WCER Working Paper No. 2002-1
April 2002

Testing Accommodations Research and Decision Making: The Case of "Good" Scores Being Highly Valued But Difficult to Achieve for All Students

Stephen N. Elliott
Department of Educational Psychology/
Wisconsin Center for Education Research
University of Wisconsin–Madison
snelliot@facstaff.wisc.edu

Ryan J. Kettler
Department of Educational Psychology/
Wisconsin Center for Education Research
University of Wisconsin–Madison
rjkettler@students.wisc.edu

Brian C. McKevitt
Heartland Area Education Agency 11
Johnston, IA
bmckevitt@aea11.k12.ia.us

Wisconsin Center for Education Research
School of Education • University of Wisconsin–Madison
http://www.wcer.wisc.edu/

The research reported in this paper was supported by the U.S. Department of Education, Office of Special Education and Rehabilitative Services under Grant No. H324D990027 and by the Wisconsin Center for Education Research, School of Education, University of Wisconsin–Madison. Any opinions, findings, or conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies, WCER, or cooperating institutions.

Testing Accommodations Research and Decision Making: The Case of "Good" Scores Being Highly Valued But Difficult to Achieve for All Students[1]

Stephen N. Elliott, Ryan J. Kettler, and Brian C. McKevitt

Achievement testing is often considered part of an accountability system that states and/or districts use to measure and report student achievement. In the past, many students with disabilities have been excluded from large-scale achievement tests. Reasons for the exclusion of students with disabilities are varied, but the most common are confusion about the use of testing accommodations, concern over causing undue stress from testing, or fear that district test scores will go down (Elliott & Braden, 2000; Elliott, Braden, & White, 2001; Heubert & Hauser, 1999). However, as recent education reform efforts concerning high standards for all students have developed, the inclusion of students with disabilities in accountability efforts has become a requirement of law and a key aspect of good testing practices. Thus, from social and accountability perspectives, the inclusion of students with disabilities in statewide accountability systems is highly valued, especially when their scores are known to be valid. Unfortunately, at this time many students with disabilities are receiving testing accommodations of unknown validity.

The inclusion of students with disabilities in assessment is deemed critical to improve the quality of educational opportunities for these students and to provide meaningful and useful information about students' performance to the schools and communities that educate them (McDonnell, McLaughlin, & Morrison, 1997). This inclusion raises important questions, however, concerning the appropriateness of common performance standards for students with disabilities, the appropriate accommodations to use, the effects of testing accommodations on the validity of assessment, and the reporting of scores when accommodations have been used (Pitoniak & Royer, 2001). Education leaders at state and district levels struggle with these issues as they work to create policies regarding testing students with disabilities.

The purposes of this article are to (a) review definitional and legal issues associated with testing accommodations, (b) focus on validity issues that confront educators who must make decisions about the use and likely effects of testing accommodations, and (c) summarize findings about the statistical effects of accommodations on test scores. A basic premise of this article is that the inclusion of all students in educational accountability systems is an important and attainable goal. The wise use of testing accommodations can be an effective tactic for increasing the meaningful participation of students who have frequently not been included in large-scale accountability assessments and to increase the likelihood of "good" (i.e., valid) test scores.

---

[1] A portion of this paper was presented as part of an invited presentation on testing accommodations research at a workshop in Washington, D.C., on November 28, 2001, entitled "Reporting Test Results for Accommodated Examinees: Policy, Measurement, and Score Use Considerations." This workshop was sponsored by the National Academies/National Research Council's Board on Testing and Assessment.

Definitional Issues and Controversies

Testing accommodations are changes in the way a test is administered to or responded to by the person tested. These changes are intended to offset or correct for distortions in scores that may be caused by a student's disability (Elliott, Kratochwill, & Schulte, 1998; McDonnell et al., 1997; Pitoniak & Royer, 2001). Thus, accommodations are used to help students show what they know on assessments without being impeded by their disability. Accommodations are not intended to change the nature of the construct being measured for the examinee; rather, they are meant to make the measurement of a particular construct comparable across examinees.

The term *modification* is occasionally used interchangeably with the term *accommodation* (Hollenbeck, Tindal, & Almond, 1998). Generally, a modification is considered a change in the content of the test, whereas an accommodation is considered a change in the way a test is administered. Content modifications likely change what the test measures (McDonnell et al., 1997). Modifications, for example, may include deleting certain items that are inappropriate for an examinee or making constructed-response questions into multiple-choice questions. These types of modifications are presumed to change the nature of what is being tested.

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999) defines an accommodation as "any action taken in response to a determination that an individual's disability requires a departure from established testing protocol" (p. 101). The *Standards* go on to say, "Depending on circumstances, such accommodation may include modification of test administration processes or modification of test content. No connotation that modification implies a change in the construct(s) being measured is intended" (p. 101). This definition of accommodations differs from that found in other accommodations literature. Namely, the *Standards* definition accepts changes in test content as accommodations. For the purposes of this article, however, the terms *accommodation* and *modification* will *not* be used interchangeably.

In many discussions with educators, we have found the use of two metaphors for testing accommodations useful in communicating their essence. The first metaphor for testing accommodations concerns *eyeglasses*. Eyeglasses are an accommodation for imperfect or poor vision. If you wanted to test the natural vision ability of a person who wears glasses for driving and outdoor activities, then having the person wear glasses during a test of distant vision would invalidate the test score assuming your purpose is to make an inference about the person's natural or uncorrected vision. On the other hand, if your purpose was to determine the same person's driving ability, then permitting the person to wear the glasses during the driving test that he or she wears daily would be a valid accommodation because it would facilitate a more accurate assessment of the person's driving skills by minimizing or eliminating problems due to vision impairments. Remember, even in the absence of disabilities or other complicating factors, tests are imperfect measures of the constructs they are intended to assess.

