accommodations needed by individual students need to be added—independent ways of measuring whether each

student who participates in an accommodation study actually needs the accommodation being studied. (pp. 14-15)

Tindal and Fuchs (2000) conducted an extensive review on the effects of accommodations on SWD. Many of the studies they cited are included in the present review; however, they also discussed numerous abstracts that appeared in Dissertation Abstracts International. Since such abstracts do not provide enough detail to critically evaluate the internal validity of the conclusions, we chose not to include them in our review and so interested readers are referred to the Tindal and Fuchs study. Their review included annotated references for over 100 studies and was organized by type of accommodation. Sets of studies pertaining to each accommodation were critiqued according to their quality from a methodological perspective.

In the majority of cases, Tindal and Fuchs noted that the research findings were complex, which prevented unequivocal conclusions. Some exceptions were the use of Braille, large-print, and oral accommodations on math tests, all of which seemed to improve the performance of SWD, but not the performance of students without disabilities. However, they noted that changes in test presentations were not always "differential for students with disabilities versus those without disabilities" (p. 50). Another conclusion they derived from their review was that the accommodation of allowing examinees to mark their answers in the test booklet, rather than on the standard answer sheet, was *not* effective in improving the test performance of SWD or students without disabilities.

Chiu and Pearson (1999) also conducted a literature review in this area. A key feature of their study was the use of meta-analysis to synthesize the findings of studies investigating the magnitude of gains due to test accommodations. They found 30 empirical studies in this area. The most commonly used research design was a repeated measure design with a comparison

group (n=16), followed by repeated measures without a comparison group (n=7), and equivalent control group design (n=7).

Their meta-analysis focused on 20 studies they considered to be acceptable in terms of research design and the provision of sufficient statistical information. With respect to an overall accommodated/standard effect, they concluded that SWD had significant score gains under the accommodation condition, but the gain was small (.16 of a standard deviation unit). This finding supported the interaction hypothesis, since the effect size for students without disabilities was negligible (.06 of a standard deviation unit). However, their analysis of the variation of effects across studies led them to conclude that there was significant variation across studies, which implied that "using the mean effect alone could be misleading because it would fail to portray the diversity of accommodation effects" (p. 15).

Looking at specific accommodations, Chiu and Pearson found that extended time only slightly helped SWD more than it helped students without disabilities. The average effect size gain for SWD under the condition of extended time was .37 standard deviation units, but it was .30 for students without disabilities. Although this finding is consistent with Thompson et al.'s (2002) claim that "research has fairly consistently concluded that extended time helps students with disabilities" (p. 15), it is important to note that the magnitude of this improvement, relative to students without disabilities, was small.

Chiu and Pearson cautioned that the average effect sizes they noted should be interpreted extremely cautiously due to the wide variety of accommodations that were used (and the quality in which they were implemented) and the heterogeneous types of students they were used for. They hypothesized that this variation is probably due differences in what constituted "standard conditions," and the nature of specific implementations of the accommodation.

Given the findings of Thompson et al. (2002) and Chiu and Pearson (1999), it appears that the most common accommodations that show promise for having positive effects on SWD test scores are extended time and oral presentation. Therefore, we next review the literature specific to these accommodations.

Extended Time Accommodation for SWD

Extended time appears to be the most frequent accommodation given to SWD. In fact, extended time, which includes the special case of unlimited time, often accompanies other accommodations such as oral presentation, Braille, separate testing location, etcetera. Therefore, it is not surprising that extended time has received the most empirical study.

Experimental Studies Focusing on Extended Time

Several studies looked at the effects of extended time using an experimental design. In this section, we review these studies.

Runyan (1991a) examined reading score differences between a small sample of college students with and without learning disabilities (LD) using extra time as an accommodation. She hypothesized that students with LD score lower on timed tests than their non-disabled peers, but will score in similar ways under untimed conditions. Her study involved 16 students with LD (identified according to the discrepancy formula approach—1.5 SD difference between IQ and achievement) all with a history of reading problems, with slow reading rates highlighted among their difficulties. Her control group comprised 15 non-LD students who were randomly selected and had no learning disabilities, speech problems, or academic probation. These groups were matched on gender, ethnicity (all white), and total SAT. The Nelson-Denny Reading test was used to derive the dependent measures.

Runyan's (1991a) design involved recording students' scores at the end of the standard test time (20 minutes) and again when the student completed the test (untimed condition). However, the students were not told that they would be given a chance to continue to work on the test after standard time had run out. Raw scores of words per minute were transformed into percentile ranks and used as the dependent measure for each time period. Using separate independent and dependent t-tests, she found that (a) under the "standard time" condition, non-LD students significantly outperformed LD students; (b) students with LD had significant score gains under the "extended time" condition, while non-LD students did not have significant gains; and (c) there was no significant difference between the scores of students with LD when they had extended time and the scores of non-LD students under the standard time condition. These findings supported the interaction hypothesis. However, Zuriff (2000) pointed out that a flaw in her design is that any students who completed the test during the standard time condition were unable to increase their scores under the extended time condition. This ceiling effect represents a significant threat to the validity of her conclusions.

Zuriff (2000) critically reviewed Runyan's (1991a) study and four other experimental studies (i.e., Hill, 1984; Halla, 1988; Runyan, 1991b; and Weaver, 1993; cited in Zuriff, 2000) that directly evaluated the interaction hypothesis with respect to extended time and students with learning disabilities. All five studies were conducted between 1984 and 1993, and only the Runyan (1991a) study was published (the other four studies were unpublished doctoral dissertations). Zurriff described the interaction hypothesis as the *maximum potential thesis*, which states that students without disabilities would not benefit from extra examination time because they are already operating at their "maximum potential" under timed conditions (p. 101).

All five studies compared the performance of students with learning disabilities with students without disabilities under extended time and standard time conditions, and all studies used the Nelson-Denny Reading Test as a dependent measure. One study also used the ACT and another study also used the GRE. The sample sizes for the studies ranged from 31 to 126.

Across the studies, Zuriff found some support for the interaction hypothesis, but his general finding was that extra time helped both students with learning disabilities and students without disabilities. Specifically, he found that students with learning disabilities performed better with extended time in 31 out of 32 comparisons (the anomaly being from an analysis that compared the performance of high-IQ students with learning disabilities on the GRE-Verbal). However, students without disabilities also showed gains under the condition of extended time in half of the comparisons. This finding led Zuriff to conclude that the maximum potential thesis "is not empirically supported" (p. 115). However, it should be noted that the Nelson-Denny Tests, which were involved in 23 of the 32 comparisons, appear to be speeded. All students would be expected to have score gains when extended time is given on a speeded test. The interaction hypothesis was supported in all 6 comparisons involving the ACT, but not in any of the four comparisons involving the GRE. Thus, the results of Zuriff's analysis are equivocal.

Schulte, Elliott and Kratochwill (2001) conducted an experimental study to evaluate the interaction hypothesis. The test used in their study was the fourth grade math test battery from the TerraNova (a standardized achievement test developed by CTB/McGraw Hill). Eighty-six fourth grade students participated in the study: 43 SWD and 43 students without disabilities. All 86 students took the test with one or more accommodations. Accommodations were selected for each individual SWD based on a review of her/his individual education plan (IEP). Students without a disability were matched to a SWD and given the same accommodation that their yoked

partner had. Extra time and reading the test aloud were the most common accommodations, and many of the students used both. The design of the study involved repeated measures for both groups. That is, SWD and students without disabilities took the test under standard and accommodated conditions.

Schulte at al. (2001) found that total scores for SWD improved more between non-accommodated and accommodated test conditions (medium effect size) than did their non-disabled peers (small effect size). For multiple-choice items, SWD benefited more from the accommodations (medium effect size) than did students without disabilities (negligible effect size). No differences were detected on scores from the constructed response items. From the perspective of state testing performance criteria, about equal percentages of SWD and students without disabilities achieved a higher proficiency classification under the accommodation condition (40% SWD and 37% students without disabilities). The remainder of the sample either demonstrated no change in proficiency rating (47% SWD, 49% non-SWD), or decreased in proficiency when allowed accommodations (about 14% for each group). These findings do not support the interaction hypothesis in that substantial score gains were observed for students without disabilities who were given accommodations. However, it should be noted that the gains for SWD were larger.

Schulte et al. also found small effects for a few SWD receiving accommodations other than extra time and oral presentation. Although the extra time/oral presentation package was the most common accommodation, it was not more or less effective than other accommodation packages (.16 for students without disabilities; .30 for SWD). For a limited number of SWD receiving accommodations other than extra time and oral presentation, the average effect size was .50, while the average effect size their non-disabled peers was .14.

Quasi-experimental Studies Focusing on Extended Time

Huesman and Frisbie (2000) conducted a quasi-experimental study to examine the effects of extended time on students with learning disabilities performance on the Iowa Tests of Basic Skills (ITBS) Reading Comprehension Test. Two groups of sixth grade students were studied: 129 students with learning disabilities (SWLD) and 397 students without disabilities. The students without disabilities came from two different school districts and were different with respect to overall achievement. Although an experimental design was planned, administration problems led to nonrandom assignment of students to conditions and some loss of student test score data. Scores under both standard time and extended time conditions were available for just under half of the SWLD. For the SWLD, only their scores under the condition of extended time were available. For the students without disabilities scores were available under both standard and extended time conditions.

Given these data, Huesman and Frisbie (2000) found that SWLD had larger gains on the ITBS Reading Comprehension Test with extended-time than students without disabilities. SWLD improved their average grade equivalent (GE) score from 4.60 to 5.21 (a gain of .61). The gains for students without disabilities were broken down by school district. In one district, the students improved their mean GE from 6.24 to 6.62 (a gain of .38); in the other district, their mean GE improved from 8.30 to 8.39. Although these findings support the interaction hypothesis, the large differences noted across the student groups leaves open the possibility of a regression-toward-the mean effect for the SWLD. Nevertheless, the authors concluded that extended time appears to promote test score validity for LD students.

Zurcher and Bryant (2001) investigated the effects of student-specific accommodations on the Miller Analogies Test, all of which involved extended time. They selected 15 SWLD

across three different colleges and 15 students without disabilities across these same colleges. The test was split into two halves and each student took one half without an accommodation and the other with an accommodation. All of the accommodations involved extended time and two accommodations also included oral presentation (one of which also used a scribe). Their results indicated no significant improvement for either group under the accommodation condition (scores across conditions differed by .06 of a point for SWLD and by .14 for students without disabilities). Thus, this study did not support the interaction hypothesis. However, there were several methodological limitations of this study. Aside from the small sample sizes, the "potency" of the accommodation effect may not have been sufficiently captured by the short half-tests. Furthermore, the student groups were not randomly selected or matched, which makes across group comparisons difficult. For example, there was a large difference in grade point average between the students without disabilities (GPA of 3.27) and the SWLD (GPA of 2.72).

Non-experimental Studies on the Effects of Extended Time

It is not surprising that a great deal of non-experimental research has been conducted to evaluate the effects of test accommodations on the test performance of SWD and ELL. In contrast to experimental studies that require manipulating the accommodation condition, randomly assigning examinees to conditions, and testing examinees at least once, non-experimental studies typically involve analysis of already existing data. Although these studies cannot implicate the accommodation as the cause of any performance difference, they do provide useful information for evaluating the effects of accommodations. As Koretz and Hamilton (2001) described

> …in the absence of experimental data that could isolate the effects of accommodations from the effects of student characteristics…simple data on the performance of students with disabilities can provide clues to the quality of measurement. For example, they can indicate the level of test difficulty for students with disabilities and can identify patterns

> in performance—such as mean differences between students with and without disabilities, differences in performance associated with accommodations, and anomalous item-level performance…that suggest hypotheses and point to needed additional investigation. (p. 15)

In this section, we review those studies that focused primarily on extended time. Many of these studies were conducted within the context of postsecondary admissions tests.

Camara, Copeland, and Rothchild, (1998) investigated score gains for students with learning disabilities (SWLD) and students without disabilities who took the SAT once in their junior year of high school and once in their senior year. Their design focused on three specific groups of students: SWLD who took the test once under standard conditions and once with extended time, SWLD who took the test twice with extended time, and students without disabilities who took the test twice under standard conditions. These groups of students were selected from the population of students taking the SAT between March 1994 and December 1995 (about 2 million students). The final samples of students included about 700,000 students without disabilities who took the test twice under standard conditions and about 9,000 SWLD who took the test twice. Within the SWLD group, about half took the test with extended time on both occasions, about 42% took the test first under standard conditions and second with extended time, and about 9% took the test first with extended time and second under standard conditions.

Camara et al. (1998) found that score gains for SWLD taking test with extended time were three times larger than score gains for students without disabilities who took the test twice under standard conditions. For SWLD who took the SAT first with standard time and second with extended time, the gain under the extended time accommodation was 38.1 points on the math test and 44.6 points on the verbal test. For students without disabilities who took the test twice under standard conditions the gain from first testing to second testing was 11.8 on math and 12.9 on verbal. SWLD who took the test under extended time first and standard time second

did worse, on average, on their second testing (i.e., under standard time). The loss for these students under standard time conditions was 6.1 points for math and 8.6 points for verbal.

Interestingly, Bridgeman, Trapani, and Curley (in press) found that the expected score gains for students without disabilities who took the SAT with extended time (specifically time-and-a-half) were about 10 and 20 points for the verbal and math tests, respectively. The relatively larger score gains for SWLD noted in the Camara et al. study taken together with the estimated gains reported in Bridgeman et al. provides some support for the interaction hypothesis.

Camara et al. also looked at how much extended time SWLD actually used in taking the test and the relationship between amount of time and standard-to-extended time score gain. They found a positive relationship between amount of extended time and score gain; that is, the more extended time given to a SWLD, the greater their score improvement.

Ziomeck and Andrews (1998) conducted a similar study on the ACT assessment. Across a three-year period between 1992 and 1995 52,667 students took the ACT with extended time. Of these students, 7,288 tested twice with at least one extended time administration. Ziomeck and Andrews looked at these SWD who took the ACT either (a) twice with extended time, (b) first with standard time and second with extended time, or (c) first with extended time and then with standard time.

Their results indicated that the group of SWD who took the ACT first under standard conditions and second with extended time had the largest score gains—3.2 scale score points, which was more than four times the score gain noted for the general population of students who take the test twice (.7 points). SWD who took the test twice with extended time exhibited a score gain of .9 scale score points. Similar to Camara et al., the group of SWD who took the test

with extended time first and standard time second performed worse, on average, upon retesting. Their mean score in the standard time condition was .6 scale score points lower than their initial mean score obtained under standard timing. To interpret the magnitude of these results it should be borne in mind that the ACT score scale ranges from 1-36 with a standard deviation of 6. The standard error of measurement on this scale is 1 point. Thus, the score gains for the group of SWD who took the test under standard conditions first and extended time second, are large. In interpreting these findings, Ziomeck and Andrews cited Mehrens (1997) who stated "After years of research, the profession has insufficient evidence to conclude the scores given under non-standard administrations mean the same thing as scores obtained under standard administrations" (cited in Ziomeck & Andrews, p. 7).

It is hard to determine the degree to which studies like Camara et al. (1998) and Ziomeck and Andrews (1998) support the interaction hypothesis. On the one hand, they demonstrate the consistent and expected finding that SWD improve their scores given extended time. On the other hand, they do not determine whether students without disabilities would also obtain better scores given extended time. As Camara et al. described

> A major problem with any analysis of the effects of accommodations for disabled examinees, such as the effects of extended time, is the difficulty in disaggregating the extent the modification compensates for the disability from the extent that it may overcompensate and introduce construct irrelevant variance…into the score (p. 13).

One potential explanation why extended time would help all students, not just those with disabilities is that these tests are speeded, at least to some extent. Again excerpting from Camara et al.

> One argument is that extended-time administrations provide a more precise estimate of students' abilities than standard-time administrations, and differences in score change are more related to the effect of speed on regular administrations, rather than providing learning disabled students with any advantage due to extended time. Such an argument would suggest that nondisabled students are slightly disadvantaged because of the time

> limits, rather than that learning disabled students are advantaged with substantial amounts of extended time. (p. 14)

Research that is not directly related to the effects of extended time on test performance, but is somewhat relevant, are studies that compared the predictive validity of scores from admissions tests taken with and without extended time. Cahalan, Mandinach, and Camara (2002) investigated the differential predictive validity of the SAT across students who took the test with and without extended time. In general, they found that (a) the predictive validity coefficients were significantly lower for SWD who took the test with extended time than for students who took the test under standard time conditions, (b) the SAT scores for students with learning disabilities who requested extended time tended to over-predict their first year college grades, and (c) this over-prediction was greatly reduced when SAT scores and high school grades were considered together. Furthermore, when looking at these results across the sexes they found that SAT scores for female SWD *under*predicted first year college grades when they were considered together with high school grades. The standardized residual (from predicting first year college GPA from the SAT I) for female students with learning disabilities was .02 (overpredicted), but for males, the residual was .21 (overpredicted).

The findings of Cahalan et al. (which are similar to Braun, Ragosta, & Kaplan, 1986—reviewed later) provide credible evidence that, from the perspective of predictive validity, in general, scores from students with learning disabilities who take the test with the accommodation of extended time are not comparable to scores from students who take the test under standard time conditions. However, a few caveats must be raised regarding the conclusion of non-comparability. First, the authors were unable to match course-taking patterns across the accommodated and non-accommodated student groups. Thus the degree to which differences in

courses taken affected first year college grades is unknown. Second, in the Cahalan et al. study, the general finding that SAT scores over-predict first year grades did not hold for women. Third, the conclusions of over-prediction and non-comparability are based on group statistics. At the individual level, many learning disabled students who took the SAT with extended time had their college grades under-predicted. Therefore, the extrapolation of this finding to all SAT scores taken under the accommodation of extended time should be made cautiously. For these reasons, research on the differential predictive validity of scores from accommodated and standard test administrations is considered irrelevant for evaluating the effects of test accommodations on students' performance. Therefore, the many other studies in this area (e.g., Wightman, 1993) are not reviewed here.

