Jurnal Pendidik dan Pendidikan, Jil. 23, 81–110, 2008

VALIDITY ISSUES IN ACCOMMODATING READING TESTS[1]

Stephen G. Sireci
School of Education, 156 Hills South,
University Of Massachusetts, Amherst, 01003, USA
E-mail: sireci@acad.umass.edu

Abstract: National Assessments seeks to include all students in the sampling frame from which students are selected to participate in the assessment. However, some students with disabilities (SWD) are either unable to take tests under standard testing conditions or are unable to perform at their best under standard testing conditions. In many testing situations, accommodations to standard testing conditions are given to SWD to improve measurement of their knowledge, skills and abilities. This practice is in the pursuit of more valid test score interpretation; however, it produces the ultimate psychometric oxymoron – an accommodated standardized test. In this paper, I review validity issues related to test accommodations and summarize some empirical studies in this area. The focus of the paper is on accommodations for reading tests because some types of accommodations on these tests are particularly controversial. The specific accommodations emphasized in this review are extended time and oral (read-aloud) accommodations. A review of professional standards, validity theory, and recent empirical research in this area suggests that extended time accommodations may be appropriate for reading tests, but read-aloud accommodations are likely to alter the construct measured. Suggestions for determining when to provide accommodations and how to report scores from accommodated test administrations are provided.

Keywords: construct validity, reading tests, test accommodations, validity

Abstrak: Pentaksiran Kebangsaan berhasrat melibatkan semua pelajar Amerika Syarikat dalam kerangka pemilihan sampel yang mengambil bahagian dalam pentaksiran. Namun, pelajar yang kurang upaya (SWD) mungkin tidak dapat menduduki ujian atau menunjukkan kebolehan mereka di bawah keadaan piawai ujian. Dalam kebanyakan keadaan, akomodasi diberi kepada SWD untuk memperbaiki pengukuran pengetahuan, kemahiran dan kebolehan. Amalan ini dilakukan supaya pentafsiran skor akan mempunyai kesahan yang lebih baik. Hasilnya ialah satu ujian piawai dengan akomodasi. Dalam kertas kerja ini, saya meninjau isu kesahan berkaitan akomodasi dan merumuskan daripada kajian empirik dalam bidang ini. Fokus kertas kerja ini adalah pada akomodasi untuk ujian membaca yang menimbulkan kontroversi. Jenis akomodasi yang ditumpukan ialah melanjutkan masa ujian dan akomodasi lisan. Tinjauan standard profesional, teori

[1] Center for Educational Assessment Research Report No. 515. Amherst, MA: School of Education, University of Massachusetts, Amherst. This paper was commissioned by the National Assessment Governing Board as part of the Conference on Increasing the Participation of SD and LEP student's in NAEP, February 26–27, 2004.

kesahan dan kajian empirik mencadangkan bahawa akomodasi melanjutkan masa mungkin sesuai untuk ujian membaca; tetapi akomodasi lisan akan mengubah gagasan yang diukur. Cadangan tentang cara dan masa yang sesuai untuk melaporkan skor ujian akomodasi juga dibincangkan.

Kata kunci: kesahan binaan, ujian membaca, akomodasi ujian, kesahan

INTRODUCTION

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 (SWD). 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, I discuss the psychometric issues related to test accommodations with a particular focus on accommodations for reading tests. Focusing on reading tests illuminates many controversial issues, because some accommodations, such as reading test material aloud to a student, may dramatically change the construct measured by the test.

For example, when reading test material is presented orally to a student, many fear the construct changes from "reading comprehension" to "oral comprehension."

PROVIDING ACCOMMODATIONS TO PROMOTE VALIDITY

One of the most authoritative validity theorists, Samuel Messick, summarized threats to the validity of interpretations based on test scores as coming from two sources: "construct under-representation" or "construct-irrelevant variance." As he put it "Tests are imperfect measures of constructs because they either leave out something that should be included…or else include something that should be left out, or both" (Messick, 1989: 34). Construct under-representation refers to the situation where a test measures only a portion of the intended construct (or content domain) and leaves important knowledge, skills and abilities untested. Construct-irrelevant variance refers to the situation where the test measures proficiencies irrelevant to the intended construct Examples of construct-irrelevant variance undermining test score interpretations are when computer proficiency affects performance on a computerized mathematics test, or when familiarity with a particular item format (e.g., multiple-choice items) affects performance on a reading test.

Test accommodations are often provided to address the problem of construct-irrelevant variance that may arise as a consequence of standardized testing conditions. In testing, "standardized" means that the test content, scoring and administration conditions are uniform for all test takers. The concept of standardization stems from the scientific method and the procedures used by the earliest scientific psychologists such as Wundt, Weber and Fechner. The idea behind standardization is to keep the measurement instrument and observation conditions constant so that any differences observed reflect true individual differences, rather than measurement artifacts. Although elegant from a research design perspective, standardization introduces a lack of authenticity into the measurement process, which provides fertile ground for construct-irrelevant variance to propagate. Therefore, the provision of test accommodations is often granted in the pursuit of more valid test score interpretations.