The second metaphor for testing accommodations is an *access ramp*. As a point of fact, an access ramp can be conceptualized as part of a package of testing accommodations for individuals with significant physical impairments. If individuals can't get to the testing room, then they certainly can't demonstrate what they know or can do! The conceptual value of an

access ramp has additional meaning, however, when addressing issues of construct validity. Testing accommodations facilitate access to a test for students with a wide range of disabilities just like a ramp facilitates access to a building for individuals with physical disabilities. The tests that students are required to take are designed to measure some specific *target cognitive skills or abilities*, such as mathematical reasoning and computations, but almost always assume that students have the skills to access the test, such as attending to instructions, reading story problems, and writing responses. Thus, these tests are designed to measure knowledge and concepts and target broad constructs like mathematics and language arts. Some students, in particular many students with disabilities, have difficulty with the *access skills* needed to get "into" the test. Thus, valid testing accommodations, just like an access ramp, should be designed to reduce problems of access to a test and enable students to demonstrate what they know and can do with regard to the skills or abilities the test is targeting.

By now, you should have a good understanding of what testing accommodations are, how they should function to improve the validity of a student's test score, and the potential value they have for meaningfully including more students in large-scale accountability systems. Assessment accommodations can include alterations to the presentation format, response format, timing/scheduling of the test, and setting and environment. One or more accommodations to an assessment can be made involving scheduling, setting, test directions, assistance during the assessment, and the use of aids (e.g., calculator, tape recorder). Most states, however, do not allow for any of these accommodations unless they are listed in a student's individual education plan (IEP). Therefore, IEP teams in many schools need support and guidance to meaningfully implement the assessment requirements of the Individuals with Disabilities Education Act (IDEA) and its related state legislation. In other words, IEP teams are entrusted with selecting valid accommodations for individual students *prior to testing*. This can be a difficult assignment!

Common types of accommodations that may be used with students include changes in the setting in which a test is taken (e.g., special education classroom, study carrel); the timing of a test (e.g., extended time to complete the test); the scheduling of the test (e.g., administration at a time most beneficial for the student); the presentation of the test (e.g., provision of an audiotape or a reader, paraphrasing of directions); or the way the student responds to the test (e.g., permission to record answers in the test book rather than the answer sheet). Elliott et al. (1998) identified eight domains of accommodations that can be used before and during assessment. These domains are (a) motivation; (b) assistance prior to the administration of the test; (c) scheduling; (d) setting; (e) assessment directions; (f) assistance during assessment; (g) use of equipment or adaptive technology; and (h) changes in format. Regardless of the way in which accommodations are categorized, all are intended to assist students in removing barriers created by disabilities when taking a test. For example, Braille text on a reading test could be considered an accommodation for a person with visual impairment. The Braille is intended to remove the barrier of the person's sight limitation during testing.

For some students (e.g., those with visual impairments), the use of accommodations (e.g., Braille) clearly offsets difficulties in measuring the construct of a test. For other students, however, the utility and effect of accommodations are less clear. A recently published set of accommodation guidelines written by CTB/McGraw-Hill (2000) featured a three part model developed by Thurlow and her associates (Thurlow, House, Boys, Scott, & Ysseldyke, 2000) for

categorizing types of accommodations by the way they are believed to affect student performance. *Category 1 accommodations* are not expected to influence how test scores can be interpreted because they are clearly not related to the construct being measured. These accommodations generally refer to changes in the location of testing (e.g., taking the test in a study carrel). *Category 2 accommodations,* however, may interfere with adequate measurement of the construct, depending on what is being measured. Extra time is an example of a Category 2 accommodation and should be noted when results are being interpreted. Finally, when a *Category 3 accommodation* is used, scores must be interpreted cautiously because the accommodation may have changed what the test is purported to measure. Accommodations in this category are specific to test content. For example, using a calculator as an accommodation on a math computation test may change the nature of what the math test is measuring. The following statement by Phillips (1994) demonstrates this problem with read-aloud accommodations:

> [A]dministration of the reading test in oral rather than written form substitutes measurement of the skill of listening comprehension for the intended skill of reading comprehension. Thus, a blind student who passes the test using a Braille edition has demonstrated competence in the intended skill of reading comprehension, while not being penalized for the unrelated physical impairment of lack of sight. However, the dyslexic applicant for whom the test was read aloud has not demonstrated competence in reading comprehension because the accommodation in this case is related to the mental skill intended to be measured. (p. 102).

Phillips views the disability and the reading construct as intertwined, leading to ineffective measurement of the construct. In contrast and by way of illustrating the challenging conflicts that confront educators, J. Elliott, Ysseldyke, Thurlow, and Erickson (1998) stated that read-aloud accommodations may be appropriate on a reading comprehension test. According to them, read-aloud would be appropriate if the test is measuring the student's ability to understand written language and interpret meaning, and not his or her ability to decode words. (It should be noted that test publishers could make all our lives easier if they clearly stated at the beginning of each subtest what construct was being measured; they would be even more helpful if they indicated the intended target skills and likely access skills needed to complete the items.) Furthermore, Harker and Feldt (1993) stated that by middle-school age, children's listening and reading comprehension skills become comparable. This finding provides some further support for the use of read-aloud accommodations on tests of comprehension for older students. Yet in another study by Meloy, Frisbie, and Deville (2000) of a read-aloud accommodation, the investigators found the accommodation to have greater positive effects for groups of students with disabilities than for those without disabilities, but they believed it changed the construct being measured for most students and thus cautioned against its use. Given the numerous questions and controversies over this and other types of accommodations, law, policy, and measurement standards are needed to provide guidelines for selecting and using testing accommodations wisely. Practicing educators should not be expected to make consistent and appropriate decisions when knowledgeable researchers cannot even come to agreement on the appropriateness of an accommodation!

Legal and Policy Issues Associated with the Use of Testing Accommodations

Federal education and civil rights laws include language related to the use of testing accommodations on state and district-wide large-scale accountability assessments. In the 1997 reauthorization of IDEA, accommodations, if necessary, are mandated to facilitate the inclusion of students with disabilities in assessments. If participation in assessment is not appropriate, even with accommodations, an alternative assessment must be developed to measure a student's progress [Pub. L. No. 105-17. § 612(17)(A)(i)-(ii)]. Furthermore, the regulations governing the IDEA cover the documentation of participation and accommodation in students' IEPs. These regulations state the IEP must include "[a] statement of any individual modifications in the administration of State or district-wide assessments of student achievement that are needed in order for the child to participate in the assessment; and . . . [i]f the IEP team determines that the child will not participate in a particular . . . assessment of student achievement . . . , a statement of . . . [w]hy that assessment is not appropriate for the child; and . . . [h]ow the child will be assessed" [34 C.F.R. § 300.347(5)(i)-(ii)].