Oral Administration Accommodations for SWD

The category of oral accommodations (e.g., read aloud protocols) usually includes adjustments to how test takers are presented with either the test directions or items when they appear in written form. Usually, the oral presentation is a verbatim translation of the directions and items. Typically, a test administrator (or computer, video, or audio tape) reads the relevant portions of the test for the student. For test directions, an oral presentation may take the form of paraphrasing or restating the directions in test taker "friendly" form. Oral presentations are typically not allowed on reading tests, or other tests where the ability to read, per se, is part of the construct of interest.

Experimental Studies on Oral Accommodations

Meloy, Deville, and Frisbie (2000) examined the effects of a read aloud accommodation on the test performance of middle school students with a reading learning disability (LD-R) and students without a disability. The tests involved in the study were the ITBS achievement tests in

Science, Usage and Expression, Math Problem-Solving and Data Interpretation, and Reading Comprehension. All tests were given on level and the read aloud accommodations were conducted by one of the authors using a script carefully designed for each test at each grade level.

A total of 260 students from two middle schools in a Midwestern school district participated, including 98 sixth graders, 84 seven graders, and 78 eighth graders. Of these students, 198 did not have a disability and 68 students had a reading disability. Students were randomly assigned to one of the two test administration conditions (read aloud or standard). To permit comparisons across subject areas, each student was administered all four tests and remained in the same condition for each.

The results of the study indicated that, on average, the LD-R students scored significantly higher under the read aloud accommodation. However, this finding held for the students without disabilities, too. Although the score gain under the read aloud condition for LD-R students (about .75 standard deviations) was larger than the gain for students without a disability (about .50 standard deviations), the interaction was not statistically significant. The only statistically significant findings were the main effects: both groups scored higher under the accommodation condition and the students without disabilities outperformed the LD-R students. These results do not support the interaction hypothesis, which led Meloy et al. to conclude that general use of the read aloud accommodation for LD students taking standardized achievement tests is not recommended.

McKevitt and Elliot (in press) conducted an experimental study where groups of students with and without disabilities took a standardized reading test (TerraNova Multiple Assessments Reading Test) twice—once under standard administration conditions and once with an oral

accommodation (audiocassette version of test content). The study involved 79 eighth-graders, 40 of whom were classified as having an educationally defined disability and were receiving services in reading/language arts, and 39 general education students. They found no statistically significant differences for the accommodation condition. Neither group of students performed better with the accommodation and the students without disabilities outperformed SWD in both conditions (i.e., main effect for student type, no interaction). The results of this study do not support the interaction hypothesis.

McKevitt and Elliot also asked 48 teachers what accommodations they thought were valid for specific students. The teachers selected extra time most frequently, with "reading the directions" next. However, no teacher selected "reading the test content aloud" as an accommodation and felt this accommodation was somewhat invalid. However, the majority of SWD (42.5%) reported they liked taking the test better with the accommodation and 40% of SWD reported they it was easier to show what they knew when given accommodations.

Johnson (2000) also used a 2X2 (accommodation-by-student group) experimental design to look at the effectiveness of an oral accommodation for helping SWD. He used data from the 1997 and 1998 Washington Assessment of Student Learning (WASL) math tests. Groups of students with and without reading disabilities were tested under conditions where the math items were read to them and where they read the math items themselves. The WASL math test included both multiple-choice and constructed-response item formats.

The study involved the participation 115 fourth grade students. Seventy-seven of these students did not have a disability; the other 38 students had a reading disability. Students were matched across SWD and non-SWD groups for ethnicity, gender and language as closely as possible. The SWD initially took the test with an oral administration and the students without a

disability took the test under standard conditions. The students without disabilities were then randomly assigned to a post-test condition—oral accommodation or standard administration. The SWD were post-tested under standard conditions. The post-tests occurred within 6 to 21 days of the initial testing.

Johnson found that the SWD scored considerably lower than the two general education groups. However, their score gains from standard to accommodation condition (gain of 19.64 points) was much larger than the gains noted for the control group (7.61 points) and the group that retested under standard conditions (2.29 points). The interaction of student group-by-test administration condition approached statistical significance ($p=.076$). Thus, results provided partial support for the interaction hypothesis.

Weston (2002) was also interested in evaluating the interaction hypothesis with respect to the accommodation of oral administration. He tested two groups of fourth grade students: 65 SWLD and 54 students without disabilities. Both groups of students took two parallel math test forms from the National Assessment of Educational Progress (NAEP). One test form was taken under standard conditions, the other with oral presentation. Although both student groups exhibited gains in the accommodation condition, the SWLD had significantly greater gain. The effect size for SWLD was .64. For students without disabilities, the effect size was .31. Although students without disabilities also benefited from the accommodation, the significantly larger gain for SWLD lends some support to the interaction hypothesis.

Students in the Weston study also took a reading test. In looking at score gains across students who differed with respect to reading proficiency, he found that students with better reading proficiency gained less. He also found that accommodations improved performance on word problems more than on calculation problems. These findings led Weston to conclude

"…differential easiness due to unexplained factors may affect a minority of students' scores, but as a group the accommodated test is a better representation of student ability than the non-accommodated test" (p. 21).

Weston also interviewed 19 teachers about the assessment. All of the teachers he interviewed spoke very favorably of the accommodation for SWLD, but 8 of the teachers felt that students without disabilities would be frustrated by the accommodation.

Tindal, Heath, Hollenbeck, Almond, and Harniss (1998) also used an experimental design to investigate the effects of oral accommodation and response format on the test scores of SWD. The specific response format investigated was allowing students to write their answers into the test booklet rather than on an answer sheet. The oral accommodation was investigated on a fourth grade statewide math test; the response format accommodation was investigated on the same math test and a statewide reading test.

The study involved 481 fourth grade students, 84% of whom were students without disabilities. There were 36 SWD who took the reading test and 38 SWD who took the math test. For the analysis of response format accommodation, all students participated in both conditions. Each student took one test (either reading or math) with an answer sheet and wrote their answers to the other test directly into the booklet. For the oral accommodation, 122 students without disabilities and 42 SWD were randomly assigned to the standard or oral presentation conditions. The results showed no effect for the response format condition. Also, students without disabilities outperformed SWD under both standard and oral accommodation conditions. However, there was a significant improvement in scores for SWD under the oral accommodation condition for SWD (effect size of .76), but not for the other student group (negative effect size of

.20). This finding supported the interaction hypothesis and led Tindal et al. to conclude "More valid inferences of math performance were possible when [SWD] had test read to them" (p. 447).

Non-Experimental Studies on Oral Accommodations

Kosciolek and Ysseldyke (2000) examined the effects of a read aloud accommodation using a quasi-experimental design on a small number of students in third through fifth grade in a suburban school district. Seventeen general education students and 14 special education students participated in the study. Efforts were made to keep the groups as comparable as possible in terms of demographic characteristics, but the students were not randomly selected. Also, due to the limited number of students willing to participate, the special education group was comprised mostly of males. All students were given two equivalent forms of the *California Achievement Tests (CAT/5)*, *Comprehension Survey*. One form was administered with the accommodation; the other was administered under standard conditions. To minimize error due to practice, the order of the accommodation was counterbalanced.

The read aloud accommodation was provided using a standard audiocassette player to maintain consistency between testing sessions. Two open-ended questions were asked at the end of the testing session to get an idea of student perception of and comfort level with the read aloud accommodation. A repeated-measure analysis of variance was conducted to determine whether there was an interaction between the test administration condition and disability status on students' test performance.

Consistent with previous research, students without disabilities outperformed SWD under both test administration conditions. However, the gain for SWD in the accommodation condition was much larger. In the standard condition, SWD obtained a mean score of 661.4; in the oral accommodation condition, they achieved a mean of 691.6. Although this gain only

approached statistical significance (p=.06) it represented a large effect size (.56). For students without disabilities, the mean test score under the standard condition was 744.6, and under the accommodation condition it was 749.8. The effect size associated with this gain was negligible (.10). Kosciolek and Ysseldyke also noted that SWD embraced the accommodation, while the students without disabilities preferred the standard administration. Although the study involved small samples, the results support the interaction hypothesis.

Helwig and Tindal (2003) investigated the degree to which teachers can predict which types of students would benefit most from specific accommodations. They asked teachers to rate students on a 5-point Likert Scale that judged each student's proficiency in reading and math and how important the accommodation would be to the student's success on the test. Then, they tested students in both standard and oral accommodation conditions. A large national sample of students participated in the study—about 1,200 students, 245 of whom were SWD. The results of the study did not match the authors' expectations. Not only were teachers unable to predict which students would benefit from the accommodation, in most of the comparisons, students with and without disabilities performed better in the standard test administration condition.

Computer-Assisted Oral Presentation

The computer is revolutionizing assessment in several ways and many researchers have explored the different ways in which computerized accommodations may lead to more valid measurement for SWD. Two of these studies are essentially oral accommodations where the computer "reads aloud" the test content that appears on the screen.

Calhoon, Fuchs, and Hamlett (2000) compared the effects of a read-aloud accommodation on a math test that was provided by either a teacher or a computer. The math test involved constructed-response items. Their study involved four conditions: standard

administration, teacher-read administration, computer-read administration, and computer-read with video. Eighty-one ninth through twelfth grade SWLD participated in the study and were tested in each of the four conditions over a four-week period (the conditions were counterbalanced). The results indicated that all read-aloud accommodations led to higher scores for these students compared with the standard administration (effect sizes ranged from about one-quarter to one-third of a standard deviation). However, there were no significant differences among the read-aloud methods. That is, the computer-read administration did not lead to score improvements relative to the teacher-read administration. However, about two-thirds of the students preferred the anonymity provided by the computer when taking the test.

Brown and Augustine (2001) also evaluated whether screen-reading software would provide a helpful accommodation to students with reading difficulties. They developed two parallel social studies test forms and two parallel science test forms from publicly released NAEP items. They administered the forms to students both with and without reading disabilities from high schools in Delaware and Pennsylvania—96 students completed science assessment and 110 students completed the social science assessment. Within each subject area, students took the test forms under standard and computer-read conditions (the screen reading software was called Authorware 5.0). After controlling for students' reading proficiency, there was no difference in student performance under standard and screen-reading conditions. However, although the authors stated both students with and without "reading difficulties" were included in the study, they did not provide information on the numbers of such students and they did not compare the effect of the accommodation across student types.

Multiple Accommodations for SWD

The literature on test accommodations includes studies that focus on a single accommodation and studies that focus on all accommodations that were allowed on a particular test. Most of the studies that focused on multiple accommodations were ex post facto studies that analyzed data from a large-scale assessment and broke out accommodated test administrations from non-accommodated administrations. These studies represent real-world conditions where many students are given two or more types of accommodations. The drawback to these studies is that they typically do not use an experimental design to investigate the effects of the accommodations on test performance.

A smaller number of studies have looked at multiple accommodations using experimental or quasi-experimental research designs. We begin our review with these studies.

Experimental and Quasi-experimental Studies on Multiple Accommodations

Elliot, Kratochwill, and McKevitt (2001) used an alternating treatment design (ATD) to evaluate the effects of several types of accommodations on the test scores of students with and without disabilities. The study involved 100 students, 41 of whom were SWD. The students were compared on 8 mathematics and science performance items relevant to the state of Wisconsin's content standards. The items were sampled from domains of knowledge defined by the National Council for Teachers of Mathematics and the National Science Standards. Trained judges scored all responses using 5-point scoring criteria scale that ranged from (1) *inadequate* to (5) *exemplary*.

Elliot et al. provided accommodations that were administered individually to students rather than using a "prepackaged" approach where every student was allowed the same accommodation, regardless of whether the accommodation was appropriate for them or not. The

study therefore relied on a "single case" analysis rather than a group design. Accommodations allowed were those that were identified by teachers and documented on the student's Individual Education Plans. Accommodations used were: verbal encouragement, extra time, individual test administration, read directions to student, read subtask directions, paraphrase directions, restate directions or vocabulary, read questions and content, restate questions, spelling assistance, mark task book to maintain place, and use of manipulatives. Students without disabilities were randomly assigned to one of three comparison or control conditions. These were (a) 25 students in a no accommodation condition, (b) 20 students in a standard accommodation condition, and (c) 14 students in a teacher-recommended accommodation condition.

ATD involves *within*-student analyses, where an effect size statistic is computed by comparing the accommodated condition with the no accommodation condition for each student. Group average effect sizes are also computed. The results indicated that as a group, SWD scored nearly 1 SD lower in the no accommodation condition relative to their performance in the accommodation condition (effect size of .83). In the accommodation condition, SWD demonstrated comparable or better scores than students without disabilities in the non-accommodation condition on 4 of the 8 items. Other results indicate similarities between groups on complex math and science items when SWD were allowed an accommodation and both groups found the performance tasks difficult.

Overall, based on effect size criteria, accommodations had a significantly higher impact on SWD (63.7%) than on students without disabilities receiving teacher-recommended accommodations (42.9%) or standard accommodations (20%). Testing accommodations had a medium to large effect on more than 75% of SWD, but also had a similar effect on more than

55% of the students without disabilities. Negative effects of accommodations were noted for 17% of SWD and 7% of students without disabilities.

The results of this study provide some support for the interaction hypothesis in that SWD exhibited greater gains than students without disabilities. However, the fact that many students with disabilities performed better under the accommodation condition suggests that this hypothesis may need modification. A positive aspect of this study is that it reflected the reality of how accommodations can influence test performance of SWD. Also, it is also one of only a few studies that investigated performance on constructed-response items.

Fuchs, Fuchs, Eaton, Hamlett, and Karns (2000) evaluated the performance of students with learning disabilities (SWLD) and students without disabilities on mathematics test performance using a variety of accommodations. The utility of teacher judgments in specifying appropriate accommodations was also investigated. Curriculum based assessment (CBM) techniques were used to supplement teacher decisions when deciding which accommodations to use. The basic research questions were concerned with a "differential boost" (greater than expected) demonstrated by SWLD when using accommodations over the "boost" (typical) demonstrated by students with out disabilities. Accommodations used were extended time, calculators, reading text aloud, and transcription.

Participants in the study were 192 SWLD in grades 4 and 5 and 181 students without disabilities in grade 4. The sample was selected to represent the school demographics for gender, race and ethnicity. Each student was administered a CBM probe in math concepts, computation, application and problem solving. Criterion validity for the probes was established at .74, using the CTBS for comparison. Data were collected in phases where in phase 1, CBM probes were administered to all students using accommodations to determine who should receive

accommodations on the commercial tests (ITBS and Stanford Achievement Tests). In phase II, an accommodation "boost" was calculated for each student without a disability by subtracting their score on the non-accommodated administration of the test from their score under the accommodation condition, and the standard deviation of that boost. For each accommodation, the standard deviation was added to the mean and corrected for regression. To receive an accommodation in the next phase, SWLD had to demonstrate an individual boost that exceeded the accommodation criterion. Based on this determination, each SWLD was assigned to either a standard or accommodation condition. In phase III, special education teachers completed questionnaires for each SWLD judging which, if any, accommodations are needed when testing. The commercial tests were administered in Phase IV.

Using CBM data, SWLD did not benefit more than students without a disability from extended time, using a calculator, or having the instructions read to them. On problem solving, SWLD profited more using accommodations than did the other students. Poor correspondence was found between teacher-identified students who needed accommodation, and those who did not, and their eventual performance. Teachers over-identified the need for accommodations for SWLD. While teacher judgments may not assist in identifying which students may profit from an accommodation, data based sources can supplement decisions. Although this study does not support the interaction hypothesis, it provides invaluable insight regarding how an accommodation may boost a test score for a student with a disability. It also provides a mechanism for determining what that "boost" might look like for a non-disabled student and uses that criterion as a baseline for determining an extended boost for SWLD.

Similar to Fuchs, Fuchs, Eaton, Hamlett, and Karns (2000), Fuchs, Fuchs, Eaton, Hamlett, Binkley, and Crouch (2000) evaluated the performance of SWLD and non-disabled

students on a reading subtest of the ITBS under both accommodated and non-accommodated conditions. They tested 181 SWLD in grades 4 and 5 and 184 students without disabilities in grade 4. Students completed four, brief assessments in reading using 400 word passages. And answered eight multiple-choice questions (six literal; two inferential). Three passages were used for each of the conditions of (1) standard, (2) extended time, (3) large print, and (4) student reads aloud. Selected teachers completed questionnaires about whether a student should complete the ITBS under standard or accommodated conditions.

For extended time and large print accommodations, SWLD did not benefit more than students without disabilities. Both groups benefited to a similar degree, which suggests that extended time may inflate test scores of SWLD. Reading aloud, however, proved beneficial to SWLD, but not to the non-disabled students. However, reading aloud was the only accommodation administered individually, and thus the individual administration may partly account for this effect. Similar to the previous study, teachers were a poor source of decision making in that they recommended many more accommodations than were necessary, as indicated by the data-based approach.

McKevitt, Marquart, Mroch, Schulte, Elliott, and Kratochwill (2000) looked at the effects of a variety of accommodations on the performance of 58 SWD and 20 students without disabilities on fourth grade math and science items from the Wisconsin Student Assessment System. The main objectives of the study were to (a) document and describe the test accommodations listed on student IEPs, (b) document accommodations actually used during testing, and (c) examine the effects accommodations have on test score for students with and without disabilities. All test items were math and science performance tasks. Items were presented in four 1-hour sessions and were scored by raters using a five-point scale ranging from

1 (inadequate) to 5 (exemplary). Four math tasks and 4 science tasks were used. Accommodations were selected from students' IEPs and from the *Assessment Accommodations Checklist* completed by teachers. Extra time, reading and paraphrasing test directions and verbal encouragement were the most common accommodations provided. Students typically received "packages" of accommodations (majority receiving 10-12 accommodations), that is, more than one allowed at different times.

An ATD (single-subject) design was used in which all students were subjected to two different testing conditions (i.e., receiving and not receiving accommodations) until all 8 tasks were completed. Starting condition and starting task were randomly assigned for both intra-individual and inter-group comparisons without the need for a baseline phase. Effect size differences between accommodated and non-accommodated conditions were computed.