If the conditions of a standardized test administration prevent some students from demonstrating their knowledge and skills, those conditions may be considered barriers to valid assessment. For example, 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. Removing those barriers, which is tantamount to accommodating the administration, is therefore, seen as removing construct-irrelevant variance and increasing test validity.

The flipside of this issue is that an accommodation may also introduce construct-irrelevant variance, if the accommodation changes the construct measured. If the construct intended to be measured by a test changes, and the new attributes measured represent a different and unintended construct, then construct-irrelevant variance is also present. Therefore, although test accommodations are often granted in the pursuit of test fairness, the degree to which the accommodation promotes validity is directly related to the degree to which the accommodation alters the construct measured. Thus, the "construct equivalence" of standard and accommodated test scores is a fundamental psychometric issue in evaluating the validity of a particular accommodation for a particular student.

Psychometric Issues in Test Accommodations

Psychometric issues in test accommodations stress the need to remove construct-irrelevant barriers to test performance while maintaining integrity to the construct being measured. Several excellent discussions of these issues appear in the published literature (e.g., Geisinger, 1994; Green & Sireci, 1999; Koretz & Hamilton, 2000; Phillips, 1994; Pitoniak & Royer, 2001; Scarpati, 1991; Sireci & Geisinger, 1998; Willingham et al., 1988), and these issues have been discussed in extensive detail in the current and previous versions of the "Standards for Educational and Psychological Testing" (APA, AERA & NCME, 1985; AERA, APA & NCME, 1999. The validity of scores from accommodated tests rests on the following issues:

1. Does providing a particular accommodation to a particular student improve measurement of that student’s knowledge, skills and abilities?

2. Does providing a particular accommodation to some, but not all, students unfairly advantage the students who receive the accommodation?

3. Does providing a particular accommodation change the construct the test is measuring?

4. Are scores from accommodated and standard test administrations comparable? That is, can they be interpreted as if they are on the same scale?

Answering "yes" to the first and last question, and "no" to the second and third question, means the test accommodations are valid from a psychometric perspective. However, these questions are complex. For example, an accommodation may facilitate valid score interpretation for some students (a "yes" to the first question) but simultaneously provide an unfair advantage, relative to students who do not receive the accommodation (a "yes" to the second question). Furthermore, there are many different types of accommodations and some students may receive more than one accommodation on a single test. To illustrate the complexities involved with these issues, I will start with the third question regarding the "construct equivalence" of scores from accommodated and non-accommodated tests.

DO TEST ACCOMMODATIONS CHANGE THE CONSTRUCT MEASURED?

The term "construct" has an important meaning in educational testing because it emphasizes the fact that we are not measuring tangible attributes of students. Educational tests attempt to measure students' knowledge, skills and abilities. Given this endeavor, it must be assumed that (a) such concepts exist within students and (b) they are measurable. Since we do not know for sure if such intangible student attributes or proficiencies really exist, we admit they are "constructs"; they are hypothesized attributes we believe exist within students. Hence, these attributes were "constructed" from educational and psychological theories, and they are subsequently operationally defined using test specifications and other elements of the testing process.

Although the current version of the Standards for Educational and Psychological Testing (AERA, APA & NCME, 1999) merely defines a construct as "the concept or characteristic that a test is designed to measure" (p. 173), its definition of "construct validity" provides greater insight into the importance of the construct in interpreting test scores. The "Standards" borrow from Messick (1989), Loevinger (1957), and other validity theorists to underscore the notion that validity refers to inferences about constructs that are made on the basis of test scores. In fact, many validity theorists describe "construct validity" as equivalent to validity in general. According to the "Standards" construct validity is:

> A term used to indicate that the test scores are to be interpreted as indicating the test taker's standing on the psychological construct measured by the test. A construct is a theoretical variable inferred from multiple types of evidence, which might include the interrelations of the test scores with other variables, internal test structure, observations of response processes, as well as the content

> of the test. In the current standards, all test scores are viewed as measures of some construct, so the phrase is redundant with validity. The validity argument establishes the construct validity of a test (AERA, APA & NCME, 1999: 174).

The construct measured by a test sets the basis for evaluating its utility as well as evaluating the validity of the interpretations that are made on the basis of its scores. For this reason, a fundamental step in educational testing is clearly defining the construct measured. All subsequent test construction steps strive to be faithful to this construct. Developing test specifications, writing items, screening items for differential item functioning, and determining the conditions under which the test is to be administered are just some examples of how construct concerns permeate all test development and validation. Therefore, it is no surprise that when accommodations are suggested on a standardized test, a major concern is that the accommodation might change the hallowed construct.

The "Standards" are clear on the importance of evaluating whether test accommodations alter the construct measured. The first standard in the chapter on testing individuals with disabilities reads "In testing individuals with disabilities, test developers, test administrators and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement" (AERA, APA & NCME, 1999). This standard provides justification for granting accommodations to obtain more valid measures of students' proficiencies, but it also underscores the notion that if an accommodation alters the construct measured, scores from accommodated tests cannot have the same meaning as scores from standardized administrations. The key question then is "When does an accommodation change the construct?"