Section 504 of the Vocational Rehabilitation Act of 1973 (Pub. L. No. 93-112) is part of a civil rights law that also contains provisions related to testing accommodations. This law states that "No otherwise qualified individual with a disability in the United States . . . shall, solely by reason of his or her disability, be excluded from the participation in, be denied the benefits of, or be subjected to discrimination under any program receiving Federal financial assistance" (29 U.S.C. § 794(a)). The regulations governing section 504 further state that admissions tests must accurately reflect the test-taker's aptitude or achievement and not his or her lack of skill related to the disability except when that skill is the factor the test purports to measure [45 C.F.R. § 84.42(b)(3)].

Case law also has contributed to the growing body of legal guidelines associated with testing accommodations. Thurlow, Ysseldyke, and Silverstein (1995) identified five cases that are commonly associated with testing accommodations issues and the inclusion of students with disabilities in assessment. The courts involved in these cases found that (a) adequate advance notice of testing requirements (1½–3 years) must be provided to students and parents (*Debra P. v. Turlington*, 1984; *Board of Educ. v. Ambach*, 1983), (b) tests must measure what is taught in the curriculum (*Debra P. v. Turlington*, 1984), (c) there must be equal opportunity to participate in testing (*Anderson v. Banks*, 1981; *Brookhart v. Illinois State Bd. of Educ.*, 1983), and (d) educational institutions do not have to change or modify standards or programs to accommodate a person with a disability (*Southeastern Community College v. Davis*, 1979).

A more recent case related specifically to read-aloud accommodations involved the Hawaii State Department of Education (1990). In the case, the Office of Civil Rights (OCR) ruled that the state department of education could not deny a student with a learning disability read-aloud accommodations on the non-reading portions of the state-mandated graduation test (Phillips, 1993). It also determined that accommodations must be judged on a case-by-case basis (Phillips, 1994) and that read-aloud accommodations may not be appropriate for all students with learning disabilities. In its ruling, OCR conceded that allowing a reader for the reading portion of the test would "defeat the purpose of the test and that denying it would not be discriminatory" (Phillips, 1994, pp. 112-113).

In response to federal and case law, state and local education agencies face the challenge of creating policies to guide decision making about participation in assessments and testing accommodations for students with disabilities. As of 1997, 40 of 50 states had active policies on the participation of students with disabilities in large-scale statewide assessment (Thurlow, Seyfarth, Scott, & Ysseldyke, 1997). Variables determining participation included relevance of the curriculum to the test, IEP goals, parent/guardian desire, percentage of time receiving special education services, and the likelihood of obtaining a reliable and valid score.

Also as of 1997, 39 of 50 states had active accommodation policies (Thurlow et al., 1997). Of these 39 states, 31 offered multiple accommodations on statewide tests. Some did not list specific allowable accommodations but rather deferred the decision to the individual student's IEP team. Presentation accommodations (e.g., reading the test aloud to a student) varied widely by state. According to Thurlow et al. (1997), "reading the test aloud is one of the most controversial accommodations" (p. 21), yet 9 of the 39 states offered reading aloud as an acceptable accommodation with no restrictions. Acceptable response accommodations (e.g., using a scribe to record answers, using a calculator) also varied widely among states. For example, 10 states allowed calculator use with no restrictions, whereas 2 prohibited it completely. Finally, scheduling accommodations (e.g., extended time) were frequently included in state policy, with 13 of the 39 states allowing extended time and several offering frequent breaks or test administration at a time most beneficial to the student. Clearly the differences in policies among states, with some allowing accommodations that others prohibit, points to the need for more information about the use and effects of testing accommodations, and also to the difference in values about the role of testing and perhaps the importance of including students with disabilities in statewide testing programs.

Psychometric Issues Associated with the Use of Testing Accommodations

Along with legal and policy issues, numerous measurement issues arise when discussing standardized testing and testing accommodations. As stated previously, testing accommodations are changes in the administration of a test, or the way a student responds to a test, to compensate for or offset distortions in performance that may occur because of a disability. Given the heterogeneity of disabilities, accommodations often are suited to the needs of individual students, so the provision of accommodations may differ for different students taking the same test. The individualized nature of accommodations seems to contrast with the purpose of standardization that allows score comparability. The lack of standardized and uniform administration procedures endangers the reliability of a test and potentially invalidates the interpretation of resulting scores. Thus, reliability and validity are important psychometric considerations that need to be thoroughly addressed when discussing testing accommodations and understanding "good" scores.

Reliability is a property of assessment that refers to a tool or technique's ability to be consistent in its measurement. A test is considered unreliable if scores are subject to random variation. Such variation can affect the accurate measurement of a trait or a characteristic. Although error is assumed in all measurement, the chances for random error can be reduced in a tool that is consistent in the way it measures a behavior. This consistency can be demonstrated by agreement between raters or observers (i.e., interrater/interobserver reliability), across time periods (i.e., test-retest reliability), or between alternate forms of the same test. Reliability is

important because it tells us whether a test or measure itself can be trusted for the purpose intended.

Validity is a property of assessment that defines the extent to which evidence and theory support the meaningful interpretation of test scores, as defined by the purpose and proposed use of a test. According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), validation is the process of developing arguments to support appropriate score interpretation. A tool that yields scores with good validity is a tool that can provide meaningful information about that which is being measured. Given that assessment can have so many purposes, and consequently so many possible decisions or actions based on it, it is of great importance that one have confidence that the results are truly related to what is intended to be measured. Without evidence of validity, the results of a test are essentially meaningless, and the consequences based on those results could be devastating. When thinking about validity, it is useful to keep the following points in mind:

1. Validity is concerned with the general question, To what extent will this assessment information or test score help me make an appropriate decision?

2. Validity refers to the decisions that are made from assessment information, not the assessment approach or test itself. Keep in mind that assessment information that is valid for one decision or group of students is not necessary valid for other decisions or groups.

3. Validity is a matter of degree; it does not exist on an all-or-nothing basis. Think of assessment validity in terms of categories: highly valid, moderately valid, and invalid.