The accommodations had a medium to large positive effect for 47 (81%) of the SWD and 26 (51%) of the students without disabilities. Small, or zero effects were revealed for 3 (5%) of the SWD and 21 (41.2%) of the students without disabilities. Negative effects were found for 8 (14%) of the SWD and 4 (7.8%) of the students without disabilities. On average, accommodated test scores compared to non-accommodated scores yielded a .94 effect size for SWD, .44 effect size for students without disabilities who received the "standard' (only one) accommodation, and .55 for students without disabilities who received teacher-recommended accommodations. Once again, the improvement of students without disabilities in the accommodation condition undermined the validity of the interaction hypothesis.

Non-experimental Research Involving Multiple Accommodations

Most of the non-experimental research on test accommodations has focused on data from large-scale tests, such as statewide educational assessments or postsecondary admissions tests. In

this section, we first review those studies pertaining to K-12 assessments. Next we review those related to postsecondary admissions tests.

Kentucky Instructional Results Information System

The Kentucky Instructional Results Information System (KIRIS) was one of the first state-mandated assessments to investigate the effects of test accommodations on the test performance of SWD. Two studies looked at the performance of students taking the KIRIS tests in some depth.

Trimble (1998) documented the performance of grade 4, 8, and 11 SWD on KIRIS tests from 1992-1995. During these years, the KIRIS assessments consisted solely of constructed response items and portfolio assessments. A variety of assessment accommodations were given to SWD including reader/oral, scribe/dictation, cueing, paraphrasing, interpreter, and technological. The most frequently used accommodations were combinations that included paraphrasing and oral presentations. Trimble recorded the scale scores for each of the content areas for (a) the total population of students, (b) students representing various accommodation or combination of accommodation conditions used by at least 100 students, and (c) SWD who used no accommodations during the two testing years.

Trimble (1998) found that, as a cohort, SWD who took the test with one or more accommodations showed greater gains across years than SWD who took the test without accommodations or students without disabilities. Although the design of this study does not allow cause for the improvement to be linked with the provision of an accommodation, this study shows one way in which it may be useful to track the intended benefits of test accommodations, particularly when their implementation in a statewide testing program is new.

Koretz and Hamilton (2000) examined KIRIS data from 1995 to 1997 to discover how SWD were included in these tests and to evaluate the effects of the accommodations test performance. They looked at differences between SWD and students without disabilities on total test scores as well as on scores based on multiple-choice and open-response items. They found that the most frequent accommodations provided were extra time, using scribes, oral presentation, paraphrasing, technology, interpreters, and separate sessions, with many students receiving more than one accommodation. Students were assessed in math, science, and language arts.

Koretz and Hamilton (2000) found that the performance of SWD was generally lower than students without disabilities with the differences increasing with grade level. While some differences were noted in the early grades, no overall strong or consistent relationship was found between test item formats for SWD. In most instances the discrepancy between groups grew for both item formats as students progressed through the grades. Performance of SWD varied substantially according to their particular disability but small numbers of students in each category made it difficult to interpret these findings.

Frequent substantially higher associations between the use of accommodations and test scores were reported, but this finding was not consistent and at times yielded "implausible" scores. For example, accommodations given to students with mental retardation boosted their scores above their non-disabled peers. Given the cognitive delays exhibited by these students and their engagement in the general curriculum, is seems unlikely that they should earn higher scores than students without disabilities.

New York State English Regents Exam

Similar to their study on KIRIS, Koretz and Hamilton (2001) evaluated the performance of SWD in large field-test of the New York State English Regents Examination. Specifically, they looked at overall performance, completion rates, and item performance by type of accommodation for about 8,700 high school seniors, 481 of whom were SWD (about 80% of the SWD were students with learning disabilities). They found that on this pilot test, 77% of the SWD were accommodated. The most common accommodation was extra time plus separate room, which was given to over half of the students who received an accommodation.

Overall, Koretz and Hamilton (2001) found that SWD performed about three-quarters of a standard deviation lower than students without disabilities. Although the completion rates for SWD and students without disabilities were comparable, they found that SWD who took the tests *without* accommodations had lower completion rates on the constructed-response items.

An interesting feature of the Koretz and Hamilton (2001) study was their comparison of performance differences across SWD who received accommodations and SWD who did not. Overall, SWD who did not receive an accommodation performed much better than SWD who received an accommodation. This result is not necessarily surprising because SWD who do not need accommodations may be less impaired that SWD who are most in need of one. However, when comparing the performance differences between these two groups, Koretz and Hamilton noted that the gap was much smaller on the constructed-response items than on the multiple-choice (MC) items. For example, the standardized mean difference on the MC items was .30 across SWD who received an accommodation and SWD who did not. This difference dropped to .10 on the constructed-response items. With respect to comparing SWD who did not have an accommodation with SWD who had only extended time, they concluded “the additional benefit

of the accommodations for [constructed-response item] performance was .23 standard deviation (p. 19). Furthermore, they concluded "Across all accommodation conditions, additional time was associated with an increase of .13 standard deviation on the [constructed-response] portion of the test, but essentially no change on the MC portion" (p. 19). These findings led them to conclude "Extended time appears to give more of a boost to performance on [constructed-response] items than on MC items" (p. 19).

Another interesting observation by Koretz and Hamilton was that SWD had more blank or unscorable responses than students without disabilities. Furthermore, on the constructed-response items, they noted that SWD had far more responses in the lowest score category (minimal or no evidence of understanding).

Given the multiple test forms used in the pilot study, Koretz and Hamilton (2001) were also able to investigate the effects of test length on students' performance. They found that the gap in performance between SWD and students without disabilities increased with test length.

SAT

In the mid-1980s, the College Board, Educational Testing Service (ETS) and the Graduate Record Examinations Board conducted a series of studies on the SAT and GRE to investigate the performance of SWD on these tests and appraise the comparability of test scores from accommodated and standard administrations. Although these studies were not experimental forays into the effects of test accommodations on test performance, they provide unique information regarding the utility of test accommodations. Many of these studies were summarized in the book *Testing Handicapped People* (Willingham et al., 1988), and many were published earlier as College Board or ETS research reports.

In one report, Braun, Ragosta, and Kaplan (1986) looked at the predictive validity of SAT scores across accommodated and standard administration conditions. They computed college grade-point-average (GPA) prediction equations using SAT scores and other preadmission criteria such as high school grades. A key aspect of their methodology was computing the prediction equations using data from students who took the test under standard conditions and then using these equations to predict the GPAs for students who took the test with one or more accommodations. Comparisons were made across accommodated and standard groups with respect to the general degrees to which their scores were over- or under-predicted.

The Braun et al. (1986) study involved 1,000 SWD who took the SAT with an accommodation and 650 SWD who it without an accommodation. Data were obtained from 145 schools that admitted sufficient numbers of students who took the test with an accommodation. Fifty students who did not have a disability were randomly selected at each school to compute separate within-school regression analyses. In general they concluded that, from a predictive validity perspective, test scores from accommodated administrations of the test were comparable to scores from standard administrations for visually impaired and physically handicapped students. For students with learning disabilities, using SAT scores from accommodated administrations alone led to over-prediction of their GPA. However, this degree of over-prediction was mediated when high school GPA was added to the prediction equation. The authors speculated that this mediation was caused by high school GPA under-predicting college GPA for these students.

Multiple-day Accommodation for SWD

Although extended time and oral presentation are the most popular accommodations given and studied, there are many other test accommodations that are given to SWD.

Unfortunately, many of these accommodations have not yet been empirically studied. One exception is a study by Walz, Albus, Thompson, and Thurlow (2000) that looked at a "multiple-day" accommodation.

A multiple-day accommodation splits up a test administration that is typically administered in one day over multiple days. Walz et al. (2000) evaluated this accommodation using a sample of 112 seventh and eighth graders from two rural and two urban schools in Minnesota. Forty-eight of these students were SWD; the other 64 were general education students. The test items came from a statewide test in Minnesota. All students took two different forms of the test. One form was taken in a single-day administration; the other form was administered over a two-day period. The students without disabilities outperformed the SWD under both conditions. Furthermore, neither student group exhibited meaningful gains under the multiple-day condition. The SWD group exhibited a gain of 0.7 points and the general education group exhibited a gain of 2.08 points. Thus, the results did not support the use of a multiple-day accommodation for improving the scores of SWD.

Summary of Accommodations Research on Students With Disabilities

At this juncture, 26 studies pertaining to test accommodations for SWD have been reviewed. These studies looked at a variety of accommodations, the most common being extended time and oral presentation. Many of the studies specifically investigated the performance of students with learning disabilities. Brief summaries of the reviewed studies are presented in Tables 4 and 5. Table 4 summarizes 14 experimental studies and Table 5 summarizes 12 quasi- and non-experimental studies.

Table 4

Summary of Experimental Studies on SWD

Study	Subgroups	Accommodations	Design	Results	Supports H_1?
McKevitt, Marquart, et al. (2000)	SWD	Extra time, oral, encouragement, "packages"	SS-ATD	Greater gains for SWD	Yes
Elliot, Kratochwill, & McKevitt, (2001)	SWD	encouragement, extra time, individual admin., various oral, spelling assist., mark to maintain place, manipulatives	SS-ATD	Moderate to large improvement for SWD.	Yes
Runyan (1991)	SWD-LD	Extra time	b/w groups	Greater gains for SWD	Yes
Zuriff (2000)	SWLD	Extra Time	5 diff. Studies	Gains for both SWD and non-SWD	No
Fuchs, Fuchs, Eaton, Hamlett, Binkley, & Crouch (2000)	SWD-LD	Extra time, large print, student reads aloud	b/w groups	Read aloud benefited SWLD but not others	Yes
Weston (2002)	SWD	Oral	w/in & b/w group	Greater gains for SWD	Yes
Tindal, Heath, et al. (1998)	SWD	Oral	w/in & b/w group	Sig. gain for SWD only	Yes
Johnson (2000)	SWD	Oral	b/w groups	Greater gains for SWD	Partial
Kosciolek & Ysseldyke (2000)	SWD	Oral	w/in & b/w group	No gain	No
Meloy, Deville, & Frisbie. (2000)	SWD	Oral	w/in & b/w group	Similar gains for SWD and non-SWD	No
Brown & Augustine (2001)	SWD	Screen reading	w/in & b/w group	No gain	No
Tindal, Anderson, Helwig, Miller, & Glasgow (1998)	SWD	Simplified English	Unclear	No gain	No
Fuchs, Fuchs, Eaton, Hamlett, & Karns, (2000)	SWD-LD	Calculators, extra time, reading aloud, transcription, teacher selected	b/w groups	Differential benefit on CR items	Partial
Walz, Albus, et al. (2000)	SWD	Multi-day/session	w/in & b/w group	No gain for SWD students	No

SS-ATD=Single- subject alternating treatment design.

Table 5

Summary of Quasi and Non-Experimental Studies on SWD

Study	Subgroups	Accommodations	Design	Selected Findings
Cahalan, Mandinach, & Camara (2002)	SWLD	Extended time	Ex post facto	Predictive validity was lower for LD students, esp. males.
Camara, Copeland, & Rothchild, (1998)	SWLD	Extended time	Ex post facto	Score gains for LD retesters w/ extended time 3X greater than standard retesters.
Huesman, R.L. & Frisbie, D. (April 2000)	SWD	Extra time	Quasi-experimental	Score gains for LD but not for NLD groups.
Ziomeck & Andrews (1998)	SWD	Extra time	Ex post facto	Score gains for LD retesters w/ extended time 3X greater than standard retesters.
Schulte, Elliot, & Kratchowill (2001)	SWD	Extra time, oral	Ex-post facto	SWD improved more b/w non-accom. & accom. (medium effect size; .40 to .80.) than non-SWD (small effect size; less than .40). No differences on CR items
Braun, Ragosta, & Kaplan (1986)	SWD	Various	Ex post facto	Predictive validity was similar across acc. And non-accom. Tests, slightly lower for LD
Koretz & Hamilton (2000)	SWD	Various	Ex-post factor	SWD performed lower than non-SWD and differences increased w/ grade level. No consistent rel. found b/w test item formats for SWD
Koretz & Hamilton (2001)	SWD	Various	Ex post facto	Accommodations narrowed gap more on CR items.
Helwig, & Tindal (2003)	SWD	Oral	Ex-post facto	Teachers were not accurate in determining who would benefit from accommodation.
McKevitt & Elliot (in press).	SWD	Oral	Ex-post facto	No sig. effect size differences b/w accommodated and non-accommodated conditions for either group.
Johnson, Kimball, Brown, & Anderson (2001)	SWD	English, visual, & native language dictionaries, scribes, large print, Braille, oral	Ex-post-facto	All SWD scored lower than non-SWD. Accommodations did not result in an unfair advantage to special education students
Zurcher & Bryant (2001)	SWD	Not specific	Quasi-experimental	No significant gains

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. It is also clear that the interaction hypothesis, as it is typically described, is on shaky ground. Students without disabilities typically benefit from accommodations, particularly the accommodation of extended time. In our discussion section, we address the issues of general and specific effects for different types of accommodations in more detail. We also propose a revision of the interaction hypothesis. These discussions are temporarily forestalled so that we can first review the literature on the effects of test accommodations for English language learners.

Studies Focusing on English Language Learners

From a psychometric perspective, the issues involved in test accommodations for English language learners (ELL) are the same as those for SWD. The comparability of scores from accommodated and non-accommodated test administrations is an important validity issue regardless of the reason for the accommodation. Also, evaluations of the utility of the accommodations are the same across these different contexts with respect to the research designs and statistical methods that can be used. The interaction hypothesis remains the critical hypothesis to be tested. Thus, research in this area has also focused on the degree to which accommodations improve the scores of ELL relative to their English proficient peers.

Although many important studies have been conducted in this area, Abedi and his colleagues have been particularly prolific. In fact, Abedi's work with his colleagues at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) has produced the most comprehensive set of reports on specific accommodations for ELL. We

review several of these research products, which, in general, represent final or revised versions of earlier papers.

A conspicuous, but unfortunate, outcome of our review is that, relative to research on SWD, little research has been conducted on the *effects* of test accommodations on the test *performance* of ELL. Instead, many of the studies in this area have focused on the psychometric properties of the accommodated tests such as the adequacy of test translation, equivalence of factor structures across accommodated and original tests, and investigating differential item functioning (e.g., Allalouf, Hambleton, & Sireci, 1999; Robin, Sireci, & Hambleton, in press; Sireci & Khaliq, 2002). As with test accommodations for SWD, we consider this research irrelevant to analysis of the effects of the accommodation on student performance. Therefore, our review focuses on those studies that look at such effects.

Linguistic Modification, Dictionaries, Glossaries

Abedi (2001)[2] evaluated the utility of providing three specific accommodations hypothesized to have the potential to improve the performance of ELL on science assessments: (a) provision of a customized English dictionary, (b) providing a "linguistically modified[3]" version of the test, or (c) providing an English-to-Spanish glossary. The latter accommodation was for Spanish-speaking ELL students only. The subject area of science was chosen because "it is a content area where language ability can grossly interfere with the measurement of science achievement" (p. 3). The study involved testing ELL and non-ELL students from southern California in fourth and eighth grade. About 1,800 students were tested in fourth grade, and about 1,600 students were tested in eighth grade. In each grade, about half the students were

[2] An earlier version of this report, which was described as a pilot study, was published as Abedi, Courtney, Mirocha, & Goldberg, 2001.

[3] This accommodation involves making minor changes in the text associated with test items. It is also sometimes referred to as "modified English" and "simplified English."

ELL (about two-thirds of whom were Spanish-language dominant). The research design used was a between-subjects design, with students randomly assigned to conditions. Main effects for accommodation and student group were studied, as was the interaction. The analyses were done with and without using a reading score as a covariate.

Four forms of NEAP science items were created and administered to the participating students. One form contained science items with no accommodation, while the remaining three included the customized English dictionary, the English-to-Spanish glossary, or the linguistically modified test. An English-to-English glossary was used as the control test booklet for non-ELL students in the bilingual glossary condition. Extra time was given to all students.

The results reported in Abedi (2001) revealed that there were no accommodation effects (main effect or interaction effect) for the fourth grade students. ELL and non-ELL students performed similarly under all test administration conditions. The largest mean gain for an accommodation relative to the standard administration was .14 for the customized dictionary condition, which produced a negligible delta effect size of about .01. For non-ELL, the largest gain was .08 (for the same condition), which also produced a negligible effect size (also about .01). For the eighth graders, the accommodation of linguistic modification improved the mean score of the ELL group, relative to non-ELL. However, the effect of this improvement was small. The mean for ELL taking the linguistically modified version of the test was about 2 points higher than the mean for ELL taking the test under the standard condition (delta effect size about .20). Interestingly, the improvement was higher for non-Spanish speaking ELL than for Spanish-speaking ELL. For non-ELL, the mean score for the linguistically modified version was about a half-point higher, which yielded a negligible effect size (delta=.05). No meaningful gains were noted under any other conditions. Another interesting finding reported by Abedi

(2001) was that ELL students were more likely to report that there were words in the science test that they did not understand, and more ELL students looked up words in the glossary and dictionary conditions than non-ELL students.

Abedi, Hofstetter, Baker, and Lord (2001) investigated similar accommodations for ELL on a math tests. They administered a short eighth grade NAEP math test to 946 eighth grade students from urban districts in California under one of 5 conditions: no modification, use of simplified (modified) vocabulary, use of a glossary allowed, extra time, and extra time with glossary. Conditions were randomly assigned to students. LEP, non-LEP, and FEP (former LEP students who are now fully English proficient) were tested. Most LEP students were Spanish-speaking. All students were also given a reading test, which was used as a covariate.

For most students, improvements were made under all accommodations, particularly extra time and extra time with glossary. However, the gains for the LEP group were small. The authors concluded that the "modified English" accommodation was "the only accommodation type that narrowed the score difference between LEP and non-LEP students." However, this "narrowing" was due to the fact that the non-LEP students performed poorest on this version, not that the LEP group did much better than other conditions. Thus, the accommodations studied did not lead to score improvements for the targeted group they intended to help.