Unfortunately, the "Standards" provide only limited guidance on this issue. Essentially, they require testing agencies to use logical and empirical methods to determine whether an accommodation alters the construct measured. Furthermore, the "Standards" acknowledge that empirical studies are not practical in many situations due to small numbers of SWD who take accommodated tests and the variety of accommodations provided. The "Standards" settle the issue by recommending that "cautionary statements," or "flags" accompany test scores when there is no evidence that scores from accommodated tests are "comparable" to scores from standard administrations. For example, Standard 10.4 reads:

> If modifications are made or recommended by test developers for test takers with specific disabilities…Unless evidence of validity for a given inference has been established for individuals with the specific disabilities, test developers should issue cautionary statements in manuals or supplementary materials regarding confidence in interpretations based on such test scores (AERA, APA & NCME, 1999: 106).

Elaborating on the concept of issuing cautionary statements if accommodations may affect the construct measured, Standard 10.11 states:

> When there is credible evidence of score comparability across regular and modified administrations, no flag should be attached to a score. When such evidence is lacking, specific information about the nature of the modification should be provided, if permitted by law, to assist test users properly to interpret and act on test scores (p. 108).

An excerpt from the comment accompanying this standard is also relevant here:

> If a score from a modified administration is comparable to a score from a nonmodified administration, there is no need for a flag. Similarly, if a modification is provided for which there is no reasonable basis for believing that the modification would affect score comparability, there is no need for a flag (p. 108).

Clearly, the issue of when to flag test scores centers on whether the accommodation changes the construct measured. Furthermore, it is clear AERA, APA and NCME (1999) recommend (a) when there is no reason to believe a modification would alter the construct, no flag is necessary; (b) when there is clear evidence of "score comparability" across scores from accommodated and non-accommodated test administrations, no flag is necessary; and (c) when such evidence is lacking, information should be provided to indicate a non-standard administration.

What is not clear from the "Standards" is how much "credible evidence of score comparability" is required to determine the construct has not been changed and scores should not be flagged. That is, how much evidence is needed before one can conclude scores from accommodated and non-accommodated tests can be interpreted similarly?

Studies Assessing Construct Equivalence of Accommodated Tests

Methods for evaluating construct equivalence, and hence comparability of scores from standard and accommodated tests include (a) comparing the dimensionality (factor structure) of test data from standard and accommodated administrations; (b) comparing the relationship between scores from accommodated and standard tests to external criteria (e.g., differential predictive validity studies); and (c) conducting experimental studies where SWD (and sometimes students without disabilities) are tested under both standard and accommodated conditions (Sireci, 2003; Thompson, Blount & Thurlow, 2002).

There have been many studies evaluating construct equivalence by using exploratory factor analysis, confirmatory factor analysis, or multidimensional scaling to look at the consistency of test structure across standard and accommodated versions of tests. Several studies involved tests translated into a second language (e.g., Allalouf, Hambleton & Sireci, 1999; Sireci & Gonzalez, 2003), bilingual test administrations (Sireci & Khaliq, 2002), or quantitative and verbal reasoning tests used for post-secondary admissions (Rock et al., 1988). The logic motivating these studies is that if the factor structures of data from accommodated and standard test administrations were the same, some evidence of construct equivalence is provided.

Although factor-analytic and other dimensionality studies partly address construct equivalence, very few of these studies have been conducted on reading tests. One study, by Huesman and Frisbie (2000) used exploratory factor analysis on small samples of students with learning disabilities and students without disabilities tested with and without extended time. Under standard time conditions they found two factors fit the data for all groups. Under the extended time condition, the second factor disappeared for the non-disabled students, but remained for the students with disabilities. Although this finding could indicate differential speededness, interpretation of these results is hindered by the fact that there were less than 100 students in each group and the analysis was exploratory rather than confirmatory. In another study, Tippets and Michaels (1997, cited in Bielinski et al., 2001) used confirmatory factor analysis to study the consistency of the factor structures of a reading test and a language usage test across standard and read-aloud administrations. They concluded a two-factor model fit both accommodated and standard administration data, thus supporting the idea that the read-aloud accommodation did not change the construct measured. Although these two unpublished studies represent important steps toward better understanding the effects of reading test accommodations on construct equivalence, clearly, much more research in this area is needed.

Although not a reading test, Rock et al. (1988) used confirmatory factor analysis to evaluate the comparability of scores from accommodated and non-accommodated administrations of the SAT and GRE. For the SAT, they found that the hypothesized two-factor (verbal and mathematical) structure fit the data "reasonably well for each of the nine handicapped (sic) groups as well as for the nonhandicapped group[2]" (p. 104). With respect to the hypothesized three-factor structure of the GRE, the only structural differences noted were for students with visual or physical impairments (data were not reported for students with learning disabilities). This study suggests that accommodations can be granted in a way that does not alter the construct, but it should be noted that several types of accommodations were involved in this study, and the effects of each type of accommodation were not isolated.