4. Validity involves an overall evaluative judgment. It requires an evaluation of the degree to which interpretations and uses of assessment results are justified by supporting evidence and in terms of the consequences of those interpretations and uses.

Numerous professionals have articulated concerns about the influence of testing accommodations on the validity of test scores for students with disabilities. As background for a detailed examination of validity issues, we document (chronologically) the essence of several of the concerns about using testing accommodations by quoting authoritative sources. Most of the quotations also suggest some of the methods and evidence needed to establish the validity of accommodated test scores.

> From a measurement point of view, the bottom line is whether the scores with and without accommodations are comparable. That is, do scores from nonstandard test administrations have the same meaning as scores from standard test administration? (Phillips, 1994, p. 96)
>
> To design an accommodation that will increase the validity (meaningfulness) of scores for students with disabilities, one must first identify the nature and severity of the distortions the accommodations will offset. These distortions depend on the disability, the characteristics of the assessment, the conditions under which the assessment is administered, and the inferences that scores are used to support. (McDonnell et al., 1997, p. 177)
>
> Professional judgment necessarily plays a substantial role in decisions about test accommodations. Judgment comes into play in determining whether a particular individual needs accommodation and the nature of such accommodation. . . . The overarching concern is the

> validity of the inference made from the score on the modified test: fairness to all parties is best served by a decision about test modification that results in the most accurate measure possible of the construct of interest. (AERA, APA, & NCME, 1999, p. 102)

The *Standards for Educational and Psychological Testing* (AERA, APA,& NCME, 1999) emphasize five sources of validity evidence: test content, response processes, internal structure, relations to other variables, and consequences of testing. These sources do not represent distinct types of validity. Consistent with Messick's (1995) proposition, validity is a unitary concept for which supportive evidence is gathered and evaluated.

Evidence based on test content stems from logical, empirical, or expert analyses of the "adequacy with which the test content represents the content domain and the relevance of the content domain to the proposed interpretation of test scores" (AERA, APA, & NCME, 1999, p. 11). Differences in the meaning or interpretation of scores can be addressed by evidence from test content. In particular, a review of test content may assist in identifying potential sources of construct underrepresentation or construct-irrelevant variance that may provide unfair advantages or disadvantages to one group over another.

Evidence based on response processes looks at the ways in which examinees respond to test questions (AERA, APA, & NCME, 1999). By analyzing examinees' response processes, one can observe the match between the test's construct (i.e., what the test is intended to measure) and performance (i.e., what the test-taker produces). For example, students with and without disabilities may perform differently on tests when accommodations are provided (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000; McKevitt & Elliott, 2001; Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998). A comparison of the performance of the two groups with and without accommodations may provide response process evidence for the validity or invalidity of score interpretation. The questions addressed in the studies cited above were intended to provide validity evidence related to response processes, as the scores of examinees with and without disabilities were compared when accommodations were and were not provided for each group.

Evidence based on internal structure provides information about the relationships among test items. The way in which those items function together or differentially provides evidence to support or refute score interpretations. In accommodations research, an analysis of differential item functioning (DIF) can be used as a mechanism to test score validity (Lewis, Green, & Miller, 1999). Items that function differently for different groups (e.g., students with and without disabilities) may indicate invalidity resulting from the use of accommodations. In addition, factor analysis evidence is also of value (see Huesman, Jr. & Frisbie, 2000).

Evidence based on relations to other variables is important to establishing validity of score interpretation (AERA, APA, & NCME, 1999). Experimental or correlational evidence of convergent and discriminant relationships between test scores and other measures provides useful information about how a test's construct matches other tests with similar or dissimilar constructs. Evidence also may be gathered by analyzing a test's ability to predict some criterion or outcome of interest, or by predicting differential outcomes for different groups. In addition, validity generalization—i.e., how well evidence of validity based on test-criterion relations can be used in a situation in which validity is not studied—is also an important source of validity evidence.

The final source of validity evidence is *evidence based on the intended and unintended consequences of testing*. According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999),

> evidence about consequences may be directly relevant to validity when it can be traced to a source of invalidity such as construct under-representation or construct-irrelevant components. Evidence about consequences that cannot be so traced—that in fact reflects valid differences in performance—is crucial in informing policy decisions but falls outside the technical purview of validity. (p. 16)

Thus, in cases in which testing accommodations are not used with students with disabilities on tests with high consequences (e.g., a graduation exam), but the disability is considered a source of construct-irrelevant variance, the consequence of omitting the accommodations may be retention. This consequence would be considered a source of evidence for score invalidation, because the accommodations should have been used to correct for the disability. Thus, there can be problems if one uses accommodations and another set of problems if one does not.

The bottom line, according to Phillips (1994), is "whether the scores with and without accommodations are comparable. That is, do scores from nonstandard test administrations have the same meaning as scores from standard test administrations?" (p. 96). Test standardization is the traditional method for making test results among examinees comparable (McDonnell et al., 1997). Standardization, however, may reduce the comparability of scores for students with disabilities because the disability itself "biases the score" by creating construct-irrelevant variance in the score (p. 173). The use of accommodations, then, reduces that variation in score caused by a disability but precludes standardization because accommodations, by their very nature, change the standard administration of the test. The essential question, then, is, *At what point does the change from standard test administration intended to improve score comparability actually change the task and harm score comparability, and how does one know the tasks and resulting scores are no longer comparable?*

Decisions about Score Comparability: The Role of Different Research Designs

Tindal (1998) proposed three models for making decisions about task comparability. These models are descriptive (Model 1), comparative (Model 2), or experimental (Model 3) in nature. The focus of all three models is to determine whether the construct that is being measured changes as a result of testing accommodations. The models represent a continuum of evidence for task comparability from the weakest evidence (descriptive) to the strongest (experimental).

In the *descriptive model*, evidence about task comparability relies on current policy for decision making. Policies, like those described above, provide descriptive evidence because they do not offer explanations or justifications of why judgments about accommodations in policy are made. Instead, policies can be created based on external information (e.g., the policies of other states) or may offer simple procedural recommendations for selecting and using accommodations without stating a rationale for doing so. As part of their policies, districts may track the implementation of various accommodations to understand how frequently accommodations are

selected and used for certain tasks. This tracking may provide further descriptive evidence about the relationship of tasks to one another when accommodations are provided.