Abedi, Lord, Hofstetter, & Baker (2000) reported the same results as Abedi, Hofstetter et al. (2001), but they added some confirmatory factor analyses to evaluate the structural equivalence of the reading and math tests across ELL and non-ELL groups. Although they found similar factor loadings for parcels of items across the two groups, they noted the correlation between reading and math was higher for ELL. They also noted that the internal

consistency reliability was lower for the ELL group (α=.81) than for the non-ELL group (α=.87).

In another study, Abedi and Lord (2001) focused on the effects of modifying the language associated with math problems on students' test performance. The logic motivating this accommodation was that by "simplifying" the language associated with math problems, construct-irrelevant variance due to English proficiency would be reduced. They simplified 20 eighth grade NAEP math items and created two test booklets. Each booklet contained the original versions of 10 of the items, modified versions of the other 10 items and 5 "control" items. These booklets were randomly spiraled and administered to 1,174 eighth grade students in Los Angeles. Approximately 31% of these students were ELL. The results indicated that the modified items were only slightly easier than the original items and the difference in difficulty was not statistically significant. Also, there was no statistically significant interaction between ELL status and accommodation.

Abedi and Lord also interviewed 36 students about their preferences for original or modified versions of eight items. They found that students "preferred" the modified version of 6 of the 8 items.

Rivera and Stansfield (in press) also examined the effects of linguistic simplification of fourth and sixth grade science test items on a state assessment. They inserted simplified versions of ten items (6 MC and 4 constructed-response items) into two forms of a science test from the Delaware Student Testing Program. Two other forms contained these same items in their original form. Within each grade, these forms were randomly spiraled in with the other operational test forms. For non-ELL students, there were no significant differences in performance on the original or simplified items. Unfortunately, the sample sizes for ELL were

too small to compare their performance on the two versions of the items (sample sizes per form ranged from 9 to 23).

Abedi, Lord, Boscardin, and Miyoshi (2001) evaluated the effects of two accommodations for ELL: customized English dictionary, and English and Spanish glosses. The dictionary was customized by including only words that were on the test. The glosses were explanatory notes in the margins of the test booklet that identified key terms. They appeared in both English and Spanish in the same booklet. A 20-item science test was created from the pool of grade 8 NAEP Science items. Three forms were created: one that included only the items (no accommodation condition), one that included the customized dictionary, and one that included the glosses. They were randomly administered to 422 eighth grade students, 183 of whom were ELL. They found that ELL students performed best in the customized dictionary condition and that their performance with the glosses was about the same as in the standard condition. The effect size for the dictionary condition was about .38. There were no significant differences across test forms for non-ELL. This finding supports the interaction hypothesis, with respect to the accommodation of providing a customized dictionary.

Albus, Bielinski, Thurlow, and Liu (2001) also evaluated the benefits of a dictionary accommodation for ELL. They examined whether using a monolingual simplified English dictionary as an accommodation on a reading test improved the performance of Hmong students who were ELL. Students for this study came from three urban middle schools in a large metropolitan area of Minnesota. There were a total of 69 regular education students in the non-ELL group and 133 students in the Hmong ELL group.

Students were administered two reading passages with the English dictionary available, and two passages without the dictionary. The passages were designed to parallel Minnesota's

Basic Standards Reading Test. The two halves of the test were divided into Form A and Form B. The passages were assigned to forms so that Form A and Form B had the same overall difficulty. Students were allowed as much time as they needed to complete each half of the test, having been given a general time limit of two hours. Students were first asked to fill out a brief pre-test questionnaire about language background to provide self-ratings of their English and Hmong proficiency in several modalities: speaking, listening, and reading. Immediately after completion of the whole test, students were given a post-test questionnaire about dictionary use during the test, their opinions on possible usefulness of an English dictionary on a reading test, and other background information on dictionary use and instruction in the classroom. A short dictionary exercise also was given after the post-test survey to determine levels of student ability in using a dictionary.

The results showed that test performance for the ELL and non-ELL students was about the same under both standard and accommodated conditions. However, for those ELL who reported using the dictionary, and self-reported that they had an intermediate level of English reading proficiency, there was a statistically significant test score gain when they took the test with the dictionary accommodation. Furthermore, about 96% of the ELL group believed that providing an English dictionary would be helpful on a reading test.

Multiple Accommodations for ELL

Shepard, Taylor, and Betebenner (1998) examined the effects of accommodations for ELL on the *Rhode Island Grade 4 Mathematics Performance Assessment* (MPA). Students' performance on the *Metropolitan Achievement Test* (MAT) was used for comparison purposes. Although both programs were mandated by the State, accommodations for ELL were provided on the MPA, but not on the MAT. The participation rates for ELL was higher on the MPA

(79%) than on the MAT (68%). With respect to types of accommodations provided, they found that the most widely used accommodations involved changes in administrative procedures, especially oral reading of the assessment, repeating directions, testing in a special classroom or small group, and extended time.

Shepard et al. (1998) found that the mean MAT score for ELL students with less than two years in the U.S. was 1.45 standard deviations below the mean for general education students, but the ELL mean on the MPA was 1.08 standard deviations below the mean for general education students. For ELL who had been in the US for more than two years, the mean MAT performance was 1.20 standard deviation below the general education mean, but their mean MPA was .92 standard deviations below the mean. Although these effect size differences are small, and there are content differences between the two tests, the relative improvement of ELL on the MPA could be due in part to the accommodations provided on this test that were not provided on the MAT. Thus, Shepard et al. concluded "Accommodations consistently raised the relative position of LEP and special education students on the performance assessment compared to where they had been, relative to the general education mean, on the MAT" (p. 53).

Castellon-Wellington (1999) investigated oral presentation and extra time accommodations for ELL on an ITBS seventh grade social studies test. The participants included 106 seventh-grade ELL in 6 social studies classrooms across three schools in a middle-class suburb. First, the students took a form of the test under standard conditions. They were then asked which accommodation they would prefer for their retest (oral or extra time). Two weeks later, they were retested with one of the two accommodations. One third of the students received the accommodation of their preference, a third received the accommodation not of their preference, and a third received one of the two accommodations at random. The results

indicated no differences across students grouped by accommodation condition or preference, as well as no differences between the scores from accommodated and standard administrations.

Hafner (2001) evaluated the effects of extended time and extended oral presentation on the performance of ELL and non-ELL students on a standardized math test. The extended oral presentation accommodation involved reading and clarifying directions, rather than oral presentation of test items. Students from fourth (n=292) and seventh (n=159) grade classrooms were given a version of the TerraNova math test under standard conditions, extended time, or extended oral presentation (of directions). The math test was split into two parallel halves and separate analyses were conducted for each half-test.

Using a math pretest as a covariate, Haffner conducted analyses of covariance with accommodation condition as an independent variable (collapsing across the two types of accommodation conditions). The results suggested that students' scores improved under the accommodation conditions. However, she did not look at the interaction of accommodation condition and ELL status and she combined the results across fourth and seventh graders. Although she did not report effect sizes, we were able to compute eta-squared effect sizes (proportion of test score variance accounted for by accommodation condition) from the ANCOVA tables. The eta-squared values for the accommodation factor were .03 and .12 for half-tests 1 and 2, respectively. Given the way the analyses were conducted and reported, the results cannot be used to evaluate the interaction hypothesis, but irrespective of ELL status, the accommodations appear to have a small, positive effect on students' test performance.

Dual-language Test Booklets

Anderson, Liu, Swierzbin, Thurlow, and Bielinski (2000) evaluated the accommodation of providing dual-language test booklets on a reading test. The dual-language booklets presented

all reading passages in English, but all other test information, including directions, items, and response options, were written in two languages and presented side-by-side. The directions, items, and response options were also presented aurally in the native language on a cassette tape. The participants were 206 eighth grade students from two consecutive eighth grade classes from five schools in Minnesota. They were separated into three test groups: an accommodated ELL group (n=53), a non-accommodated ELL group (n=52), and a control group of general education students (n=101).

Anderson et al. found no statistically significant difference for ELL students between the standard and accommodated conditions. They also found that students tended to primarily use one version of the written test questions (either English or Spanish) and then refer to the other version when they encountered difficulties, and that students made little use of the oral presentation of the test questions in Spanish. They conjectured that, given the cost of producing translated tests, glossaries or dictionaries may be a more efficient accommodation for ELL.

Garcia et al. (2000) also evaluated the utility of a dual-language booklet for improving measurement of Spanish-speaking ELL on mathematics tests. Their dual-language accommodation involved translating (adapting) the English versions of NAEP math items into Spanish, and placing both versions of the items side-by-side in the same test booklet. Their study involved three groups of students: native English speakers, native Spanish speakers who received three or more years of academic instruction in English, and native Spanish speakers who received less than three years of academic instruction in English. A standard or dual-language version of the test was randomly administered to students in the second group (i.e., English-proficient native Spanish speakers); however, only the dual-language version was administered to the third group (i.e., the limited English proficient group), and only the standard

version was administered to the native English-speaking group. Thus, this design precluded analysis of the effect of the accommodation on the group of students to whom it could be hypothesized it would give the most benefit.

The results of the Garcia et al. study indicated that the native Spanish-speakers who received instruction for three or more years did not perform better on the dual-language test, in fact, they performed slightly worse. This finding held across both the multiple-choice and constructed-response sections of the test. The native Spanish-language students who received instruction for less than three years performed worse than their Spanish-speaking counterparts, but given these nonequivalent samples and absence of an English-language control group, the degree to which the dual-language accommodation affected their performance cannot be determined.

Garcia et al. found that Spanish-speaking LEP students appreciated the dual language booklets, with 85% of the students finding the booklets "useful" or "very useful." However, they also found that students who had three years or less instruction in English predominantly read and responded only to the Spanish versions of the test items. They also found that LEP students with high levels of English proficiency scored slightly lower when taking the dual language version of the test. These results suggest that the use of dual language booklets is promising, but more research is needed to determine its utility.

Summary of Research on ELL

A variety of accommodations have been investigated for improving the performance of ELL on standardized tests. Among these accommodations are providing customized dictionaries or glossaries, providing extra time, translating/adapting the test into a student's native language, and simplifying the English text associated with specific items. Many of these accommodations

have not been extensively studied, but the research that has been done indicates that some accommodations, such as linguistic modification and provision of a dictionary or glossary, show promise. A summary of the studies we reviewed is presented in Table 6.

Table 6

Summary of Experimental and Non-experimental Studies on ELL

Study	Accommodations	Design	Results	Supports H_1?
Abedi (2001)	Simplified Eng., bilingual glossary, customized dictionary	b/w group	No effects at 4th grade. Small gain for simpl. Eng. in 8th grade	8th grade only
Abedi, Hofstetter, Baker, & Lord (2001)	Simplified Eng., glossary, extra time, extra time + glossary	b/w group	Extra time w/ and w/out glossary helped *all* students	No
Abedi & Lord (2001)	Simplified Eng.	b/w group	Small, but insignificant gains	No
Abedi, Lord, Boscardin, & Miyoshi (2001)	Eng. dictionary; Eng. glosses, Span. translation	b/w groups	ELL gains assoc. with dictionary; no gains for others	Yes
Rivera & Stansfield (in press)	Linguistic modification of test	b/w groups	No differences for non-ELL	No
Albus, et al. (2001)	Dictionary	w/in & b/w group	No effect on validity; no significant overall gain for ELL students	No
Abedi, Courtney, et al.. (2001)	Dictionary; Bilingual dictionary; Linguistic modification of the test; Extended time	b/w group	Gain for ELL under dictionary conditions.	Yes
Shepard, Taylor, & Betebenner (1998)	Various	Ex post facto	Gains for both ELL and non-ELL.	Partial
Anderson et al. (2000)	Dual-language booklet	w/in & b/w group	No gain	No
Garcia et al. (2000)	Dual-language booklet	Quasi-experimental	Not relevant	N/A
Castellon-Wellington, (1999)	Extended time, Oral	Quasi-experimental	No gain.	No
Hafner (2001)	Extended time, oral directions	Quasi-experimental	Score gains for both ELL and non-ELL groups.	Unable to determine

Discussion

In reviewing the literature on the effects of test accommodations on SWD or ELL test performance, we found over 150 papers related to the topic. Of these studies, 40 involved empirical analysis at some level. Three papers were literature reviews, 14 papers were experimental studies involving SWD, 12 papers were non-experimental studies involving SWD, 8 papers were experimental studies on ELL, and 3 were non-experimental studies on ELL.

Our review focused on evaluating the interaction hypothesis. Specifically, we investigated whether evidence was available that test accommodations improved the performance of SWD or ELL, but did not improve the performance of students who were not ELL or SWD. The vast majority of studies pertaining to the interaction hypothesis showed that *all* student groups (SWD, ELL, and their general education peers) had score gains under accommodation conditions. Moreover, in general, the gains for SWD and ELL were *greater* than their general education peers under accommodation conditions. These conclusions varied somewhat across student groups and accommodation conditions, as we discuss below. However, it appears that *the interaction hypothesis needs qualification*. When SWD or ELL students exhibit greater gains with accommodations than their general education peers, an interaction is present. If the gains experienced by SWD or ELL are significantly greater than the gains experienced by their general education peers, the fact that the general education students achieved higher scores with an accommodation condition does not imply that the accommodation is unfair. It could imply that the standardized test conditions are too stringent for all students.

Examples of how the interaction hypothesis could be manifested are portrayed in Figures 1 and 2. In Figure 1, we see the hypothesis as stated by Shepard et al. (1998), Zuriff (2000), and

others. In this scenario, SWD or ELL experience a gain under the accommodation condition, but the general education students to not. In Figure 2, we see gains for both groups of students, but the gain for the SWD/ELL group is larger. Such differences in gains can be both statistically significant and meaningful, which signifies an important interaction. The type of interaction displayed in Figure 2 was the most common finding in the experimental studies we reviewed. Such results suggest that many accommodations are justified and are effective for reducing construct-irrelevant barriers to students' test performance. They also support the move toward universal test design, which we discuss later.

Figure 1

Illustration of Interaction Hypothesis

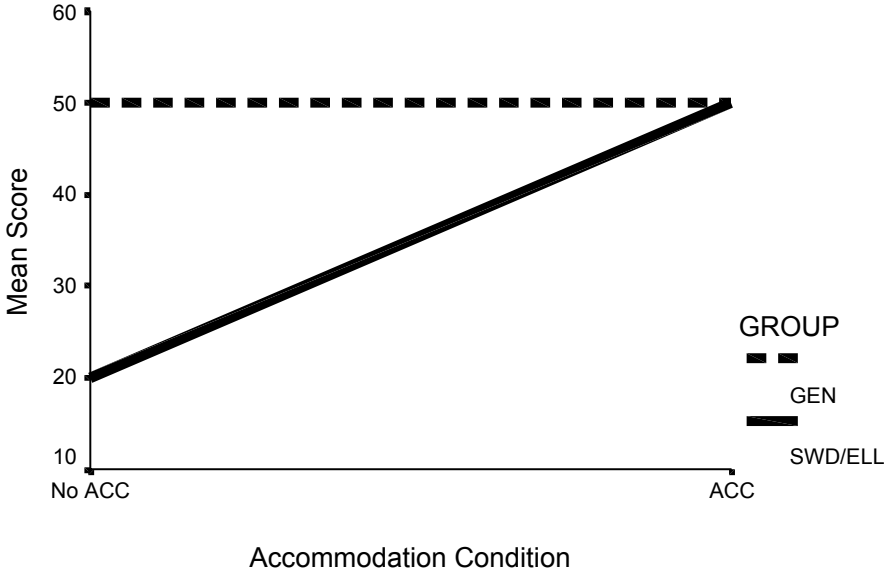
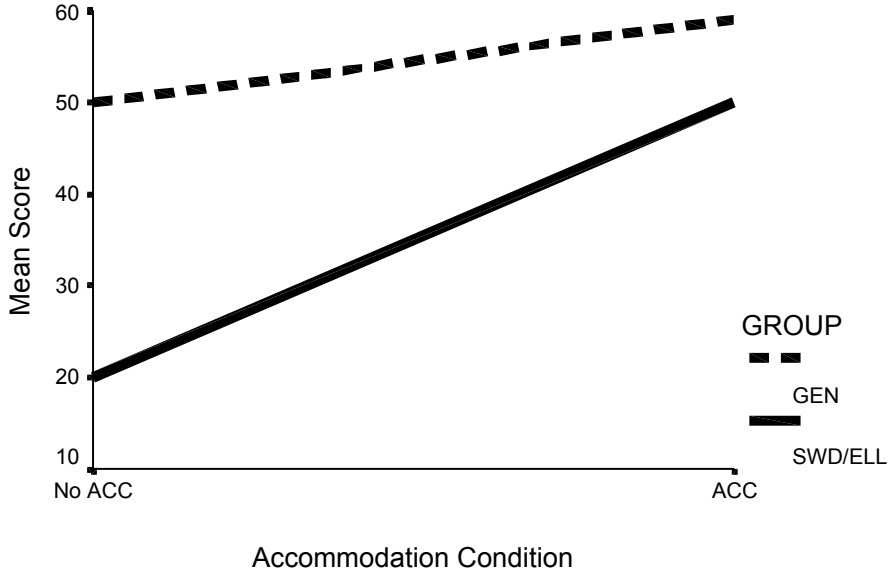


Figure 2

Illustration of Revised Interaction Hypothesis



Another finding that is clear from the literature is that there is great diversity within SWD and ELL groups, and great diversity in the way accommodations are created and implemented. Such variations create problems for generalizing the findings of particular studies to the accommodation condition in general. A good example of this variation is the accommodation of linguistic modification, also known as linguistic simplification, modified English, or simplified English. Excellent examples of how to modify English text to make it simpler exist (e.g., Abedi, 2001), but there are multiple methods that could be used (e.g., Tindal, Anderson, et al. 1998) and great differences in the quality with which the modification is done. Therefore, even with the extensive research conducted, it is difficult to make unequivocal statements about the effects of a particular accommodation on a particular type of student.