Before leaving our discussion of construct equivalence, it is interesting to note that the National Center on Educational Outcomes (NCEO) suggests use of the term "accommodation" to refer to changes in a test or test administration that do not change the construct measured. For example Thurlow and Weiner (2000) state. "The term accommodation when used for testing generally refers to a change in procedures or materials that does not change the construct being tested or the comparability of scores obtained from accommodated and non-accommodated testing (p. 1)." However, they go on to state "there are some changes in testing that may alter the construct being tested…A commonly cited example is reading aloud a reading test to a student when the purpose of the test is to measure decoding skills (p. 1)." They refer to such construct-altering accommodations as "modifications" or "non-standard admissions" (p. 2).

It is interesting to note that Thurlow and Weiner (2000) use the example of a read-aloud accommodation on a reading test as one of construct alteration. The "Standards" use the example of a written administration of an oral comprehension test as an example of an accommodation that changes the construct (p. 103). These examples suggest that in many cases it may be possible to base the conclusion that an accommodation alters the construct measured on professional judgment. However, Thurlow and Wiener echo the acknowledgement in the "Standards" that in many cases it is difficult to determine construct equivalence:

[2] It should be noted that these groups were defined by type of disability, rather than by type of accommodation. All groups, including those with learning disabilities, received extended time.

> Determining which constructs to allow (because they provide comparability) and which not to permit (because they change what is being tested) has been the subject of ongoing research and much debate. Not everyone agrees on what constitutes a change that either alters what is measured or the comparability of the scores (p. 2).

Although it is difficult, testing agencies must distinguish between accommodations that change the construct measured and those that do not, before interpreting scores from these different administrations. In the case of a national exam, such as the National Assessment of Educational Progress (NAEP) in the United States, accommodations should be provided to include as many students as possible in the assessment, but scores from test administrations that are deemed to change the construct measured should not be combined with scores from standard administrations as if they are on the same scale. For example, if reading experts agree that an oral administration of a NAEP reading test changes the construct measured from reading comprehension to listening comprehension, and if reading and listening comprehension are not perfectly correlated in the general population, scores from the standard and read-aloud accommodation administrations should not be considered comparable.

DO ACCOMMODATIONS PROMOTE FAIRNESS OR PROVIDE AN UNFAIR ADVANTAGE?

The construct equivalence of accommodated and standard test administrations is obviously related to the issue of how fair it is to grant accommodations to some, but not all students. However, it is possible that an accommodation does not change the construct measured, or actually improves measurement of the construct, but still provides an advantage to the students who receive the accommodation. This could occur, for example, when extra time is granted as an accommodation on a test that is unintentionally speeded (Sireci, Li & Scarpati, 2003). In such a situation, speed of response is not part of the construct measured, but the overly strict time limit affects scores for many students.

To defend the use of accommodations for only the SWD who need them, an "interaction hypothesis" has been proposed, which states that SWD need the accommodations and will benefit from them while students without disabilities will not benefit from them. This hypothesis (also referred to as the "maximum potential thesis" by Zuriff, 2000) has been posited by many researchers (e.g., Malouf, 2001, cited in Koenig, 2002; Shepard, Taylor & Betebenner, 1998; Weston, 2002) as one means for defending the validity of accommodations. The interaction hypothesis states that when test

accommodations are given to the SWD who need them, their test scores will improve, relative to the scores they would attain from taking the test under standard conditions, but students without disabilities will not exhibit higher scores when taking the test with an accommodation. Thus, the interaction specified in the hypothesis is between student group (SWD or non-SWD) and test administration condition (accommodated versus standard).

An illustration of the interaction hypothesis is presented in Figure 1, which depicts hypothetical mean test scores for SWD and non-SWD groups of students who take a test under both standard and accommodated conditions. The mean scores for the non-SWD group are equal under both test administration conditions, but the mean for SWD is higher under the accommodation condition. Advocates of test accommodations for SWD postulate this hypothesis as one means of arguing that test accommodations are needed for SWD so that they can demonstrate their true knowledge, skills and abilities.

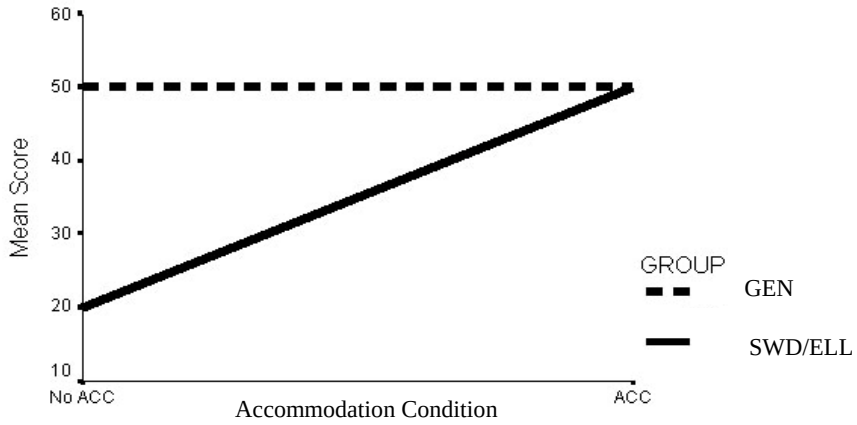

Figure 1. Illustration of interaction hypothesis.