The *comparative model* moves beyond descriptive information by relying on multiple sources of data to judge comparability of tasks (Tindal, 1998). These data provide retrospective information about the use of accommodations, judgments about their appropriateness, and performance outcomes to make relational statements about accommodations and performance. Because of its post hoc nature, however, the comparative model does not allow one to infer cause and effect relationships between accommodations and outcomes. Threats to internal validity—such as selection bias, participant maturation, or lack of comparison groups—compromise the utility of this approach for making meaningful and appropriate decisions about task comparability.

The *experimental model* for determining task comparability accounts for some of the aforementioned threats to validity by establishing a research design prior to the collection of data (Tindal, 1998). Experimental models include research designs and technically sound measurements to enhance the likelihood that meaningful inferences about outcomes can be made. Both group and single-case designs fall within the realm of the experimental model. To be informative, these designs must be able to provide inferences about cause and effect on the general population with statistical conclusions. In group designs, groups of students are compared with themselves or with other groups under varying test-taking conditions. In single-case designs, the performance of the same student under varying test-taking conditions is compared. The results of experimental studies are likely to provide the best evidence for task and score comparability because these studies overcome the limitations of the descriptive and comparative methods.

We believe, however, that most educators are unlikely to undertake an experimental approach to determine if a testing accommodation is effective and valid. They most often will be confronted with questions about the need and use of testing accommodations during an IEP meeting. They will need assistance in selecting, planning, and implementing testing accommodations with integrity. Our experience and research suggests they will benefit from a structured process for making these decisions (Elliott, Kratochwill, & Schulte, 1999). And as Phillips (1994) observed,

> When considering requested departures from standard testing conditions, measurement specialists might consider the following questions:
>
> 1. Will format changes or alterations in testing conditions change the skill being measured?
>
> 2. Will the scores of examinees tested under standard conditions have a different meaning than scores for examinees tested with the requested accommodation?
>
> 3. Would nondisabled examinees benefit if allowed the same accommodation?
>
> 4. Does the disabled examinee have any capability for adapting to standard test administration conditions?

5. Is the disability evidence or testing accommodations policy based on procedures with doubtful validity and reliability?

> Answering yes to any of these questions suggests that an accommodation is not appropriate. The final decision of whether to grant a requested testing accommodation will require the measurement specialist to balance the individual rights of the disabled requester against the obligation to maintain the integrity of the testing enterprise. The goals of providing maximum participation in society for the disabled and maintaining the validity of the testing program may be at odds. (p. 104)

Phillips' thoughts on testing accommodations have been insightful and constructive. However, we disagree with her position that saying "yes" to any of the above questions would indicate that an accommodation was inappropriate. In particular, we are concerned about her Question #3: *Would nondisabled examinees benefit if allowed the same accommodation?* This question in particular has influenced the investigative methods and decisions of several researchers (e.g., Fuchs et al., 2000; Tindal et al., 1998). Specifically, it has led to the design of experimental studies in which (a) students with and without disabilities are tested under conditions with and without accommodations and (b) a decision rule about the appropriateness of an accommodation is based on the expectation that accommodations should have more of an improvement effect (or boost, as Fuchs et al. call it) on the test scores of students with disabilities than on the scores of students without disabilities. We understand the logic of this paradigm and believe it does yield useful intra- and interindividual comparative information about the effects of testing accommodations for groups. In fact, our research team at the University of Wisconsin has been using this paradigm in our single subject–focused work for the past 5 years (Elliott, Kratochwill, & McKevitt, 2001; McKevitt & Elliott, 2001; Schulte, Elliott, & Kratochwill, 2000, 2001). Our concern, however, is with judging the appropriateness of an accommodation based solely on whether it brings about a greater change for a student with a disability than it does for the average of students without a disability. This is a questionable criterion that is not feasible for practitioners to apply and is likely to deny students with disabilities some appropriate accommodations.

An example may prove useful to illustrate our concern about the normative change criterion that both Phillips and Fuchs et al. are advocating for making a decision about the appropriateness of an accommodation. Consider the accommodation of extra time, perhaps the most frequently requested testing accommodation (Marquart, 2000). Now envision the use of extra time with two students, one a student with a disability who has poor math skills and the other a student without a disability with good to very good math skills. They both are given the same test with 30 math items, 20 multiple choice and 10 constructed response. The standard administration time is 45 minutes. Without accommodations, the students get raw scores of 10 and 24, respectively. A few days later, both students are given an equivalent form of the same math test, but this time they are both allowed an extra 30 minutes to take the test. Their raw scores are 13 and 28, respectively. This results in an improvement of 3 raw score points for the student with disabilities and an improvement of 4 raw score points for the student without a disability. Is the accommodation valid?

Given the above case, one would most likely decide that extra time is an invalid accommodation because the nondisabled student benefited more from the accommodation than the student with a disability. This result suggests, in keeping with the Phillips and Fuchs et al.

approach, that the accommodation affected construct-relevant variance rather than construct-irrelevant variance. But what if the test directions or manual never stated that the speed with which math problems are processed was an important skill the test measured? What if the content and performance standards with which the test was aligned never mentioned mathematical processing time? What would one decide about the validity of the testing accommodation of extra time if these guiding documents did not mention time for responding as an explicitly valued aspect of the mathematical skills to be assessed?

It is a simple fact that students use time differently and students have different mathematical skills. In the above example, both the student with a disability and the student without a disability benefited from the accommodation of extra time. (It should be noted, however, that many students, regardless of ability, do not benefit from extra time. See Marquart, 2000, for a review of this variable.) This result suggests that time is an important part of the construct being measured by the test, yet the publisher of the test did not intend it to be so. And according to the state's testing guidelines, only the student with the disability could be given the accommodation. This may seem unfair, but it is the current policy in the majority of states. The issues involved in this case concern fairness, a well-defined construct (mathematical computation and reasoning) to measure, and the valid measurement of the construct. Teachers cannot be expected to have a detailed grasp of the constructs being measured by the various tests that they are required to administer in a state's testing program. Of course, they must understand the purpose of the test, but it is the responsibility of the test publisher to clearly communicate what the test measures and to provide guidance about the access skills needed to take the test. Teachers can grasp the concept of access skills and will then be far more prepared to accommodate students appropriately.