The problem of variation among accommodations and student groups was mentioned in several studies and many researchers called for more research to inform decisions about

matching accommodations to specific students (e.g., Abedi et al., 2001; Anderson et al. 2000; Shepard et al. 1998; Thompson et al. 2002). With respect to ELL, Shepard et al. stressed the need to avoid a "one size fits all" approach in providing test accommodations and concluded "Better training is needed to make both better classification decisions (who is LEP?) and better accommodation decisions" (p. 54). This finding was echoed by Anderson et al. and Chiu and Pearson (1999) who recommended that ELL accommodations be based on individualized education plans. Thompson et al. went one step further by recommending that students be involved in making decisions about test accommodations. They stated "Studies are also needed that explore the desirability and perceived usefulness of accommodations by students themselves." (p. 15)

Notwithstanding caveats about heterogeneity across accommodation condition and students groups, we draw some very general conclusions across studies for the popular accommodations of extended time and oral presentation, as well as the few accommodations designed for ELL.

General Conclusions on the Accommodation of Extended Time

The accommodation of extended time was studied for both SWD and ELL, but was most often studied as an accommodation for SWD. As mentioned earlier, Thompson et al (2002) concluded "research has fairly consistently concluded that extended time helps students with disabilities" (p. 15). Our review supports this contention, although it is clear that extra time appears to improve the performance of all student groups, not just those with disabilities. Experimental studies that looked at the effects of extended time found gains for students without disabilities, although the gains for SWD were significantly greater (as depicted in Figure 2). This finding suggests two things: (a) many SWD need extra time to demonstrate their true

knowledge, skills, and abilities; and (b) many educational tests are speeded to some extent. Given that most of the tests studied are not intended to be speeded, extending the time limits for all examinees may reduce the need for extended time accommodations. It is also important to note the recommendations of Huesman and Frisbie (2000) who stated that, when extended time accommodations are provided, they should be based on individual student needs and prescribe specific time extensions, rather than a universal rule such as time-and-a-half.

With respect to ELL, the effects of extended time are less clear. None of the studies reviewed looked at extended time in isolation, rather, it was included along with other accommodations such as dual-language booklet, glossary, dictionary, or linguistic modification. Where score gains for ELL were found, extra time was typically involved. Although this finding is not surprising, more research on extended time for ELL, without other forms of accommodation, will better determine its effectiveness as an accommodation.

General Conclusions on the Accommodation of Oral Presentation

The studies we reviewed involving oral presentation used a variety of methods for presenting test material orally to examinees. These methods included teacher reading, student reading, and screen-reading software. As indicated in Table 5, about half the studies that focused on oral accommodations found positive effects (i.e., Fuchs et al. 2000; Weston, 2002, Tindal et al., 1998; Johnson, 2000). However, several studies found either no gains for SWD (Kosciolek & Ysseldyke, 2000) or similar gains for SWD and students without disabilities (Meloy et al. 2000; Brown & Augustine, 2001). Therefore, the benefits of oral accommodation remain unclear. This finding could be due to the fact that only specific subsets of SWD need this type of accommodation and administering it to a larger group of SWD obscures its effects.

General Conclusions on Accommodations for ELL

Linguistic modification (simplified or modified English) is an attractive accommodation for ELL in subject areas such as science or math. Minor changes in the text associated with test items should not change the construct measured by such tests, and as Abedi (2001) stated "modifying test questions to reduce unnecessary language complexity should be a priority in the development and improvement of all large-scale assessment programs" (p. 106).

The research on the effects of linguistic modification is mixed. Abedi, Hofstetter, Baker, and Lord (2001) claimed that this accommodation was the most effective in reducing the score gap between ELL and non-ELL, but in their study, the gap was narrowed because non-ELL scored worse on the linguistically modified test, not because the ELL performed substantially better. Significant, but small, gains were noted for eighth grade students in the Abedi (2001) study, but not for fourth grade students. Abedi explained this finding by hypothesizing "With an increase in grade level, more complex language may interfere with content-based assessment" (p. 13) and "…in earlier grades, language may not be as great a hurdle as it is in the later grades" (p. 14). In other studies, providing ELL with customized dictionaries or glosses seemed to improve the performance of ELL (e.g., Abedi, Lord, Boscardin, & Miyoshi (2001). At this juncture, there is little support for dual-language test booklets.

Limitations of Studies

Although our review was extensive, there are some general limitations across many of the studies that should be mentioned. First, much of the research on SWD has focused on relatively small, and ethnically homogeneous groups of students. The proportion of under-represented minority students in these studies is typically small. Furthermore, much of the work on ELL has

focused on children in the Los Angeles area. Thus, the generalizability of the experimental findings across various groups of students is limited.

A second general limitation of the studies reviewed is that most of the studies focused on elementary school grades. As illustrated in Table 3, nearly two thirds of the studies reviewed focused on students in grades 3 to 8, while the remainder of the studies evaluated the effect of accommodations on test performance for students in grades 9 to 12. We highlight this disproportion to draw attention to two points that should be raised. First, virtually no *experimental* studies involved secondary school students. This may be an indication that traditional designs are more difficult to apply when secondary school students are involved[4]. Second, the lack of research on secondary school students is unfortunate, given that fact that many states have implemented high school graduation tests. We only need to look at the depressing data that compares drop out and graduation rates between students with and without disabilities to realize that more empirical evidence is needed to better understand not only the interaction effect of accommodations on test performance, but also the underlying efficiency and usefulness of the accommodations with respect to SWD performance.

A third limitation of many studies is that effect sizes were not reported. Where it was possible, we computed effect sizes by transforming mean differences across groups and conditions to standard deviation units. However, this was not possible for all studies.

[4] The nature of the high school curriculum (i.e., differentiated curriculum), where academic versus career-oriented tracks tend to naturally sort students according to their ability, motivation, and interests may affect the variance attributed to the effects of a particular accommodation. Thus, there is much more to consider when designing studies with older students given the complex interaction of cognitive ability and academic requirements. For one, random assignment becomes an extraordinary challenge. And for SWD their special education history (i.e., years being served, how their learning and tests needs have been accommodated in the past, etc.) could be the source of unexplainable error.

Beyond Test Accommodations: Universal Test Design

The common finding that many accommodations improved the performance of all student groups suggests that the standardized administration conditions of many tests may reduce the validity of the inferences derived from the test scores. Such findings are congruent with the idea that tests should be constructed and administered more flexibly so that accommodations become unnecessary. Borrowing from the field of architecture, the concept of Universal Design for Learning (UDL) has emerged as a way to construct tests that would not need to be accommodated for SWD or ELL. UDL seeks to create learning and assessment activities that accommodate the widest spectrum of users possible.

Under UDL, tests designed for SWD or ELL would increase access for all students. A non-psychometric corollary is the common example of building televisions with closed caption capabilities for the hearing impaired. Once thought of as an accommodation for only these individuals, closed-captioned televisions are now commonplace in airports, restaurants, and health clubs for everyone to gain access to audio information when their hearing has been "impaired." With respect to testing, Thompson et al. (2002) embraced this notion of universal test design and encouraged more research in this area.

> Future research should…explore the effects of assessment design and standardization to see whether incorporating new item designs and incorporating more flexible testing conditions reduces the need for accommodations while facilitating measurement of the critical constructs for students with disabilities. It is possible that through implementation of the principles of universal test design…the need for accommodations will decrease, and the measurement of what students know and can perform will improve for *all* students." (Thompson et al., p. 17).

Technology-based accommodations have tremendous potential to improve the accessibility of tests for students with disabilities, but adding technology to an assessment (i.e., retrofitting) does not embrace the tenets of UDL and is otherwise limited (Dolan & Hall, 2001).

The UDL approach presents us with a rich area of new research possibilities that could improve the likelihood that SWD and ELL could gain more equitable access to tests and ultimately produce more meaningful and valid information about what they know and can do.

It is interesting to note that some state-mandated testing programs, such as the Massachusetts Comprehensive Assessment System, have already moved in this direction by administering their tests without time limits. Although UDL represents an exciting development that may eliminate or reduce the need for test accommodations, it is important to point out that there is virtually no research on the benefits of universal test design at this time.

Closing Remarks

Our extensive review of the literature covered many types of students and many types of accommodations. Although unequivocal effects were not found, it is important to remember that the provision of test accommodations is an important area of promoting equity and validity in educational assessment. The research in this area is difficult to do and often provides inconsistent results, but much progress has been made. We know that extended time improves the scores for many students. However, many other questions remain such as what is the appropriate amount of extended time and what other accommodations are best, given the myriad of student conditions and accommodation options. Our review indicates that many accommodations have positive, construct-valid effects for certain groups of students. The remaining challenge is to implement these accommodations appropriately and identify which accommodations are best for specific students. Another challenge is developing more flexible tests that would make accommodations unnecessary. These challenges appear surmountable. Thus, it appears that research in the area of test accommodations will continue to result in more valid assessment practices.

References

Abedi, J. (2001, December). *Language accommodation for large-scale assessment in science: Assessing English language learners (Final Deliverable, Project 2.4 Accommodation).* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, University of California Los Angeles.

Abedi, J., Courtney, A., Mirocha, S.L., & Goldberg, J. (2001). Language accommodation for large-scale assessment in science: a pilot study. *CSE technical report*.

Abedi, Hofstetter, Baker, & Lord (2001, February): NAEP math performance and test accommodations: Interactions with student language background. *CSE technical report 536*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Lord, C. (2001). The Language Factor in Mathematics Tests. *Applied Measurement in Education, 14(3),* 219-234.

Abedi, J., Lord, C., Boscardin, C.K., & Miyoshi, J. (2001). The effects of accommodations on the assessment of limited English proficient students in the National Assessment of Educational Progress. *National Center for Education Statistics Working Paper, Publication No. NCES 200113*, Washington, DC: National Center for Education Statistics.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19 (3)*, 16-26.

Albus, A., Bielinski, J., Thurlow, M., & Liu, K. (2001). The effect of a simplified English language dictionary on a reading test *(LEP Projects Report 1*). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [today's date], from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/LEP1.html

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*, 185-198.

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

Anderson, M., Liu, K., Swierzbin, B., Thurlow, M., & Bielinski, J. (2000). Bilingual accommodations for limited English proficient students on statewide reading tests: Phase 2 (*Minnesota Report No. 31*). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [today's date], from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/MnReport31.html

Braun, H., Ragosta, M., & Kaplan, B. (1986). The predictive validity of the Scholastic Aptitude Test for disabled students (*Research Report 86-38*). New York: College Entrance Examination Board.

Bridgeman, B., Trapani, C., & Curley, E. (in press). Impact of fewer questions per section on SAT I scores. *Journal of Educational Measurement*.

Brown, P.J., & Augustine, A. (April 2001). *Screen reading software as an assessment accommodation: Implications for instruction and student performance*. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Cahalan, C., Mandinach, E., & Camara, W. (2002). Predictive validity of SAT I: Reasoning Test for test takers with learning disabilities and extended time accommodations. *College Board Research Report (RR 2002-05).* New York, NY: College Board.

Calhoon, M.B., Fuchs, L.S., & Hamlett, C.L. (Fall 2000). Effects of computer-based test accommodations on mathematics performance assessments for secondary students with learning disabilities. *Learning Disability Quarterly; 23 (4),* 271-82.

Camara, W., Copeland, T., & Rothchild, B. (1998). Effects of extended time on the SAT I: Reasoning Test: Score growth for students with learning disabilities *(College Board Research Report 98-7)*. New York, NY: The College Board.

Castellon-Wellington, M. (1999). *The impact of preference for accommodations: The performance of English language learners on large-scale academic achievement tests*.

Chiu, C. WT, & Pearson, P.D. (1999, June*). Synthesizing the effects of test accommodations for special education and limited English proficient students*. Paper presented at the National Conference on Large Scale Assessment, Snowbird, UT.

Dolan, R. P. and Hall, T. E. (2001). Universal Design for Learning: Implications for Large-Scale Assessment. *IDEA Perspectives, 27(4)*: 22-25.

Elliott, S.N., Kratochwill, T.R., & McKevitt, B.C. (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., Binkley, E., & Crouch, R. (Fall 2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children; 67,* 67-81.

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. *School Psychology Review, 29,* 65-85.

Garcia, T., del Rio Paraent, L., Chen, L., Ferrara, S., Garavaglia, D, Johnson, E., Liang, J., Oppler, S., Searcy, C., Shieh, Y., & Ye, Y. (2000, November). *Study of a dual language test booklet in 8th grade mathematics: Final report*. Washington, DC: American Institutes for Research.

Geisinger, K. F. (1994). Psychometric issues in testing students with disabilities. *Applied Measurement in Education*, 7, 121-140.

Green, P., & Sireci, S. G. (1999). Legal and psychometric issues in testing students with disabilities. *Journal of Special Education Leadership, 12(2),* 21-29.

Hafner, A. (2001, April). *Evaluating the impact of test accommodations on test sores of LEP students & Non-LEP students*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

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

Huesman, R.L., & Frisbie, D. (2000, April). *The validity of ITBS reading comprehension test scores for learning disabled and non learning disabled students under extended-time conditions*. Paper presented at the Annual Meeting of the National Council on measurement in Education, New Orleans, LA.

Johnson, E. (2000). The effects of accommodations on performance assessments. *Remedial and Special Education, 21,* 261-267.

Johnson, E., Kimball, K., Brown, S. O., & Anderson, D. (2001). A statewide review of the use of accommodations in large-scale, high stakes, assessments. *Exceptional Children, 67,* 251-264.

Koenig, J. A. (Ed.) (2002). *Reporting test results for students with disabilities and English language learners: Summary of a workshop*. Washington, DC: National Research Council.

Koretz, D., & Hamilton, L. (2000). Assessment of students with disabilities in Kentucky: Inclusion, student performance, and validity. *Educational Evaluation and Policy Analysis, 22,* 255-272.

Koretz, D., & Hamilton, L. (2001, April). The Performance of Students With Disabilities on New York's Revised Regents Comprehensive Examination in English. Center for the Study of Evaluation, *CSE Technical Report 540*.

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. Retrieved January 2003, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Technical28.htm

McKevitt, B. C., Elliot, S. N. (in press). The effects and consequences of using testing accommodations on a standardized reading test. *School Psychology Review*.

McKevitt, B.C., Marquart, A.M., Mroch, A.A., Schulte, A., Elliott, S., & Kratochwill, T. (2000). *Understanding the effects of testing accommodations: a single-case approach*. Paper Presented at the annual meeting of the National Association of School Psychologists, New Orleans, LA.

Meloy, L., Deville, C., & Frisbie, D. (2000, April). *The effects of a reading accommodation on standardized test scores of learning disabled and non learning disabled students*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Phillips, S.E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7,* 93-120. Needham Heights, MA: Allyn and Bacon.

Pitoniak, M., & Royer, J. (Spring 2001). Testing accommodations for examinees with disabilities: a review of psychometric, legal, and social policy issues. *Review of Educational Research. 71 (1)*, 53-104.

Rivera, C., and Stansfield, C.W. (in press). The effects of linguistic simplification of science test items on performance of limited English proficient and monolingual English-speaking students. *Educational Assessment*.

Robin, F., Sireci, G. S., & Hambleton, R. K. (in press). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing*.

Runyan, M. K. (1991). The effect of extra time on reading comprehension scores for university students with and without learning disabilities. *Journal of Learning Disabilities, 24,* 104-108.

Scarpati, S. (1991). Current perspectives in the assessment of the handicapped. In R.K. Hambleton & J.N. Zall (Eds.). *Advances in educational and psychological testing* (pp. 251-276)., Norwell, MA: Kluwer.

Scarpati, S. (2003). Test accommodations for disabilities. *Encyclopedia of psychological assessment* (pp. 957-960). London: Sage.

Schulte, A.A., Elliott, S.N., & Kratochwill, T.R. (March, 2001). Effects of testing accommodations on students' standardized mathematics test scores: An experimental analysis. *School Psychology Review, 30,* 527-547.

Shepard, L., Taylor, G., & Betebenner, D. (1998). *Inclusion of limited-English-proficient students in Rhode Island's grade 4 mathematics performance assessment*. Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Sireci, S.G., & Geisinger, K.F. (1998). Equity issues in employment testing. In J.H. Sandoval, C. Frisby, K.F. Geisinger, J. Scheuneman, & J. Ramos-Grenier (Eds.). *Test interpretation and diversity* (pp. 105-140). American Psychological Association: Washington, D.C.

Sireci, S.G., & Khaliq, S.N. (2002, April). *An analysis of the psychometric properties of dual language test forms*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

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. Retrieved January 2003, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Technical34.htm

Tindal, G., Anderson, L., Helwig, R., Miller, S., & Glasgow, A. (Sum 1998). Accommodating students with learning disabilities on math tests using language simplification. *Exceptional Children, 64,* 439-50.

Tindal, G., & Fuchs, L. (2000) *A summary of research on test changes: An empirical basis for defining accommodations.* Lexington, KY: University of Kentucky Mid-south Regional Resource Center, Interdisciplinary Human Development Institute.

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.

Trimble, S. (1998). Performance Trends and Use of Accommodations on a Statewide Assessment: Students with Disabilities in the KIRIS On-Demand Assessments from 1992-93 through 1995-96. *State Assessment Series, Maryland/Kentucky Report 3*. National Center on Educational Outcomes, Minneapolis, MN.; Maryland State Dept. of Education, Baltimore.

Walz, L., Albus, D., Thompson, S., & Thurlow, M. (December 2000). *Effect of a multiple day test accommodation on the performance of special education students*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved January 3, 2003 from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/MnReport34.html

Weston, T.J. (2002, July). The validity of oral accommodation in testing. NAEP Validity Studies (NVS) Panel.

Wightman, L. (1993). *Test takers with disabilities: A summary of data from special administrations of the LSAT (LSAC Research Report 93-03)*. Newton, PA: Law School Admissions Council.

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

Ziomek, R.L., & Andrews, K.M. (1998). ACT Assessment Score Gains of Special-Tested Students Who Tested at Least Twice (*Report No.: ACT-RR-98-8*). Iowa city, Iowa: ACT Research Report Series.