Based on a review of the literature on the effects of test accommodation on test performance, Sireci, Li and Scarpati (2003) concluded a modification of the interaction hypothesis was needed to better reflect findings in the literature. They found that the most common test accommodation, which was extended time, led to the improvement of test scores for both SWD and students without disabilities. However, they found that generally, the score gains between standard and accommodated test administrations were greater for SWD than for other students. They hypothesized that this finding could be due in part to test speededness; that is, many of the tests studied had time limits that were too restrictive for many students, irrespective of disability category. Given these findings, they suggested that test accommodations for

SWD may be warranted, even in those situations where students without disabilities achieve gains under an accommodation condition, if the gains for SWD were greater. This finding is consistent with the concept of "differential boost" (Fuchs, Fuchs, Eaton, Hamlett & Karns, 2000; Phillips, 1994; Thompson Blount & Thurlow 2002), which states accommodations will lead to greater score improvements for students with disabilities than for students without disabilities. The differential boost hypothesis is presented in Figure 2.

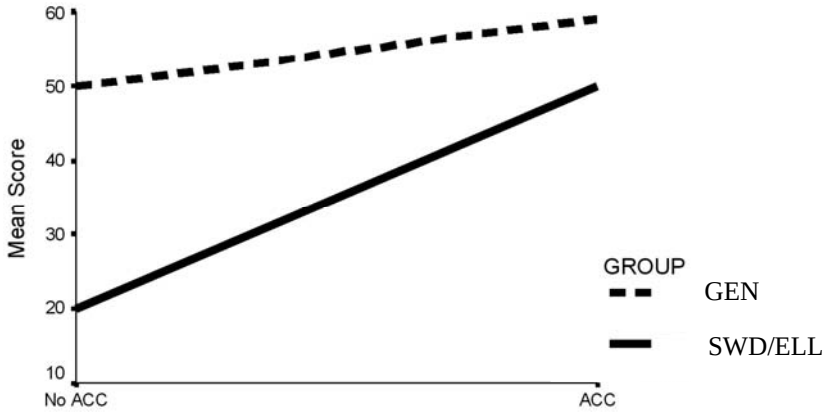

Figure 2. Illustration of differential boost hypothesis.

If test accommodations result in the type of interaction depicted in Figure 1, then they do not advantage students who are accommodated over students who are not accommodated. If the accommodation is beneficial to all students (Figure 2), then it may not be fair to limit the accommodation to SWD. As the "Standards" state "While test takers should not be disadvantaged due to a disability not relevant to the construct the test is intended to assess, the resulting accommodation should not put those taking a modified test at an undue advantage over those tested under regular conditions" (p. 105).

So, what does it mean when an accommodation, such as extended time increases the scores for all students? To answer this question, we must consider the construct measured and the accommodation. If the accommodation is extended time, and the construct measured does not involve the ability to answer test items quickly, it could mean that the standardized test conditions were unduly contaminated by overly strict time limits. In such a case, all students should be given extra time. However, if answering items quickly is part of the construct purportedly measured by the

test, then the accommodation dilutes measurement of the construct and the scores from accommodated tests are probably inflated.

Returning to the example of accommodations on NAEP reading tests, if speed of responding to reading material is not included in NAEP's definitions of reading proficiency, the accommodation of extra time probably does not result in a construct change. The degree to which SWD and students without disabilities do better on NAEP tests with extended time will help determine the fairness of the accommodation.

Accommodations for Reading Tests

As mentioned earlier, my colleagues and I reviewed the literature on test accommodations in search of empirical studies that evaluated the interaction hypothesis (Sireci Li & Scarpati., 2003). A summary of the types of accommodations used in these studies is presented in Table 1. The most common accommodations studied by researchers 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.

Table 1. General description of studies reviewed by Sireci, Li and Scarpati (2003).

Type(s) of Accommodation	# of Studies
Presentation:	
Oral*	22
Paraphrase	2
Technological	2
Braille/Large Print	1
Sign Language	1
Encouragement	1
Cueing	1
Spelling assistances	1
Manipulatives	1
Timing:	
Extended time	12
Multi day/sessions	1
Separate sessions	1
Response:	
Scribes	2
In booklet vs. answer sheet	1
Mark task book to maintain place	1
Transcription	1
Setting:	
Separate room	1
Total	52

*Includes read-aloud, audiotape, or videotape, and screen-reading software.

Table 2. Grade by subject cross-tabulation of studies reviewed by Sireci, Li and Scarpati (2003).

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	3.6
4	10	4	5	1	--	--	2	--	--	1	1	24	26.8
5	4	2	1	--	--	--	1	--	--	1	1	10	35.7
6	2	2	2	--	--	--	--	1	--	--	--	7	42.0
7	4	2	1	1	--	--	--	1	--	--	--	9	50.0
8	1	4	3	--	--	1	1	1	--	--	--	11	59.8
9	1	--	--	--	--	--	--	--	--	--	--	1	60.7
10	3	--	1	1	--	1	--	--	--	--	--	6	66.1
11	2	1	1	--	--	--	1	--	--	--	--	5	70.5
12	2	1	1	--	--	1	1	--	--	--	--	6	75.9
HS	--	1	1	--	--	--	1	--	--	--	--	3	78.6
C/U	--	1	--	--	--	--	--	--	--	--	--	1	79.5
PAT	10	3	--	--	--	--	--	--	10	--	--	23	100.0
Total	40	22	17	3	0	3	8	3	10	3	3	112	

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.