Testing Accommodations: What We Know

Elliott et al. (2001) conducted a study designed to (a) describe the nature of information on testing accommodations listed on students' IEPs, (b) document the testing accommodations educators actually use when assessing students via performance assessment tasks, and (c) examine the effect accommodations have on the test results of students with and without disabilities. Participants in the study included 218 fourth-grade students from urban, suburban, and rural school districts. Of the 218 participants, 145 students did not have disabilities, and 73 students had disabilities in a variety of categories (including learning disabilities, speech and language impairments, etc.). The researchers asked teachers to list accommodations that would be helpful for each student who had a disability. Teachers used the *Assessment Accommodations Checklist* (*AAC*; Elliott et al., 1999), a list of accommodations often used in classroom and testing situations. Project staff and teachers then administered a set of math and science performance tasks to the students, utilizing an alternating treatments design, over the course of four 1-hour sessions.

These performance tasks were designed to draw on a full range of knowledge from each content area, were shown to have known psychometric values, and were found to be nearly equivalent and nonbiased among a group of over 200 students with disabilities. The tasks were scored on a 5-point continuum from "inadequate" to "exemplary" by trained project assistants using established criteria. All students with disabilities performed half of the tasks with accommodations and half of the tasks without accommodations. Students without disabilities

were separated into three groups by accommodation status: no accommodations, standard accommodations, and teacher-recommended accommodations. Students in the no accommodations group did not receive accommodations on any of the performance tasks. Students in the standard accommodations group received a standard set of accommodations. The alternating treatments design allowed for both intraindividual and intergroup comparisons without the need for baseline conditions. An individual's performance during the accommodated condition could be compared with his or her performance during the nonaccommodated condition. Also, the effect of accommodations on students with disabilities could be compared with the effect of accommodations on students without disabilities. The researchers used effect sizes to make comparisons both within individuals and between groups.

The Elliott et al. (2001) study indicated that the most common accommodations recommended by teachers were "verbal encouragement" and "read the directions," followed by "simplify language," "reread subtask directions," and "read test questions and content." Teachers typically recommended packages of between 10 and 12 accommodations for each student. The average effect size between accommodated and nonaccommodated conditions for students with disabilities was .88, approximately double the comparable effect size for students without disabilities. On an individual level, accommodations had "medium" to "large" positive effects for 78.1% of students with disabilities and 54.5% of students without disabilities. Accommodations had "small" effects or no effect on 9.6% of students with disabilities and on 32.3% of students without disabilities, and they had negative effects on 12.3% of students with disabilities and on 13.1% of students without disabilities.

The results of this study indicate that accommodations are recommended in packages for students, rather than independently. Accommodation packages have moderate to large effects on performance assessment scores for most students with disabilities and for some students without disabilities. This increase in scores for students without disabilities raises questions about the validity of the accommodations. If changes in testing procedure affect students without disabilities in the same direction and degree that they affect students with disabilities, these changes are not truly acting as accommodations.

Schulte et al. (2001) conducted a study to determine whether accommodations on standardized tests would affect students with disabilities differently than they affect students without disabilities. The authors predicted that accommodations would significantly improve the test scores of students with disabilities but would not significantly improve the test scores of students without disabilities. Participants in the study were 86 fourth-grade students, including 43 students with disabilities (entitled students with mild disabilities) and 43 students without disabilities. The students' performance was measured on two equivalent versions of the TerraNova math test, a math subtest aligned with the National Council of Teachers of Mathematics standards (NCTM, 1989).

Teachers of participants who had disabilities reviewed their IEPs to determine which accommodations the research team would use. Each student who did not have a disability was paired with a student who did have a disability, and the research team administered the TerraNova to the students in pairs. Both students in each pair received the accommodations outlined on the IEP of the student who had the disability. All students participated in a practice session to become familiar with the testing procedures and accommodations, and all students

took one version of the test with accommodations and one version of the test without accommodations. The researchers randomly assigned the order of accommodated and nonaccommodated conditions, as well as the pairs of students. The key independent variables in the study were testing condition (accommodated versus nonaccommodated) and disability status (with disability versus without disability) The dependent variables in the study were the scores from the TerraNova Multiple Assessments.

Both groups improved significantly when the accommodated condition was compared to the nonaccommodated condition. Students with disabilities benefited more from accommodations on multiple-choice questions, and both groups benefited equally on constructed-response questions. For multiple-choice questions alone, students with disabilities yielded an effect size of .41 between accommodated and nonaccommodated conditions, whereas students without disabilities yielded an effect size of 0. On constructed-response questions alone, those effect sizes were .31 and .35, respectively. On an individual level, there was essentially no difference between the effects of accommodations on students with disabilities and the effects of accommodations on students without disabilities. Twenty-seven out of 43 students with disabilities and 29 out of 43 students without disabilities achieved higher scores on the test when accommodations were available. Seventeen out of 43 students with disabilities and 16 out of 43 students without disabilities achieved higher proficiency levels on the test when accommodations were available. Twenty out of 43 students with disabilities and 21 out of 43 students without disabilities experienced no change in proficiency levels on the test when accommodations were available.

The finding that both groups of students experienced benefits from testing accommodations indicates that the changes in test procedure may be affecting both construct-relevant and construct-irrelevant variance. The differential interaction between accommodation group and question type could indicate that constructed-response questions are more difficult for all students and that accommodations remove barriers to these questions that are not present in multiple-choice questions. These findings reinforce the notion that research on testing accommodations must take an *individual perspective*, and that all students must take the tests in both accommodated and nonaccommodated conditions, for researchers to determine whether accommodations truly help performance.