Zurcher, R., & Bryant, D.P. (2001) The validity and comparability of entrance examination scores after accommodations are made for students with LD. *Journal of Learning Disabilities, 34 (5),* 462-471.

Zuriff, G.E. (2000). Extra examination time for students with learning disabilities: An examination of the maximum potential thesis. *Applied Measurement in Education, 13 (1),* 99-117.

Appendix A

Annotated Bibliography

Abedi, J., Courtney, A., Mirocha, S.L., & Goldberg, J. (2001). *Language accommodation for large-scale assessment in science: a pilot study*. CSE technical report.

The study examined the effects of language accommodation on four NAEP Science forms–one form without accommodation, while the remaining three included extra time and either an English dictionary, a bilingual dictionary or linguistic modification. The 1994 NEAP reading assessment was used to measure the reading ability of the participants. A student background questionnaire and a follow-up questionnaire were also administered. The participants consisted of 611(317 ELL students) 4th- and 8th- grade students in the language groups of Spanish, Chinese, Filipino, and Korean. The results revealed that some of the accommodation strategies employed were effective in increasing the performance of the ELL students and reducing the performance gap between ELL and non-ELL student, though the effectiveness of accommodation may vary across grade levels. Also, the validity of assessment was not compromised by the use of accommodation.

Abedi, Hofstetter, Baker, & Lord (2001, February): *NAEP Math Performance and Test Accommodations: Interactions with student language background*. CSE technical report 536. Los Angeles, CA: National Center for research on Evaluation, Standards, and Student Testing.

The authors administered a short 8th grade NAEP math test to 950 students under one of 5 conditions: no modification, use of simplified (modified) vocabulary, use of a glossary allowed, extra time, and extra time with glossary. Conditions were randomly assigned to students. LEP, non-LEP, and FEP (former LEP students who are now fully English proficient) were tested. Most LEP students were Spanish-speaking. For most students, improvements were made under all accommodations, particularly extra time and extra time with glossary. The gains for the LEP group were small. The authors concluded that the "modified English" accommodation was "the only accommodation type that narrowed the score difference between LEP and non-LEP students." However, this "narrowing" was due to the fact that the non-LEP students performed poorest on this version, not that the LEP group did much better than other conditions.

Abedi, J., & Lord, C. (2001). The Language Factor in Mathematics Tests. *Applied Measurement in Education*, *14(3),* 219-234.

The authors revised 20 items from an 8th grade NAEP Math test to simplify the English text associated with the items. Over 1,100 students participated, about half of whom were ELL. Both ELL and non-ELL responded to sets of original and simplified items. There were no statistically significant differences between performance on original and

simplified items for either group. However, interviews with 36 students revealed that ELL students preferred the modified version of 6 of 8 items.

Abedi, J., Lord, C., Boscardin, C.K, & Miyoshi J. (2001, September). *The Effects of Accommodations on the Assessment of Limited English Proficient (LEP) Students in the National Assessment of Educational Progress (NAEP).* National Center for Education Statistics Working Paper, Publication No. NCES 200113, Washington, DC: National Center for Education Statistics.

Evaluated the effects of two accommodations for ELL: customized English dictionary, and English and Spanish glosses. The dictionary was customized by including only words that were on the test. The glosses were explanatory notes in the margins of the test booklet that identified key terms. They appeared in both English and Spanish in the same booklet. A 20-item science test was created from the pool of grade 8 NAEP Science items. Three forms were created: one that included only the items (no accommodation condition), one that included the customized dictionary, and one that included the glosses. They were randomly administered to 422 eighth grade students, 183 of whom were ELL. They found that ELL students performed best in the customized dictionary condition and that their performance with the glosses was about the same as in the standard condition. The effect size for the dictionary condition was about .38. There were no significant differences across test forms for non-ELL. This finding supports the interaction hypothesis, with respect to the accommodation of providing a customized dictionary.

Abedi, Lord, Hofstetter, & Baker (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice*, *19(3)*, 16-26.

This is an extension of the previous study (Abedi, Hofstetter, Baker, & Lord (2001, February), but it added confirmatory factor analysis (CFA) on test forms with math and reading parcels to see if structure differed across ELL and non-ELL groups. It is unclear if multi-group or separate-group CFAs were done. Concluded there were structural differences because correlation between reading and math was higher for ELL group. However, the parcels exhibited similar factor loadings across groups. Noted interaction between accommodation and type of student.

Albus, A., Bielinski, J., Thurlow, M., & Liu, K. (2001). *The effect of a simplified English language dictionary on a reading test* (LEP Projects Report 1). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [today's date], from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/LEP1.html

Examined whether using a monolingual simplified English dictionary as an accommodation on a reading test improved the performance of Hmong ELL students. The participants were 69 regular education students and 133 ELL students from three urban middle schools in a large metropolitan area of Minnesota. Students were

administered two reading passages with the English dictionary available, and two passages without the dictionary. The passages were designed to parallel Minnesota's Basic Standards Reading Test. The results showed that test performance for the ELL and non-ELL students was about the same under both standard and accommodated conditions. However, for those ELL who reported using the dictionary, and self-reported that they had an intermediate level of English reading proficiency, there was a statistically significant test score gain when they took the test with the dictionary accommodation. Furthermore, about 96% of the ELL group believed that providing an English dictionary would be helpful on a reading test.

Anderson, M., Liu, K., Swierzbin, B., Thurlow, M., & Bielinski, J. (2000). *Bilingual accommodations for limited English proficient students on statewide reading tests: Phase 2* (Minnesota Report No. 31). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [today's date], from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/MnReport31.html

Evaluated the accommodation of providing dual-language test booklets on a reading test, in which all reading passages were presented in English, but all other test information were written in two languages and presented side-by-side. The directions, items, and response options were also presented aurally in the native language on a cassette tape. The participants were 206 eighth grade students from two consecutive eighth grade classes from five schools in Minnesota and were formed into three groups: an accommodated ELL group (n=53), a non-accommodated ELL group (n=52), and a control group of general education students (n=101). They found no statistically significant difference for ELL students between the standard and accommodated conditions. They also found that students tended to primarily use one version of the written test questions (either English or Spanish) and then refer to the other version when they encountered difficulties, and that students made little use of the oral presentation of the test questions in Spanish. They conjectured that, given the cost of producing translated tests, glossaries or dictionaries may be a more efficient accommodation for ELL.

Barton & Huynh (2000). Patterns of errors made on a reading test with oral accommodation.

Looked at patterns of errors made by SWD across different accommodations on the South Carolina Basic Skills Assessment. Three types of oral accommodation were studied (reader, audiotape, video w/ sign). No non-SWD comparison group. Used log-linear analysis to look at test score, response choice, and disability group. 30 items were found to be moderately associated with disability category. Literal comprehension caused most difficulty for students. "Diversion" errors most distracting. Research design and results do not provide data on effect of accommodation on test performance.

Braun, H., Ragosta, M., & Kaplan, B. (1986a). *The predictive validity of the Scholastic Aptitude Test for disabled students.* Research Report 86-38, New York: College Entrance Examination Board.

NOTE: Reproduced in 1988 Book, *Testing Handicapped People* (Willingham et al., 1988). Looked at differential predictive validity across SWD SAT scores from accommodated tests and standard tests. Used regression equation from non-SWD for predicting freshman GPA of SWD. Concluded that visually impaired and physically handicapped had comparable predictive validity. HSGPA alone underpredicted SWD, SAT alone overpredicted SWD, especially LD. Using both SAT and HSGPA balanced the prediction errors.

Brown, P.J., & Augustine, A. (April 2001). *Screen reading software as an assessment accommodation: Implications for instruction and student performance*. Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).

Evaluated the effect of providing screen-reading software to students with reading difficulties. Two parallel social studies test forms and two parallel science test forms developed from publicly released NAEP items were delivered to students both with and without reading disabilities from high schools in Delaware and Pennsylvania. Within each subject area, students took the test forms under standard and computer-read conditions. After controlling for students' reading proficiency, there was no difference in student performance under standard and screen-reading conditions. The specific numbers of students with and without "reading difficulties" were not provided the effect of the accommodation across student types were not compared.

Cahalan, C., Mandinach, E., & Camara, W. (2002). *Predictive validity of SAT I: Reasoning Test for test takers with learning disabilities and extended time accommodations.* College Board Research Report (RR 2002-05). New York, NY: College Board.

Investigated the differential predictive validity of the SAT across students who took the test with and without extended time. Concluded that (a) the predictive validity coefficients were significantly lower for SWD who took the test with extended time than for students who took the test under standard time conditions, (b) the SAT scores for students with learning disabilities who requested extended time tended to over-predict their first year college grades, and (c) this over-prediction was greatly reduced when SAT scores and high school grades were considered together. When looking at these results across the sexes they found that SAT scores for female SWD *under*predicted first year college grades when they were considered together with high school grades. The standardized residual (from predicting first year college GPA from the SAT I) for female students with learning disabilities was .02 (overpredicted), but for males, the residual was .21 (overpredicted).

Calhoon, M.B., Fuchs, L.S., & Hamlett, C.L. (Fall 2000). Effects of computer-based test accommodations on mathematics performance assessments for secondary students with learning disabilities. *Learning Disability Quarterly; 23 (4)*, 271-82.

Compared the effects of a read-aloud accommodation on a math test that was provided by either a teacher or a computer, or a computer with video. Eighty-one 9th through 12th grade SWLD participated in the study and were tested in each of the four conditions (with the forth one being standard condition) over a four-week period (the conditions were counterbalanced). The results indicated that all read-aloud accommodations led to higher scores for these students compared with the standard administration (effect sizes ranged from about one-quarter to one-third of a standard deviation). However, there were no significant differences among the read-aloud methods. Also, about two-thirds of the students preferred the anonymity provided by the computer when taking the test.

Camara, W., Copeland, T., & Rothchild, B. (1998). *Effects of extended time on the SAT I: Reasoning Test: Score growth for students with learning disabilities.* College Board Research Report 98-7, New York, NY: The College Board.

Investigated score gains for students with learning disabilities (SWLD and students without disabilities who took the SAT once in their junior year of high school and once in their senior year. Design focused on three specific groups: SWLD who took the test once under standard conditions and once with extended time, SWLD who took the test twice with extended time, and students without disabilities who took the test twice under standard conditions. Score gains for SWLD taking test with extended time were three times larger than score gains for students without disabilities who took the test twice under standard conditions. For SWLD who took the SAT first with standard time and second with extended time, the gain under the extended time accommodation was 38.1 points on the math test and 44.6 points on the verbal test. For students without disabilities who took the test twice under standard conditions the gain from first testing to second testing was 11.8 on math and 12.9 on verbal. SWLD who took the test under extended time first and standard time second did worse, on average, on their second testing (i.e., under standard time). The loss for these students under standard time conditions was 6.1 points for math and 8.6 points for verbal.

Castellon-Wellington, M. (1999). The Impact of Preference for Accommodations: The Performance of English Language Learners on Large-Scale Academic Achievement Tests.

Investigated oral presentation and extra time accommodations for ELL on an ITBS seventh grade social studies test. Participants included 106 seventh-grade ELL in 6 social studies classrooms across three schools in a middle-class suburb. After taking a form of the test under standard condition, the students were asked which accommodation they would prefer for their retest (oral or extra time). Two weeks later, they were retested with one of the two accommodations. One third of the students received the accommodation of their preference, a third received the accommodation not of their preference, and a third received one of the two accommodations at random. The results

indicated no differences across students grouped by accommodation condition or preference, as well as no differences between the scores from accommodated and standard administrations.

Chiu, C. WT, & Pearson, P.D. (June 1999*). Synthesizing the Effects of Test Accommodations for Special Education and Limited English Proficient Students*. Paper presented at the National Conference on Large Scale Assessment (Snowbird, UT, June 13-16, 1999). 51p. American Institutes for Research in the Behavioral Sciences, Palo Alto, CA.

Using meta-analysis, the authors conducted a literature review of 20 empirical studies on both SWD and ELL students. They concluded that overall SWD had significant score gains under the accommodation condition, but the gain was small (.16 of a standard deviation unit), but there was significant variation across studies, and that extended time only slightly helped SWD more than it helped students without disabilities. They also cautioned that the average effect sizes they noted should be interpreted extremely cautiously due to the wide variety of accommodations that were used (and the quality in which they were implemented) and the heterogeneous types of students they were used for.

Elliot, S. N., Kratochwill, T. R., & McKevitt, B. C. (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

One hundred fourth grades in which 59 students were without disabilities were compared on 8 mathematics and science performance test items that were relevant to the state of Wisconsin's content standards. Accommodations used were: verbal encouragement, extra time, individual test administration, read directions to student, read subtask directions, paraphrase directions, restate directions or vocabulary, read questions and content, restate questions, spelling assistance, mark task book to maintain place, and manipulatives. Alternating treatment single case designs were used. As a group, the performance of students with disabilities in the condition were no accommodations were provided scored nearly 1 SD lower to their performance when they were provided accommodations. Accommodations had a significantly higher impact on SWD than students without disabilities receiving teacher-recommended accommodations or standard accommodations. Testing accommodations had a medium to large effect on more than 75% of SWD, but also had a similar effect on more than 55% of the students without disabilities.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. L., Binkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgements about test accommodations. *Exceptional Children, 67*, 67-81.

Students with and without learning disabilities in grades 4 and 5 completed four, brief assessments in reading using 400 word passages. Students answered eight multiple choice questions (six literal; two inferential). Three passages were used for each of the conditions of (1) standard, (2) extended time, (3) large print, and (4) student reads aloud.

For extended time and large print accommodations students with LD did not benefit more that students with out LD. Reading aloud proved beneficial to students with LD but not to students with out LD but reading aloud was the only accommodation that was administered individually. Teachers were a poor source of decision making in that they recommended may more accommodations than were necessary, as indicated by a data-based approach.

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

Students with learning disabilities (LD) and without disabilities were compared on mathematics test performance using a variety of accommodations along with the utility of teacher judgements in specifying appropriate accommodations. Standardized data based on curriculum assessment techniques were used to supplement teacher decisions. Students with LD did not benefit more that students without LD from extended time, from using a calculator or from having the instructions read to them. These accommodations added to the existing advantage students with out disabilities had over the students with LD. On problem solving, students with LD did profit more using accommodations than did the non-disabled. Data based decisions rather than teacher based decisions better predicted who would benefit from using accommodations.

Grise, P., Beattie, S., and Algozzine, B. (1982). Assessment of minimum competency in fifth grade learning disabled students: Test accommodations make a difference. *Journal of Educational Research, 76*(1), 35-40.

The authors modified a version of FL statewide competency test and administered a portion to a sample of students with learning disabilities. A variety of modifications were made, most of them related to the format of the test, and most represent good test construction practices rather than large changes to the test (e.g., ordering items according to difficulty, vertical format of answer options, shading reading comprehension passages, etc.). A large print version of the modified test was also created. There is no indication that the sample was randomly selected or was representative of 5th grade LD students in FL. The authors do not report sample sizes for the LD group based on each test form taken. Comparisons are made to students with LD who took the original version and to the general population who took the test. The authors concluded that test modifications improved the performance of LD students; however, this conclusion is not justified due to non-random/representative sample of LD students, unknown sample sizes for statistical comparisons, and confound between disability and accommodation.

Hafner, Anne. (April, 2001). *Evaluating the impact of test accommodations on test sores of LEP students & Non-LEP students*. Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA).

Examined the effects of providing extra time and extra time with extended oral presentations as accommodations to LEP and non-LEP students. The instrument used was

CTB/McGraw Hill TerraNova math test, and background and demographic information were collected via survey, along with other outcome variables to enable validation of the instrument. A sample of fourth (n=292) and seventh grade students (n=159) from California and New Mexico participated and a quasi-experimental analysis of variance design was used. The results showed that overall the use of accommodations did affect student test scores, and that both LEP and non-LEP groups benefited.

Hambleton, R. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, *10 (3)*, 229-244.

This is an issues paper regarding how to ensure appropriate test translation/adaptation procedures. Focuses on test translation issues, but these issues generalize to accommodations within the same language. Provides 22 guidelines for adapting tests and for evaluating adapted tests that are endorsed by the International Test Commission. Concludes that proper adaptation guidelines should be followed and that both statistical and qualitative analyses are necessary for evaluating the comparability of original and accommodated tests.

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

This study proposed that the accuracy with which teachers can predict which students with disabilities would benefit from test accommodations would influence the types, fairness and validity of the accommodations during math testing. Pretest reading and math scores were also used as predictors of who would benefit from read-aloud accommodations. Traditional test booklets and a video of a person reading the test items and directions were used. Modest effects were detected between accommodated and non-accommodated test conditions. Teachers were unable to predict which students would benefit from the accommodation. Student profiles using pre test math and reading scores did not match their performance profiles from accommodated math test.

Hollenbeck, K., Tindal, G., Stieber, S., & Harniss, M. ( ). Handwritten versus word processed compositions: Do judges rate them differently?

This study was designed to discern if raters would judge performance assessments differently in which students either typed (i.e., computer generated) their compositions or wrote them using conventional handwriting. Data were collected from a state-wide sample of typed and handwritten samples of essays for eighty students of which seven students were in special education and the remainder was in general education using raters selected from a pool of trained judges. Differences at different administration times and by writing "traits" were detected. The results pointed to overall scale instability for the scoring rubrics across response formats.

Huesman, R.L., & Frisbie, D. (2000, April). *The validity of ITBS reading comprehension test scores for learning disabled and non learning disabled students under extended-time*

conditions. Paper presented at the Annual Meeting of the National Council on measurement in Education. New Orleans, LA.

Conducted a quasi-experimental study to examine the effects of extended time on the performance of 129 students with learning disabilities (SWLD) and 397 students without disabilities on the ITBS Reading Comprehension Test. Concluded that SWLD had larger gains with extended-time than students without disabilities, and that extended time appears to promote test score validity for LD students.

Johnson, E. (2000). The effects of accommodations on performance assessments. *Remedial and Special Education, 21*, 261-267.