The studies we reviewed were also categorized by grade and subject area. A cross-tabulation of these variables is presented in Table 2. It should be noted that some studies investigated more than one subject area. most of the studies focused on elementary school grades and 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 evaluated the effect of accommodations on test performance for students in grades 9 to 12.

Table 3 presents a summary of the 10 studies that focused on reading tests. Some type of oral accommodation was used in three of the ten studies, two studies used extended time, and one study used both (along with large-print as a third accommodation for some students). The accommodation conditions for the other four studies were provision of a simplified English dictionary (for limited English proficiency (LEP) students, translating test material other than the reading passages – also for LEP students), breaking the test session into multiple days or sessions, and changing the means with which students recorded their answers. Since these ten studies represent the only empirical analysis of reading test accommodations found in the literature, they will be briefly reviewed.

Oral Administration Accommodations

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, computer, video, or audiotape 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. Although 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, there have been a few studies that investigated this accommodation for use on reading tests.

McKevitt and Elliott (2003) 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

Table 3. List of recent studies on accommodations for reading tests.

Study	Accommodation(s)	Design	Findings
Kosciolek and Ysseldyke (2000)	Read-aloud	Repeated measures w/ SWD and non-SWD	No gains for either group.
Meloy, Deville and Frisbie (2000)	Read-aloud	Repeated measures w/ SWD and non-SWD	Similar gains for SWD and non-SWD
McKevitt and Elliott (2003)	Audiotape presentation	Repeated measures w/ SWD and non-SWD	No effects for either student group.
Fuchs, Fuchs, Eaton, et al., (2000)	Extended time, large print, read-aloud	Repeated measures w/ LD and non-LD	Extended time & large print benefited both groups, read-aloud benefited LD only.
Runyan (1991)	Extended time	Repeated measures w/ SWD and non-SWD	SWD exhibited larger gains.
Huesman and Frisbie (2000)	Extended time	Quasi-experimental	Score gains for LD but not for NLD groups.
Anderson et al. (2000)	Bilingual test booklets and audiotape translation of non-passage material	Between-group	No gains for LEP students.
Albus et al. (2001)	Simplified English Dictionary	Between-group	No gains for LEP or non-LEP students in general, some gains for lower-LEP students.
Walz et al. (2000)	Multiple days, sessions	Repeated measures w/ SWD and non-SWD	No gains for either student group.
Tindal et al. (1998)	Response format	Repeated measures w/ SWD and non-SWD	No score differences when using answer sheet or writing in booklet.

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). There was no interaction or differential boost between student group and accommodation condition.

McKevitt and Elliott 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 that it was easier to show what they knew when given 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 Iowa Tests of Basic Skills (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 seventh 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 0.75 standard deviations) was larger than the gain for students without a disability (about 0.50 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 led Meloy, Deville and Frisbie to conclude that general use of the read-aloud accommodation for LD students taking standardized achievement tests is not recommended.

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 to 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. Each student took two equivalent forms of the California Achievement Tests (CAT/5),

Comprehension Survey. One form was administered with a read-aloud accommodation, the other was administered without an accommodation, and the order of the accommodation condition was counterbalanced. To maintain consistency between testing sessions, the read-aloud accommodation was provided using a standard audiocassette player. Two open-ended questions were asked of the students at the end of the testing session to get an idea of student perception of and comfort level with the read-aloud test 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. 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 ($p = 0.06$) it represented a large effect size (0.56). 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 (0.10). Kosciolek and Ysseldyke also noted that SWD embraced the accommodation, while the students without disabilities preferred the standard administration. Of the three studies that looked at only at oral accommodations for reading tests, this was the only one that provided slight evidence in support of the interaction hypothesis. However, given the small sample sizes, and the results of the other two studies, there is little data to support oral accommodations on reading tests.

Extended Time Accommodations

Runyan (1991) examined reading test 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 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.

Earlier, I discussed the factor analytic results of Huesman and Frisbie (2000). In that same study Huesman and Frisbie also conduct a quasi-experimental analysis of the effects of extended time on test scores for both students with learning disabilities and students without disabilities. The test studied was the 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 0.61). 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 0.38); 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. This finding appears to be consistent with the other studies that empirically evaluated extended time accommodations for reading tests.

Oral and Extended Time Accommodations

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. 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.

Dual-language Booklets

Anderson et al. (2000) evaluated the accommodation of providing dual-language test booklets on a reading test to limited English proficient students. 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 English language learner (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.

Response format

Tindal et al. (1998) used an experimental design to investigate the effects of oral accommodation on a math test and response format on a reading test. I only comment on the reading test results here. The specific response format investigated was allowing students to write their answers into the test booklet rather than on an answer sheet.

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.