In a dissertation study conducted by Marquart (2000), the use of an extended-time accommodation on a mathematics test was examined. Marquart predicted that (a) students with disabilities, but not students without disabilities, would score significantly higher in the extended-time condition than in the standard time condition, (b) students with low math skills, but not students with higher math skills, would score significantly higher in the extended-time condition, and (c) all student groups would perceive the extended-time condition as helpful in reducing anxiety, in allowing them to exhibit what they know, and in increasing their motivation to finish tests. Participants in the study included 69 eighth-grade students, 14 of their parents, and 7 of their teachers. Among the students, 23 were classified as having disabilities, 23 were classified as educationally at risk in the area of mathematics, and 23 were classified as students performing at grade level. Teachers used the *Academic Competence Evaluation Scales (ACES)* to classify students without disabilities as at risk or as performing at grade level. Student participants completed the TerraNova Multiple Assessments—Mathematics, as well as a survey about the effects of the extended-time accommodation. Each testing session included students

from each of the three groups. Marquart randomly assigned the order of conditions (accommodated and nonaccommodated) in which each student performed the test. When performing in the accommodated condition, students had up to 40 minutes to complete the test. When performing in the nonaccommodated condition, students had 20 minutes to complete the test. Parents and teachers of students in the study also completed the survey about the effects of the extended-time accommodation.

Marquart found that the effect of the extended-time accommodation was not significant for students without disabilities, who yielded an effect size of .34. The accommodation was not significant for students with disabilities, either, as their effect size was .26. The three groups (students with disabilities, at risk, and at grade level) were not significantly different in their amount of change between accommodated and nonaccommodated conditions, either. When students without disabilities were considered as at-risk and grade-level groups, the students in the at-risk group experienced an effect size of .48 between accommodation conditions, and students in the grade-level group experienced an effect size of .20.

However, according to the survey, most students felt more comfortable, were more motivated, felt less frustrated, thought they performed better, reported that the test seemed easier, and preferred taking the test under the extended-time condition. Most teachers (88%) but few parents (21%) indicated that a score from an accommodated test is as valid as a score for the same test without accommodations. Many parents (43%) but no teachers believed that the score from an accommodated test is less valid, and some members from both groups (parents = 36%, teachers = 12%) were uncertain. Most members of each group (teachers = 63%, parents = 56%) believed that if accommodations are used on a test, those accommodations should be reported with the test results.

McKevitt and Elliott (2001) studied the effects of testing accommodations on standardized reading test scores and the consequences of using accommodations on score validity and teacher and student attitudes about testing. The following predictions were tested: (a) teachers would select accommodations they considered valid and fair for use on standardized reading tests; (b) individualized packages of testing accommodations, including a read-aloud accommodation, would have a positive impact on the reading test scores of students with disabilities, but not on the scores of students without disabilities; (c) students with disabilities would score higher when the test was read aloud to them than when other accommodations were used; and (d) students would perceive the accommodations to be helpful, and teachers would have a positive attitude about testing and accommodations. Although read-aloud accommodations are considered invalid by the testing policies in many states, to date there have been no published studies that actually analyzed their effects on reading test performance. To test the above hypotheses, the reading performance of 79 eighth-grade students was tested on the TerraNova Multiple Assessment Reading Battery—Research Version (Form A; CTB/McGraw Hill, 1999). Forty of those students were diagnosed with an educationally defined disability and received special education services in the area of reading and/or language arts. The other 39 students were general education students used for comparison purposes. Four special education teachers and one general education teacher participated by recommending testing accommodations for these students using the *Assessment Accommodations Checklist* (Elliott et al., 1999). They also rated students' reading achievement levels using the *Academic Competence Evaluation Scales* (DiPerna & Elliott, 2000). An additional 43 teachers and all tested students

completed surveys about their perceptions of and attitudes about testing accommodations and standardized testing.

Once students were identified, they were divided into two groups (students with disabilities and students without disabilities). Within those groups, students were then divided into two test conditions (students receiving teacher-recommended accommodations and students receiving teacher-recommended accommodations plus a read-aloud accommodation). Students in each group and each condition completed two alternate parts of the reading test—one with accommodations (either teacher-recommended accommodations or teacher-recommended accommodations plus read-aloud) and the other without accommodations. The part of the test that was accommodated was determined by random assignment. This design yielded a repeated measures ANOVA analysis with effect size calculations used to test the predictions. Overall, the results of the McKevitt and Elliott (2001) study indicated mixed support for the predictions. First, as predicted, teachers selected accommodations they considered valid and fair for use on a standardized test. They did not recommend using a read-aloud accommodation, as this accommodation would interfere with the purpose of the test (i.e., to measure reading ability) and thus would invalidate resulting test scores. Next, the accommodations that teachers recommended did not significantly affect test scores for either group of students. However, the read-aloud accommodation, when used in addition to those recommended by the teacher, did positively and significantly affect test scores for both groups of students. There was no differential benefit from the read-aloud accommodation, indicating overall score boosts for both groups of students, rather than the boost only for students with disabilities that was predicted.

Interestingly, there was much individual variability in the accommodation effects. As indicated by effect size statistics, the accommodations positively affected the scores for half of all students with disabilities and 38% of all students without disabilities. Furthermore, neither group of students scored significantly higher when the test was read aloud to them as compared to the groups that received other accommodations. Although the read-aloud accommodation helped both groups compared to their own performance without accommodations, there was not a significant effect from the read-aloud when groups receiving the read-aloud were compared to those receiving only the teacher-recommended accommodations.

Finally, McKevitt and Elliott (2001) found that students and teachers had mixed feelings about the accommodations. Students were generally positive about their use but expressed some concern that the read-aloud accommodation was too difficult to follow. Teachers felt positive about the use of accommodations for students with disabilities but also were concerned about how accommodations would affect test score validity. Teachers reported they rely primarily on professional judgment when making accommodations decisions, rather than on their own empirical testing of accommodations effects. Therefore, it is very important to ensure teachers are knowledgeable about the use and effects of testing accommodations.

In summary, the McKevitt and Elliott (2001) study contributed to the increasing evidence that accommodations may have positive or negative effects for individual students with *and* without disabilities. It also lends support to the popular belief that reading a reading test aloud to students as an accommodation invalidates test scores. The lack of differential boost (i.e., the finding that both groups of students profited from a read-aloud accommodation) observed in the study is one piece of evidence of the invalidating effect of a read-aloud accommodation. But

the lack of differential benefit alone may not be sufficient to conclude invalidity of scores resulting from the use of accommodations. In the case of the students receiving the teacher-recommended accommodations alone, a differential boost also was not observed and scores did not improve significantly for either group. One may not conclude, however, just from this evidence that the accommodations were invalid. The accommodations may have served to remove a disability-related barrier for the student tested, yet still may not have had a significant effect on scores. Thus, evidence to support the validity of accommodations needs to come from multiple sources, examining student factors, test factors, and the accommodations themselves.