Math performance items from the 1997 and 1998 Washington Assessment of Student Learning (WASL) were subjected to a read aloud accommodation for one hundred and fifteen fourth grade students participated in which students without disabilities were assigned to one of two groups. A third group of 38 students consisted of special education students receiving services for reading disabilities. While all groups scored below standard on the math WASL, SWD scored considerably lower than the two general education groups. Math differences were noted between the two general education groups but no condition or interaction effect was found. A differential effect for reading the test for SWD was detected beyond the effect for general education student but the small sample size limits the capability to detect a significant effect for the accommodation.

Johnson, E., Kimball, K., & Brown, S. O. (2001). American Sign Language as an accommodation during standards-based assessments. *Assessment for Intervention, 26*, 39-47.

This study examined if (1) American Sign Language (ASL) changes the psychometric integrity of the "listening" portion of the Washington State Assessment of Student Learning, ASL. (2) if translation of the mathematics items into ASL affects the validity of the results, and (3) the practical implications of using ASL as an accommodation for students who are deaf or hard of hearing. Videotaped versions of the reading (listening) passages in ASL and SEE II were presented to students who answered 6 multiple choice and 2 open-ended questions. No text was provided to them. Mathematics items at each grade level were signed by certified ASL interpreters. Differences appeared as a function of the type of translation and grade (i.e., loss of information at grade 4, but not at grade 7) and whether a native ASL speaker did the translation (no loss). The ability of the interpreter was more influential in altering the information than the sign code itself.

Johnson, E., Kimball, K., Brown, S. O., & Anderson, D. (2001). A statewide review of the use of accommodations in large-scale, high stakes, assessments. *Exceptional Children, 67*, 251-264.

A post hoc analysis was conducted on the 1999 Washington State Assessment of Student Learning large-scale test results for special populations (i.e., Special Education, Section 504, Bilingual, and Migrant students) in reading and math, writing, and listening at the 4th

and 7th grades. Accommodations such as English, visual or native language dictionaries, scribes, large print or Braille versions and oral presentations were used. All special population students scored lower than students in general education, with special education students scoring the lowest. The general conclusion was that accommodations did not result in an unfair advantage to special education students.

Koretz, D. & Hamilton, L. (2000). Assessment of students with disabilities in Kentucky: Inclusion, student performance, and validity. *Educational Evaluation and Policy Analysis, 22*, 255-272.

This study was designed to examine how SWD were included in Kentucky's statewide assessment system and what were the effects of accommodations. Accommodations such as extra time, using scribes, oral presentation, paraphrasing, technology, interpreters, and separate sessions are among those allowed, with many students receiving more than one accommodation. Overall, the performance of SWD was generally lower than students with out disabilities with the differences increasing with grade level. Performance of SWD varied substantially according to their particular disability but small numbers of students in each category make it difficult to interpret these findings. Frequent substantially higher associations between the use of accommodations and test scores were reported but this finding was not consistent and at times yielded implausible scores and some scores may be untrustworthy based on the misuse of accommodations.

Koretz & Hamilton (2001). *The Performance of Students With Disabilities on New York's Revised Regents Comprehensive Examination in English.* Center for the Study of Evaluation, CSE Technical Report 540.

Explored performance of SWD in large field-test of NY regents exam. Looked at overall performance, completion rates, and item performance by type of accommodation. SWD performed about three-quarters of a standard deviation lower than students without disabilities. Although the completion rates for SWD and students without disabilities were comparable, they found that SWD who took the tests *without* accommodations had lower completion rates on the constructed-response items. Concluded "Across all accommodation conditions, additional time was associated with an increase of .13 standard deviation on the [constructed-response] portion of the test, but essentially no change on the MC portion" (p. 19).

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. Retrieved January 2003, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Technical28.htm

Examined the effects of a read aloud accommodation using a quasi-experimental design on 17 general education and 14 special education students in third through fifth grade in a suburban school district. The read aloud accommodation was provided using a standard

audiocassette player to maintain consistency between testing sessions. Two open-ended questions were asked at the end of the testing session to get an idea of student perception of and comfort level with the read aloud accommodation. A repeated-measure analysis of variance was conducted to determine whether there was an interaction between the test administration condition and disability status on students' test performance. Students without disabilities outperformed SWD under both test administration conditions. However, the gain for SWD in the accommodation condition was much larger, albeit not statistically significant (p=.06).

McKevitt, B. C., Elliot, S. N. (in press). The effects and consequences of using testing accommodations on a standardized reading test. *School Psychology Review*.

The purpose of the study was to analyze what accommodations teachers believe are valid and appropriate and what effect do teacher-recommended and packages of reading accommodations (more than one) have on students with and without disabilities. The effect of reading the reading test content on reading scores for students with and without disabilities and the perceived consequences of using a variety of accommodations on score validity and student attitude were also tested using an alternate treatment design. Teachers selected accommodations they consider to be valid and fair, with extra time selected most frequently, with "reading the directions" next. However, no teacher selected "reading the test content aloud" as a valid accommodation. The reading accommodation was not necessarily effective for SWD or their non-disabled peers.

McKevitt, B.C., Marquart, A.M., Mroch, A.A., Schulte, A., Elliott, S., & Kratochwill, T. (2000). *Understanding the effects of testing accommodations: a single-case approach*. A Paper Presented at the Annual Meeting of the National Association of School Psychologists. New Orleans, LA.

This study provides information about the use of a single approach for testing the effects of test accommodations on student test scores. Information about the test accommodations listed on student IEPs, accommodations actually used during testing, and the effects accommodations have on test score for students with and without disabilities are presented. An alternating treatment design was used in which all students were subjected to two different testing conditions receiving and not receiving accommodations. Effect sizes for differences between groups and testing conditions are reported.

Meloy, L., Deville, C., & Frisbie, D. (April, 2000). *The effects of a reading accommodation on standardized test scores of learning disabled and non learning disabled students*. A Paper Presented for the National Council on Measurement in Education Annual Meeting. New Orleans, LA.

Examined the effects of a read aloud accommodation on the test performance of 198 middle school students with a reading learning disability (LD-R) and 68 students without a disability who were randomly assigned to one of the two test administration conditions (read aloud or standard). The tests involved were the ITBS achievement tests in Science,

Usage and Expression, Math Problem-Solving and Data Interpretation, and Reading Comprehension. The results indicated that, on average, the LD-R students scored significantly higher under the read aloud accommodation, but this finding held for the students without disabilities too. The authors concluded that general use of the read aloud accommodation for LD students taking standardized achievement tests is not recommended.

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

Issues paper focusing on legal issues—when accommodations should be granted, flagging issues, etc.

Pitoniak, M, & Royer, J. (Spring 2001). Testing accommodations for examinees with disabilities: a review of psychometric, legal, and social policy issues. *Review of Educational Research*, *71 (1),* 53-104.

Literature review that focuses on psychometric, legal, and policy issues. Studies reviewed focused on test score comparability issue. Concludes that future legal decisions will determine assessment accommodations policies and that more research is needed on test comparability.

Ragosta, M. & Wendler, C. *Eligibility issues and comparable time limits for disabled and nondisabled SAT examinees.* College Board Report No. 92-5 (Report No.: ETS-RR-92-35). NY: College Board Publications.

Investigated appropriateness of time extensions for SAT by looking at SAT test takers from 1986-1988 of which about 17,600 had accommodations. Four-fifths of accommodations were for LD. Looked at completion rates for various time extensions. Also surveyed students. Non-equivalent groups who took standard or accommodated tests. Concluded that comparable testing time for SWD seems to be between 1.5 and 2 times standard time, except for Braille version which required 2-3 times more.

Rivera, C., Stansfield, C.W., Scialdone, L., & Sharkey, M. (2000). *An analysis of state policies for the inclusion and accommodation of English language learners in state assessment programs during 1998-1999.* Arlington, VA: George Washington University Center for Equity and Excellence in Education.

This is a descriptive study providing a nationwide picture of state practices in the 1998-1999 school year concerning inclusion, exemption, and accommodation of LEPs – the extent of inclusion and exclusion practices as well as the accommodations used. It is based on the direct analysis of state documents provided by Title VII Bilingual and English as a Second Language (ESL) directors in state education agencies.

Runyon, M. K. (1991). The effect of extra time on reading comprehension scores for university students with and without learning disabilities. *Journal of Learning Disabilities, 24*, 104-108.

Reading score differences between a small sample of college students with and without learning disabilities (LD) using extra time as an accommodation were tested. Higher mean percentile reading comprehension scores were detected for students with LD when given extra time but extra time had no significant effect on normally achieving students reading scores. No significant differences were found between students with LD reading scores when untimed with normally achieving students' scores when timed.

Scarpati, S., & Stoner, G. (2001). *Alternative assessment for vocational education*. Laboratory for Psychometric and Evaluative Research Report No. 419, Amherst, MA: University of Massachusetts, School of Education.

This study was conducted to compare how general and special education in vocational education Massachusetts compared to general education students on the 10th grade Massachusetts Comprehensive Assessment System (MCAS) test. Comparisons among students, levels of proficiency, and differential item functioning (DIF) were computed for students in vocational education (approximately 4300) and students in vocational education receiving special education services (approximately 1300). Overall, students in vocational education scored between one and two standard deviations below the means of students not in vocational education in all academic areas. Students in special education performed more poorly than their vocational education peers in all academic areas. The DIF analyses revealed virtually no significant item bias across each academic area of the MCAS.

Schulte, A. G., Elliot, S. N., & Kratchowill, T. R. (2001). Effects of testing accommodations on students' standardized mathematics test scores: An experimental analysis. *School Psychology Review, 30*, 527-547.

Fourth grade students with and without disabilities were allowed accommodations during a version of the TerraNova Multiple Assessment Battery in mathematics. Accommodations were selected for each individual SWD based on a review of their IEP. When possible, a student without a disability was matched and then given the same accommodation given to their yoked partner. Extra time and reading the test aloud (and their combination) constituted the most common accommodations. Total scores for SWD improved more between non-accommodated and accommodated test conditions than did their non-disabled peers. Other differences were detected for multiple choice but not constructed response items.

Shepard, L., Taylor, G., & Betebenner, D. (1998). *Inclusion of limited-English-proficient students in Rhode Island's grade 4 mathematics performance assessment*. Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Examined the effects of accommodations for ELL on the *Rhode Island Grade 4 Mathematics Performance Assessment* (MPA). Students' performance on the *Metropolitan Achievement Test* (MAT) was used for comparison purposes. Although both programs were mandated by the State, accommodations for ELL were provided on the MPA, but not on the MAT. The mean MAT score for ELL students with less than two years in the U.S. was 1.45 standard deviations below the mean for general education students, but the ELL mean on the MPA was 1.08 standard deviations below the mean for general education students. For ELL who had been in the US for more than two years, the mean MAT performance was 1.20 standard deviation below the general education mean, but their mean MPA was .92 standard deviations below the mean. Although these effect size differences are small, and there are content differences between the two tests, the relative improvement of ELL on the MPA could be due in part to the accommodations provided on this test that were not provided on the MAT.

Sireci, S.G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practices, 16 (1)*, 12-19.

Issues paper on difficulties in comparing examinees who took different (translated or accommodated) versions of a test. Concludes that it is difficult, if not impossible, to put scores from accommodated tests on the same scale. IRT and other methods are insufficient for such linking. Promising approaches include bilingual group designs and the use of nonverbal anchor items. Quality control in test translation/adaptation is particularly important.

Sireci, S.G., & Khaliq, S.N. (2002, April). *An analysis of the psychometric properties of dual language test forms*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Performed psychometric analyses on a dual-language (English-Spanish) version of a state-mandated 4th grade math test. Large group differences on total test performance b/w non-dual language and dual language groups were observed, with the DL group performing about 1.5 standard deviations lower. Before accounting for large differences in overall test performance, structural differences were pronounced and many items exhibited DIF. Most, but not all, of these differences disappeared when the analyses were repeated using matched samples of examinees. Concludes that benefits of dual-language accommodation are not clear. Also shows that we cannot assume such forms are equivalent to monolingual forms.

Thompson, Blount, & Thurlow (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. Retrieved January 2003, http://education.umn.edu/NCEO/OnlinePubs/Technical34.htm

This is a really terrific literature review on test accommodations for SWD. It does not include accommodations for LEP students. The authors identified 46 empirical research studies that addressed the effects of accommodations on test performance, although not all studies actually looked at effects. Sample sizes across the reviewed studies ranged from 3 to 21,000. The authors claimed that "three accommodations showed a positive effect on student test scores across at least four studies: computer administration, oral presentation, and extended time. However, additional studies on each of these accommodations also found no significant effect on scores or alterations in item comparability" (p.2). Good suggestions for future research.

Tindal, G., Anderson, L., Helwig, R., Miller, S., & Glasgow, A. (Sum 1998). Accommodating students with learning disabilities on math tests using language simplification, *Exceptional Children*, *64,* 439-50.

Applied simplified language to SWD rather than ELL. Two parallel math forms were developed, but rather than simplify same items one form was simplified. Thus, differences in difficulty across forms were not statistically adjusted using equating methods. The authors concluded that the original form was more difficult and so they deleted 10 items. It appears that an independent groups design was used but it is not clear whether students were randomly assigned to groups. Participants were 48 7th graders, mostly boys, low achievers, all fluent in English, 16 SWLD. Simplified language did not improve performance of either group (SWLD or non-SWD). Should be noted that study had very low power both in terms of sample size and "treatment" potency (16 items). No effect sizes were reported.

Tindal, G., & Fuchs, L. (1999). *Summary of research on test changes: An empirical basis for defining accommodations*. Lexington, KY: University of Kentucky Mid-south Regional Resource Center, Interdisciplinary Human Development Institute.

The document summarizes the research on changes to test administration in order to provide an empirical basis to various test accommodations. Research studies were synthesized and organized according to types of accommodations. These were: changes in schedule, presentation, test directions, use of assistive devices/supports, and test setting. The review concludes that research on accommodations is just beginning and that policy makers use the report to draw conclusions about accommodations used in their particular setting. The authors also suggest that research on accommodations take the form of experimental rather than descriptive or comparative studies.

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.

Investigated the effects of oral accommodation (on a fourth grade math test)and response format (fourth grade reading test). The response format investigated was recording answers in the test booklet versus on an answer sheet. The study involved 481 fourth grade students, 84% were students without disabilities. There were 36 SWD who took the reading test and 38 SWD who took the math test. The results showed no effect for the response format condition. Also, students without disabilities outperformed SWD under both standard and oral accommodation conditions. However, there was a significant improvement in scores for SWD under the oral accommodation condition for SWD (effect size of .76), but not for the other student group (negative effect size of .20).

Trimble, S. (1998). *Performance Trends and Use of Accommodations on a Statewide Assessment: Students with Disabilities in the KIRIS On-Demand Assessments from 1992-93 through 1995-96*. State Assessment Series, Maryland/Kentucky Report 3. National Center on Educational Outcomes, Minneapolis, MN; Maryland State Dept. of Education, Baltimore.

Documented the performance of grade 4, 8, and 11 SWD on KIRIS tests from 1992-1995. The most frequently used accommodations were combinations that included paraphrasing and oral presentations. The author found that, as a cohort, SWD who took the test with one or more accommodations showed greater gains across years than SWD who took the test without accommodations or students without disabilities.

Walz, L., Albus, D., Thompson, S., & Thurlow, M. (December 2000). *Effect of a Multiple Day Test Accommodation on the Performance of Special Education Students*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [January 3, 2003], from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/MnReport34.html

Evaluated the effect of a multi-day accommodation using a sample of 112 (48 SWD students and 64 genera education students) 7th and 8th graders from two rural and two urban schools in Minnesota. The test items came from a statewide test in Minnesota and all students took the test both in a single-day administration and a two-day period. The students without disabilities outperformed the SWD under both conditions and neither student group exhibited meaningful gains under the multiple-day condition. The results did not support the use of a multiple-day accommodation for improving the scores of SWD.

Weston (2002, July). The validity of oral accommodation in testing. NAEP Validity Studies (NVS) Panel.

Evaluated the interaction hypothesis with respect to the accommodation of oral administration. Tested two groups of fourth grade students: 65 SWLD and 54 students without disabilities. Both groups took two parallel NAEP math test forms. One test form was taken under standard conditions, the other with oral presentation. Although both student groups exhibited gains in the accommodation condition, the SWLD had significantly greater gain. The effect size for SWLD was .64. For students without

disabilities, the effect size was .31. Although students without disabilities also benefited from the accommodation, the significantly larger gain for SWLD lends some support to the interaction hypothesis.

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

This is a book that came out of several studies done at ETS between 1982 and 1986 to look at the comparability of scores from accommodated and non-accommodated SAT and GRE tests. Many of the chapters exist as individual ETS reports. Looked at 4 principal groups: hearing impaired, LD, physically handicapped, visually impaired. Data came from thousands of non-SWD, hundreds of students with physical, visual, or hearing disabilities, and thousands of LD aggregated over several years. Looked at reliability differences, factor structures, DIF, and differential predictive validity. Concluded that nonstandard versions of SAT and GRE administered to SWD are generally comparable to standard tests in most respects: similar reliability, factor structure, no DIF, and similar predictive validity. Lower-scoring SWD tended to be slightly underpredicted and higher scoring SWD tended to be overpredicted. An excellent set of studies.

Ziomek, R.L., & Andrews, K.M. (1998). *ACT Assessment Score Gains of Special-Tested Students Who Tested at Least Twice*. Report No.: ACT-RR-98-8. Iowa city, Iowa: ACT Research Report Series.

Looked at SWD who took the ACT either (a) twice w/ extended time, (b) first with standard time and second with extended time, or (c) first with extended time and then with standard time. Standard-first/extended-second group had largest score gain (3.2 scale score points) compared with extended-extended group who had gain of .9 points, which was similar to general re-testers (.7). Third group declined .6 on retest. Note: ACT composite SEM is 1 point.

Zurcher, R., & Bryant, D.P. (2001) The validity and comparability of entrance examination scores after accommodations are made for students with LD. *Journal of Learning Disabilities, 34 (5)*, Sep-Oct 2001, 462-471.