Multiple-day Accommodation

Walz et al. (2000) looked at a "multiple-day" accommodations for SWD on reading tests. 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 Empirical Analysis of Accommodations for Reading Tests

As the summaries provided in Table 3 imply, extended time is a potentially reasonable accommodation for SWD when they take reading tests. However, read-aloud accommodations do not produce results consistent with the interaction or differential boost hypotheses, and the unpublished factor analytic studies done in this area (i.e., Tippets & Michaels, 1997, cited in Bielinski et al., 2001; Huesman & Frisbie, 2000) do not provide enough evidence to suggest the accommodation does not alter the construct. Thus, there is little evidence in support of oral accommodations for reading tests. The other accommodations studied, bilingual portions of test booklets, multiple testing sessions, provision of simplified dictionaries, and easier response formats also did not lead to increased scores for SWD. However, very few studies have been conducted on these accommodations and so more research is warranted.

Are scores from accommodated and standard test administrations comparable? That is, can they be interpreted as if they are on the same scale? Up to this point I reviewed validity issues in test accommodations and reported on the results of some empirical studies that looked at the validity of specific accommodations for reading comprehension tests. There is one more issue to be addressed, namely, if an accommodation does alter test scores, is there a way to adjust these scores so that they can be made comparable to scores from a standard administration? This question puts us in the realm of scaling and equating.

Powers and Willingham (1988) addressed the issue of whether test scores taken under accommodated conditions could be "rescaled" (equated) to make them comparable to those taken under standard conditions. They considered two equating strategies and rejected them both. The first strategy involved equating test scores obtained from individuals with disabilities who took the test under non-standard conditions with those who took the test under standard conditions. This approach is not feasible due to simultaneous differences in examinees and test difficulty. The second proposal involved equating the scores through an external criterion such as college grades. This proposal was also rejected, primarily due to the insufficiency of college grades as a valid equating criterion.

However, a more recent idea is a third equating strategy: equating test scores administered under the condition of extended time to those administered under standard time conditions using representative samples of non-disabled students (Sireci, 2001). The logic underlying this idea is that equating can be used to adjust for differences in overall difficulty between two parallel tests.

A recent study by Bridgeman, Trapani and Curley (2004) suggests giving tests with extended time is analogous to taking an easier test form. It may be possible to adjust for this difference in difficulty through statistical equating.

One possibility for accomplishing such equating is to use a randomly equivalent groups equating design. For example, a representative group of students registered to take a test would get a note describing the special study and informing them that they could have a specific accommodation (e.g., extra time), if they like. They would also be told that this accommodation would probably not result in a score increase (since the equating would ultimately adjust for such an increase). This group would take a specific form of the test that others were also taking on the same day under standard conditions. Thus, there would be two randomly equivalent groups of examinees taking the same form on the same day, but one group would have an accommodation. The scores on the extended time version could be equated onto the scale of the standard time group using equipercentile equating.

The issue of how to use the equating adjustment on all subsequent extended time administrations would also need to be addressed. One way this could be accomplished is to repeat this study several times to get an average increase due to extended time that could be used to adjust the scores on these tests. Another idea is to repeat this study for each administration, with people with disabilities who apply for extended time taking a predetermined test form.

More practical approaches may also be possible, such as allowing for extended time on separate sections of the test for some representative groups and then adjusting each section. Or perhaps one section could be given with extended time to a representative group and then used as an anchor in an anchor-item equating design. The key to these propositions is to have a representative group of examinees take the test with extended time, rather than a group of examinees with disabilities, or any other potentially non-representative sample. If equating of scores from standard and extended-time administrations of tests were accomplished, then SWD who desire extended time could be given the accommodation, and there would be no reason to flag their scores, since they would be on the same scale as scores from the standard administration.

CONCLUSIONS

In an earlier section of this paper I raised the question "Do test accommodations change the construct measured?" I also raised the question "Do accommodations promote fairness or provide an unfair advantage?" Clearly, the appropriate questions are not "Do" questions, but "which" questions. That is, research and standards in educational testing require us to determine which accommodations change the construct measured and which accommodations promote, rather than hinder, fairness. Therefore, testing agencies must examine several factors before making decisions about whether to grant an accommodation and how to report scores from accommodated test administrations.

Our review of the issues and research in this area suggests several sensible directions regarding accommodations on reading tests.

1. Read-aloud and other oral accommodations to reading tests are likely to change the construct measured. Although it may be appropriate to provide this accommodation to some students with reading disabilities, scores from orally accommodated reading comprehension tests should not be combined with scores from standard administrations of the test.

2. More flexible time limits are likely to reduce unintended speededness effects on educational tests. Extended time accommodations may be appropriate on reading tests, assuming reading speed is not part of the construct purportedly measured. However, if the tests are unintentionally speeded, accommodating only some students is unfair to other students.

3. The principles of universal test design, which suggest building tests with greater content validity and more flexible administration conditions should be considered for future development of reading tests. As Thompson, Blount and Thurlow (2002) describe:

> 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, Blount & Thurlow, p. 17).

4. Both qualitative and quantitative approaches should be used to determine whether a particular test accommodation changes the construct measured. Qualitative approaches include convening groups of subject matter experts to determine the effects of the accommodation on the construct. Quantitative methods include dimensionality analyses, differential predictive validity studies, and studies of differential item functioning. Experimental designs to compare the gains for SWD and other students under accommodation and non-accommodation conditions should also prove helpful for evaluating the equivalence of accommodated and standard test administrations.