Conclusions

Among measurement experts, there is a consensus that the purpose of using testing accommodations is to increase the validity of the inferences one makes from the test scores of students with disabilities. This relationship between testing accommodations and the validity of the resulting scores is theoretically sound, yet only a few data-based reports of this relationship have been published. Meanwhile, educators on the front lines who are administering tests to all students must make test participation decisions for students with disabilities and then select and implement testing accommodations that are judged a priori to be valid. Educators are very capable of making participation decisions and are knowledgeable about the instructional accommodation needs of their students, but are seriously challenged to make accommodation decisions that lead to "good" (i.e., valid) test scores.

In our work with hundreds of educators who are responsible for administering large-scale assessments to students with disabilities, we have found that they need a wide range of knowledge to select and use testing accommodations reliably. In particular, these educators collectively need to have:

- Knowledge of the abilities and disabilities of the students they are testing,
- Knowledge about the students' instructional accommodations,
- Knowledge about the state's or district's testing guidelines,
- Familiarity with the test's item content and format,
- An understanding of the concept of validity and what it means to invalidate a test score, and
- Knowledge of any previous accommodations successfully used with the students.

Note that the educators with whom we have interacted have rarely requested a summary of research on the effective use of testing accommodations. Perhaps they recognize there is little research on this issue. If research is going to guide practice, researchers and test publishers interested in seeing all students participate meaningfully in assessments will need to help frontline educators more. These educators need to understand which testing accommodations are most likely to be valid and how they can go about making decisions about the validity of testing accommodations for individual students prior to testing. This is relatively difficult and very important work because "good" test scores are hard to come by and highly valued.

References

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

Anderson v. Banks, 530 F. Supp. 472, 510-11 (S.D. Ga. 1981).

Board of Educ. v. Ambach, 436 N.Y.S.2d 564 (1981), *aff'd with modifications,* 458 N.Y.S.2d 680 (App. Div. 1982), *aff'd,* 457 N.E.2d 775 (N.Y. 1983).

Brookhart v. Illinois State Bd. of Educ., 534 F. Supp 725 (C.D. Ill. 1982), *rev'd,* 697 F.2d 179 (7th Cir. 1983).

CTB/McGraw-Hill. (1999). *TerraNova multiple assessment reading battery—research version.* Monterey, CA: Author.

CTB/McGraw-Hill. (2000). *Guidelines for using the results of standardized tests administered under nonstandard conditions.* Monterey, CA: Author.

Debra P. v. Turlington, 474 F. Supp. 244 (M.D. Fla. 1979), *modified and remanded,* 644 F. 2d 397 (5th Cir. 1981), *on remand,* 564 F. Supp. 177 (M.D. Fla. 1983), *aff'd,* 730 F.2d 1405 (11th Cir. 1984).

DiPerna, J. C., & Elliott, S. N. (2000). *Academic competence evaluation scales*. San Antonio, TX: The Psychological Corporation.

Elliott, J., Ysseldyke, J., Thurlow, M., & Erickson, R. (1998). What about assessment and accountability? Practical implications for educators. *Teaching Exceptional Children, 31*(1), 20-27.

Elliott, S. N., & Braden, J. P. (2000). *Educational assessment and accountability for all students: Facilitating the meaningful participation of students with disabilities in district and statewide assessment programs.* Madison, WI: Wisconsin Department of Public Instruction.

Elliott, S. N., Braden, J. P., & White, J. (2001). *Assessing one and all: Educational accountability for students with disabilities*. Alexandria, VA: Council for Exceptional Children.

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*(1), 3-24.

Elliott, S. N., Kratochwill, T. R., & Schulte, A. G. (1998). The assessment accommodations checklist: Who, what, where, when, why, and how? *Teaching Exceptional Children, 31*(2), 10-14.

Elliott, S. N., Kratochwill, T. R., & Schulte, A. G. (1999). *Assessment accommodations checklist.* Monterey, CA: CTB/McGraw Hill.

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.

Harker, J. K., & Feldt, L. S. (1993). A comparison of achievement test performance of nondisabled students under silent reading and reading plus listening modes of administration. *Applied Measurement in Education, 6,* 307-320.

Hawaii State Dep't of Educ., 17 EHLR 360, 361 (OCR Oct. 1990).

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

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

Huesman, Jr., R., & Frisbie, D.A. (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.

Lewis, D. M., Green, D. R., & Miller, L. (1999, June). *Using differential item functioning analyses to assess the validity of testing accommodated students with disabilities.* Paper presented at the National Conference on Large-scale Assessment of the Council of Chief State School Officers, Snowbird, UT.

Marquart, A. M. (2000). *The use of extended time as an accommodation on a standardized mathematics test: An investigation of effects on scores and perceived consequences for students with various skill levels*. Unpublished manuscript, University of Wisconsin–Madison.

McDonnell, L. M., McLaughlin, M. J., & Morison, P. (Eds.) (1997). *Educating one and all: Students with disabilities and standards-based reform.* Washington, DC: National Academy Press.

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

Meloy, L. L., Frisbie, D., & Deville, C. (2000, April). *The effect 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.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

Phillips, S. E. (1993). Testing condition accommodations for disabled students. *West's Education Law Quarterly, 2,* 366-389.

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

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

Schulte, A. G., Elliott, S. N., & Kratochwill, T. R. (2000). Educators' perceptions and documentation of testing accommodations for students with disabilities. *Special Services in the Schools, 16,* 35-56.

Schulte, A. G., Elliott, S. N., & Kratochwill, T. R. (2001). Experimental analysis of the effects of testing accommodations on students' standardized achievement test scores. *School Psychology Review, 30*(4), 527-547.

Southeastern Community College v. Davis, 442 U.S. 397, 406 (1979).

Thurlow, M. L., House, A., Boys, C., Scott, D., & Ysseldyke, J. (2000). *State participation and accommodation policies for students with disabilities: 1999 update*. Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., Seyfarth, A. L., Scott, D. L., & Ysseldyke, J. E. (1997). *State assessment policies on participation and accommodations for students with disabilities: 1997 update* (Synthesis Report 29). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

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

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

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*, 439-450.