Using a sample of college students with LD and a control sample, the authors investigated the effects of testing accommodations in general for students with LD on the criterion validity and the comparability of scores from a standardized, norm-referenced entrance examination – the *Miller Analogies Test (AMT)*. The focus was on the comparability of scores after accommodations are made. Their results indicated no significant improvement for either group under the accommodation condition (scores across conditions differed by .06 of a point for SWLD and by .14 for students without disabilities). Thus, this study did not support the interaction hypothesis. However, there were several methodological limitations of this study.

Zuriff, G.E. (2000). Extra examination time for students with learning disabilities: An examination of the maximum potential thesis. *Applied Measurement in Education, 13 (1)*, 99-117.

Examined "maximum potential thesis" which states that non-SWD would not benefit from extra time accommodation they are already working at their maximum potential under timed conditions (p. 101). Evaluated 5 studies that looked at this question—4 studies were unpublished dissertations. General finding was that extra time helped both students with learning disabilities and students without disabilities.

Appendix B

List of Initial Citations (Unformatted)

Abedi, J. (1999). *NAEP Math* Test Accommodations *for Students with Limited English Proficiency*. Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).

Abedi, J. (2001). *Assessment and Accommodation for English language learners: Issues and recommendation* (Policy Brief 4). Los Angeles: University of California, Los Angeles, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J. (1999). *The Impact of Students' Background Characteristics on Accommodation Results for Students with Limited English Proficiency*. Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999). ED431786

Abedi, J. (2001, July). *Examining ELL and non-ELL student performance differences and their relationship to background factors: Continued Analyses of Extant Data.* National Center for Research on Evaluation, Standards, and Student Testing; Center for the Study of Evaluation; Graduate School of Education & Information Studies; University of California; Los Angeles, CA

Abedi, J., (2001, July). *Impact of selected background variables on students' NAEP math performance.* National Center for Research on Evaluation, Standards, and Student Testing; Center for the Study of Evaluation; Graduate School of Education & Information Studies; University of California; Los Angeles, CA.

Abedi, J. (2001, April). Validity of Accommodation for English Language Learners. Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).

Abedi, J., Leon, S., & Mirocha, J. (no date). *Impact of students' language background on content-based performance: Analyses of extant data.* National Center for Research on Evaluation, Standards, and Student Testing; Center for the Study of Evaluation; Graduate School of Education & Information Studies; University of California; Los Angeles, CA.

Abedi, J., & Lord, C. (2001). The Language Factor in Mathematics Tests. *Applied Measurement in Education, 14(3), 219-234.*

Abedi, J., Lord, C., Hofstetter, C., and Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice*, *19*(3), 16-26.

Abedi, J., Lord, C., Hofstetter, C., and Baker, E. (2000). NAEP Math Performance and Test Accommodations: Interactions With Student Language Background. *CRESST*

Abedi, J., Lord, C., Kim, C., and Miyoshi, J. (2000). *The effects of accommodations on the assessment of LEP students in NAEP*. Los Angeles: University of California, Los Angeles, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J.,Lord, C., Koscardin, C.K, &Miyoshi J. (2001). The Effects of Accommodations on the Assessment of Limited English Proficient (LEP) Students in the National Assessment of Educational Progress (NAEP). *CRESST.*

Abedi, J., Lord, C., & Plummer, J.R. (2001, July). *Language background as a variable in NAEP mathematics performance.* National Center for Research on Evaluation, Standards, and Student Testing; Center for the Study of Evaluation; Graduate School of Education & Information Studies; University of California; Los Angeles, CA.

Barton, K. & Huynh Huynh. (April 2000). Patterns of Errors Made on a Reading Test with Oral Reading Administration. Paper presented at the National Council on Measurement in Education Conference New Orleans, Louisiana.

Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J., & Freidebach, M. (2001). *Read-aloud accommodations: Effects on multiple-choice reading and math items* (TechnicalReport 31). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved January 2003, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Technical31.htm

Bielinski, J. and Ysseldyke, J. (2000). Interpreting trends in the performance of special education students (Technical Report 27). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Braun, H., Ragosta, M., & Kaplan, B. (1986a). *The predictive validity of the Scholastic Aptitude Test for disabled students* (Research Report 86-38). New York: College Entrance Examination Board.

Brown, P.J., & Augustine, A. (Apr 2001). *Screen Reading Software as an Assessment Accommodation: Implications for Instruction and Student Performance*. Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).

Calhoon, M. B., Fuchs, L.S., & Hamlett, C.L. (Fall 2000). Effects of computer-based test accommodations on mathematics performance assessments for secondary students with learning disabilities. *Learning Disability Quarterly*, *v23* n4 p271-82.

Castellon-Wellington, M. (1999). *The Impact of Preference for Accommodations: The Performance of English Language Learners on Large-Scale Academic Achievement Tests.*

Chiu, C. WT, & Pearson, P D. (Jun 1999*). Synthesizing the Effects of Test Accommodations for Special Education and Limited English Proficient Students*. Paper presented at the National Conference on Large Scale Assessment (Snowbird, UT, June 13-16, 1999). 51p. American Institutes for Research in the Behavioral Sciences, Palo Alto, CA

Disability Rights Advocates. Do Not Harm- High Stakes Testing and Students with Learning Disabilities. 2001.

Elliott, S.N. (no date). *Testing accommodations: Legal and technical issues challenging educators or "good" test scores are hard to come by.*

Elliott, S. (1999). Valid test accommodations: Fundamental assumptions and methods for collecting validity evidence. Wisconsin Center for Educational Research.

Elliott, S.N., Kratochwill, T.R., & McKevitt, B.C. (2001). Experimental analysis of the effects of testing accommodations on the scores of students with and without disabilities. Wisconsin Center for Education Research and University of Wisconsin-Madison. *Journal of School Psychology*, Vol. 39, No. 1, pp. 3-24.

Elliot. J.L., Ysseldyke, J.E., Thurlow, M.L. (1998). What about assessment and accountability. *Teaching Exceptional Children*, 31, 20-17.

Fuchs, L.S., & Fuchs, D. (Aug 2001). Helping teachers formulate sound test accommodation decisions for students with learning disabilities. *Learning Disabilities: Research & Practice,* 16(3) p174-81 Learning Disabilities: Research & Practice; v16 n3 p174-81.

Fuchs, L.S., Fuchs, D., Eaton, S.B., Hamlett, C., Binkley, E. & Crouch, R. (Fall 2000). Using objective data sources to enhance teacher judgments about test accommodations. Exceptional Children; v67 n1 p67-81.

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

Grise, P., Beattie, S., and Algozzine, B. (1982). Assessment of minimum competency in fifth grade learning disabled students: Test accommodations make a difference. *Journal of Educational Research, 76*(1), 35-40.

Hall, J., Griffin, H., Cronin, M., & Thompson, B. (1985). Factors related to competency test performance for high school learning disabled students. (Summer, 1985). *Educational Evaluation and Policy Analysis. 7*(2), 151-160.

Hambleton, R. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10*(3), 229-244.

Helwig, R., Stieber, S., Tindal, G., Hollenbeck, K., Heath, B., & Almond, P. (2000). *A comparison of factor analyses of handwritten and word-processed writing of middle school students.*

Hollenbeck, K., Rozek-Tedesco, M. A., Tindal, G. & Glasgow, A. (Spr 2000). An exploratory study of student-paced versus teacher-paced accommodations for large-Ssale math tests. (Spr 2000). Journal of Special Education Technology; v15 n2 p27-36.

Hollenbeck, K., Tindal, G., Stieber, S., & Harniss, M. (1999). *Handwritten versus word processed statewide compositions: Do judges rate them differently?*

Huesman, R.L. & Frisbie, D. (April 2000). *The validity of ITBS reading comprehension test scores for learning disabled and non learning disabled students under extended-time conditions.* Paper presented at the Annual Meeting of the National Council on measurement in Education. New Orleans, LA.

Johnson, E. (no date). The effects of accommodations on performance assessments. Remedial & Special Education. Vol 21(5), Sep-Oct 2000, pp. 261-267.

Johnson, E., Kimball, K., & Brown, S.O. (2001) American sign language as an accommodation during standards-based assessments. *Assessment for Effective Intervention*; v26 n2 p39-47.

Kiplinger, V.L., Haug, C.A., & Abedi, J.(2000). Measuring Math—Not Reading—on a Math Assessment: A Language **Accommodations** Study of English Language Learners and Other Special Populations. Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).

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

Kopriva, R.J., and Lowrey, K. (1994*). Investigation of language sensitive modifications in a pilot study of CLAS, the California Learning Assessment System* (Technical Report). Sacramento, CA: California Department of Education, California Learning Assessment System Unit.

Koretz, D., and Hamilton, L. (1999). Assessing students with disabilities in Kentucky: The Effects of Accomodations, Format, and Subject. *Center for the Study of Evaluation, CSE Technical Report 498.*

Koretz, D., and Hamilton, L. (2000). Assessing students with disabilities in Kentucky: Inclusion, student Performance, and validity. *Educational Evaluation and Policy Analysis, 22*(3), 255-272.

Koretz, D., & Hamilton, L. (April 2001). The Performance of Students With Disabilities on New York's Revised Regents Comprehensive Examination in English. *Center for the Study of Evaluation, CSE Technical Report 540.*

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. Retrieved January 2003, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Technical28.htm

Marquart, A. & Elliot, S.N. (2000). *Extra time as an accommodation.*

Mazzeo, J., Carlson, J.E., Voelkl, K.E., and Lutkus, A.D. (2000). Increasing the participation of special needs students in NAEP: A report on 1996 NAEP research activities.

McDonnell, L.M., McLaughlin, M.J., & Morison, P. (1997). EDS, Educating One & All.

McKevitt, B.C. & Elliot, S.N. (1999). What we have learned about the use of testing accommodations. *Inclusive Education Programs*, 7(1), LRP Publications. Available http://web.lexis-nexis.com

McKevitt, B.C., & Elliott, S.N. (2001, September). *The effects and consequences of using test accommodations on a standardized reading test.* Manuscript submitted for publication.

McKevitt, B.C., Marquart, A.M., Mroch, A.A., Schulte, A., Elliott, S., & Kratochwill, T. (2000). Understanding the effects of testing accommodations: a single-case approach. A Paper Presented at the Annual Meeting of the National Association of School Psychologists. New Orleans, LA.

Meloy, L., Deville, C., & Frisbie, D. (April, 2000). *The effects of a reading accommodation on standardized test scores of learning disabled and non learning disabled students*. A Paper Presented for the National Council on Measurement in Education Annual Meeting. New Orleans, LA.

National Center for Education Statistics. (2000). *Becoming a more inclusive NAEP*. Available: <http://nces.ed.gov/nationsreportcard/pubs>. [May 17, 2002].

National Institute of Statistical Sciences. (2000, July). NAEP inclusion strategies: The report of a workshop at the National Institute of Statistical Sciences, July 10-12.

National Research Council (2000). *Testing English language learners in U.S. schools: Report and workshop summary*. Kenji Hakuta and Alexandra Beatty, Eds. Board on Testing and Assessment, Center for Education. Washington, DC: National Academy Press.

National Research Council. (2001). *NAEP reporting practices: Investigating district level and market-basket reporting.* P.J. DeVito and J.A. Koenig, (Eds.). Washington DC: National Academy Press.

National Research Council (1997). *Educating One and All: Students with disabilities and standards-based reform*. Lorraine M. McDonnell, Margaret J. McLaughlin, and Patricia Morison, Eds. Board on Testing and Assessment, Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Olson, J.F., and Goldstein, A.A. (1997). *The inclusion of students with disabilities and limited English proficiency students in large-scale assessments: A summary of recent progress* (NCES 97-482). Washington, DC: National Center for Education Statistics.

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

Pitoniak, M, & Royer, J. (Spring 2001). Testing accommodations for examinees with disabilities: a review of psychometric, legal, and social policy issues. *Review of Educational Research. 71* (1), 53-104.

Quenemoen, R.F., Lehr, C.A., Thurlow, M.L. and Massanari, C.B. (2001). Students with disablities in standards-based assessment and accountability systems: Emerging issues, strategies, and recommendations (Synthesis Report 37). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Ragosta, M.& Kaplan, B.A. (1986). A Survey of Handicapped Students Taking Special Test Administrations of the SAT and GRE. Report No. 5, Studies of Admissions Testing and Handicapped People (Report No.: ETS-RR-86-5). Princeton, N.J.: College Entrance Examination Board.

Ragosta, M.& Wendler, C. (1992). Eligibility Issues and Comparable Time Limits for Disabled and Nondisabled SAT Examinees. College Board Report No. 92-5 (Report No.: ETS-RR-92-35). NY: College Board Publications.

Rivera, C., and Stansfield, C.W. (2001, April). *The effects of linguistic simplification of science test items on performance of limited English proficient and monolingual English-speaking students*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Rivera, C., Stansfield, C.W., Scialdone, L., and Sharkey, M. (2000). *An analysis of state policies for the inclusion and accommodation of English language learners in state assessment programs during 1998-1999*. Arlington, VA: George Washington University Center for Equity and Excellence in Education.

Runyan, M. K. (1991). The effect of extra time on reading comprehension scores for university students with and without learning disabilities. *Journal of Learning Disabilities,* 24, 104-108.

Scarpati, S. & Stoner, G. (2001). Alternative assessment for vocational education. *Laboratory of Psychometric and Evaluative Research Report No. 419*, Amherst, MA: University of Massachusetts, School of Education.

Schulte, A.A., Elliott, S.N. & Kratochwill, T.R. (2001, March*). Effects of testing accommodations on standardized mathematics test scores: An experimental analysis of the performances of students with and without disabilities.* Manuscript submitted for publication.

Schulte, A. A.,Gilbertson. (2001).Experimental analysis of the effects of testing accommodations on the students' standardized mathematics test scores. Dissertation Abstracts International Section A: Humanities & Social Sciences. Vol 61(8-A), Mar 2001, pp. 3056

Shepard, L., Taylor, G., & Betebenner, D. (1998). *Inclusion of limited-English-proficient students in Rhode Island's grade 4 mathematics performance assessment.* Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Shriner, J.G. (2000). Legal perspectives on school outcomes assessment for students with disabilities. The Journal of Special Education, 33, 232-239.

Sireci, S.G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practices, 16*(1), 12-19.

Sireci, S.G., & Khaliq, S.N. (2002, April). An analysis of the psychometric properties of dual language test forms. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Siskind, T. G. (Apr 1993). The Instructional Validity of Statewide Criterion-Referenced Tests for Students Who Are Visually Impaired. *Journal of Visual Impairment and Blindness; v87* n4 p115-17

Pomplun, M., & Omar, M. H.. (2001).Score comparability of a state reading assessment across selected groups of students with disabilities. *Structural Equation Modeling*. Vol 8(2), 2001, pp. 257-274.

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. Retrieved January 2003, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Technical34.htm

Thurlow, M., & Bolt, S. (2001). *Empirical support for accommodations most often allowed in state policy* (Synthesis Report 41).Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved January 2003, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis41.html

Thurlow, M.L., Elliot, J.L., Scott, D.L., and Shin, H. (1997). An analysis of state approaches to including students with disabilities in assessments implemented during educational reform (Technical Report No. 18). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M.L., McGrew, K.S., Tindal, G. Thompson, S.L., Ysseldyke, J.E. and Elliott, J.L. (2000). Assessment accommodations research: Considerations for design and analysis (Technical Report 26). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, Martha L. & Others. (Mar 1993). *Testing **Accommodations** for Students with Disabilities: A Review of the Literature. Synthesis Report 4*. Special Education Programs (ED/OSERS), Washington, DC. ED358656

Thurlow, M., Ysseldyke, J., Bielinski, J., House, A., Trimble, S., Insko, B. & Owens, C. (Jan 2000). Instructional and assessment accommodations in Kentucky State assessment series, Maryland/Kentucky Report 7. Special Education Programs (ED/OSERS), Washington, DC.; Maryland State Dept. of Education, Baltimore.

Tindal, G., Anderson, L., Helwig, R., Miller, S., & Glasgow, A. (Sum 1998). *Accommodating students with learning disabilities on math tests using language simplification.* Exceptional Children. V64 n4 p439-50.

Tindal, G. & Fuchs, L. (2000) A summary of research on test changes: an empirical basis for defining accommodations. Commissioned by the Mid-South Regional Resource Center Interdisciplinary Human Development Institute.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. *Exceptional Children*, Vol. 64, No. 4, pp. 439-450.

Trimble, S. (1998). *Performance Trends and Use of Accommodations on a Statewide Assessment: Students with Disabilities in the KIRIS On-Demand Assessments from 1992-93 through 1995-96. State Assessment Series, Maryland/Kentucky Report 3*. National Center on Educational Outcomes, Minneapolis, MN.; Maryland State Dept. of Education, Baltimore.

Walz, L., Albus, D., Thompson, S., & Thurlow, M. (December 2000). Effect of a Multiple Day Test Accommodation on the Performance of Special Education Students. *National Center on Educational Outcomes, Minnesota Report 34.*

Weston, T.J. (2002). *The validity of oral accommodation in testing.* NAEP Validity Studies (NVS) Panel.

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

Ysseldyke, J., Thurlow, M., Bielinski, J., House, A., Moody, M., & Haigh, J. (2001). The relationship between instructional and assessment accommodations in an inclusive state accountability system. *Journal of Learning Disabilities, 34* (3), 212-220.

Ziomek, R.L., & Andrews, K.M. (1998). ACT Assessment Score Gains of Special-Tested Students Who Tested at Least Twice (Report No.: ACT-RR-98-8). Iowa city, Iowa: ACT Research Report Series.

Zurcher, R., & Bryant, D.P.. (2001) The validity and comparability of entrance examination scores after accommodations are made for students with LD. *Journal of Learning Disabilities*. Vol 34(5), Sep-Oct 2001, pp. 462-471

Zuriff, G.E. (2000) Extra examination time for students with learning disabilities: An examination of the maximum potential thesis, *Applied Measurement in Education*, 13 (1), 99-117.