5. Finally, testing agencies must develop clear definitions of the constructs measured on a test, as well as potential sources of construct-irrelevant variance. These definitions will help test users better evaluate the utility of the test and will help facilitate understandings of how accommodations may alter the construct.

In closing, it is clear that in some cases the provision of a test accommodation to a particular student with a particular disability will increase test validity and not provide an unfair advantage to that student; but in other cases, a particular accommodation may not promote validity and may be unfair to students who do not receive the accommodation. Thus, accommodation decisions must take into account the construct measured by a test, the degree to which the accommodation is likely to alter the construct, and the specific needs of a particular student. Research to date has provided some information on what types of accommodations are likely to maintain fidelity to the construct and remove construct-irrelevant variance. However, ultimately, accommodation and score-reporting decisions must be made on a case-by-case basis.

REFERENCES

Albus, D., Bielinksi, J., Thurlow, M. and Liu, K. (2001). *The effect of a simplified English languange dictionary on a reading test. LEP Projects report 1.* Minneapolis, MN: National Center on Educational Outcomes, University of Minnesota.

Allalouf, A., Hambleton, R. K. and Sireci, S. G. (1999). Identifying the sources of differential item functioning in translated verbal items. *Journal of Educational Measurement, 36,* 185–198.

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME). (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Psychological Association (APA), American Educational Research Association (AERA) and National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC : American Educational Research Association.

Anderson, M., Liu, K., Swierzbin, B., Thurlow, M. and 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 24 January 2003 from http://education.umn.edu/NCEO/OnlinePubs/MnReport31.html.

Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J. and Freidebach, M. (2001). *Read-aloud accommodations: Effects on multiple-choice reading and math items (Technical Report 31)*. Minneapolis, MN: University of Minnesota, National Center on Educational Coutcomes. Rerieved 24 January 2004 from http://education.umn.edu/NCEO/OnlinePubs/Technical31. htm

Bridgeman, B., Trapani, C. and Curley, E. (2004). Impact of fewer questions per section on SAT I scores. *Journal of Educational Measurement, 41,* 291–310.

Chiu, C. W. T and Pearson, P. D. (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, Utah, 13–15 June.

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

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

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

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

Huesman, R. L. and Frisbie, D. (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, Los Angeles, 24–28 April.

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

Kosciolek, S. and 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 6 January 2003, from http://education.umn.edu/NCEO/OnlinePubs/Technical28.htm

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635–694 (Monograph Supplement 9).

McKevitt, B. C., Elliott, S. N. (2003). Effects and perceived consequences of using read-aloud and teacher-recommended testing accommodations on a reading achievement test. *School Psychology Review, 32(4),* 13–26.

Meloy, L., Deville, C. and Frisbie, D. (2000). *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, Los Angeles, 24–28 April.

Messick, S. (1989). Validity. In R. Linn (ed.), *Educational measurement* (3rd ed.). Washington, DC: American Council on Education, 13–100.

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

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

Powers, D. E. and Willingham, W. W. (1988) The feasibility of rescaling. In W. W. Willingham, M. Ragosta, R. E. Bennett, H. Braun, D. A. Rock, and D. E. Powers (eds.). *Testing handicapped people*. Needham Heights, MA: Allyn and Bacon, 133–142.

Rock, D. A., Bennett, R. E., Kaplan, B. A. and Jirele, T. (1988). Construct validity. In W. W. Willingham, M. Ragosta, R. E. Bennett, H. Braun, D. A. Rock and D. E. Powers (eds.). *Testing handicapped people*. Needham Heights, MA: Allyn and Bacon, 99–107.

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 and J. N. Zall (eds.). *Advances in educational and psychological testing*. Norwell, MA: Kluwer, 251–276.

Shepard, L., Taylor, G. and 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. (2001, December). *Equating non-standard and standard administrations of the SAT*. Unpublished opinion paper submitted to the Blue Ribbon Panel on Flagging.

Sireci, S. G. (2003). *Unlabeling the disabled: A psychometric perspective on flagging scores from accommodated test administrations. Center for Educational Assessment Research Report No. 502.* Amherst, MA: School of Education, University of Massachusetts.

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

Sireci, S. G. and Gonzalez, E. J. (2003). *Evaluating the structural equivalence of tests used in international comparisons of educational achievement.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, Illinois, 12–14 April.

Sireci, S. G. and Khaliq, S. N. (2002). *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, Los Angeles, 2–4 April.

Sireci, S. G., Li, S. and Scarpati, S. (2003). *The effects of test accommodations on test performance: A review of the literature. Center for Educational Assessment Research Report No. 485*. Amherst, MA: School of Education, University of Massachusetts, Amherst.

Thompson, S., Blount, A. and 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 6 January 2003 from http://education.umn.edu/NCEO/OnlinePubs/Technical34.htm

Thurlow, M. and Weiner, D. (2000). *Non-approved accommodations: Recommendations for use and reporting (Policy Directions 11).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

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

Walz, L., Albus, D., Thompson, S. and Thurlow, M. (2000 December). *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 3 January 2003 from 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 (NCES 200306).* Washington, DC: National Center of Education Statistics.

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

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.