Extended Time Testing Accommodations for Students With Disabilities: Answers to Five Fundamental Questions

Benjamin J. Lovett
Elmira College

Extended time is one of the most common testing accommodations provided to students with disabilities. It is also controversial; critics of extended time accommodations argue that extended time is used too readily, without concern for how it changes the skills measured by tests, leading to scores that cannot be compared fairly with those of other students. Advocates argue, instead, that many students with disabilities are only able to demonstrate their skills with extended time. This article reviews a wide variety of empirical evidence to draw conclusions about the appropriateness of extended time accommodations. The evidence reviewed raises concerns with the way that extended time accommodations are currently provided, although the same literature also points to potential solutions and best practices.

Keywords: validity, reliability, high stakes testing, psychometrics.

Providing testing accommodations to students with disabilities has become increasingly common in a variety of settings. In K–12 settings, legal mandates concerning high-stakes testing have led to greater use of accommodations (Cox, Herner, Demczyk, & Nieberding, 2006). On admissions and certification tests in postsecondary settings, examinees’ requests for accommodations have increased, leading to greater use on these tests as well (Ranseen & Parks, 2005). Controversies about the practice continue (e.g., Lerner, 2004; Zuriff, 2000), especially regarding extended testing time, one of the most commonly requested and provided accommodations (Bolt & Thurlow, 2004; Stretch & Osborne, 2005). This article aims to inform the controversy by identifying and tentatively answering the key questions regarding the appropriateness of extended time testing accommodations for students with disabilities.

The Controversy Over Extended Time Testing Accommodations

The practice of standardized assessment of student learning is continually evolving in response to novel concerns. Testing accommodations were designed to address two distinct but related concerns: the need for participation of all students in large-scale assessment programs and the need for fairness in measuring the achievements of students with special needs. Recently, issues of participation have
been a special concern, in part because of work on achievement gaps. Researchers and policymakers have been interested in why groups of students differ in achievement, and standardized assessment has been viewed as a tool for better understanding this issue (e.g., R. S. Johnson, 2002; Lee, 2006). In the case of certain student groups (viz., students with disabilities and English language learners), one barrier to the exploration of achievement gaps has been the low rate of participation by these students in standardized assessment programs (Thurlow, Albus, Liu, & Rivera, 2006; Thurlow & Thompson, 2004). Accommodations are viewed as a tool to increase participation and, through that participation, to increase researchers’ understanding of why these students tend to have more academic problems (Kieffer, Lesaux, Rivera, & Francis, 2009; Koretz & Barton, 2003–2004). Accommodations are also viewed as a tool for ensuring the fair treatment of students with special needs. The skills of students with disabilities or low levels of proficiency in English may be underestimated by conventional standardized tests (Goh, 2004), thus denying these students benefits associated with high performance or subjecting them to burdens associated with low performance.

The psychometric logic of extended time accommodations for students with disabilities is fairly straightforward. Disability conditions may keep students from demonstrating their skills within standard testing time limits; a student with a learning disability in reading, for instance, may read the test items too slowly to complete the test in the allotted time. Testing these students under standard conditions leads to a problem of construct-irrelevant variance (Haladyna & Downing, 2004) in the resulting test scores; the scores will vary across students in part because of variability in students’ reading speed rather than solely because of variability in the construct under investigation (knowledge of American history, for example). Extended time accommodations are meant to reduce this construct-irrelevant variance by keeping construct-irrelevant skills such as reading speed from being prerequisites for accessing the test content (Fuchs & Fuchs, 2001).

Accommodations such as these are now required by law and are given based on a student’s disability diagnosis or other classification (for reviews of these legal foundations, see Colker & Milani, 2005; Phillips & Camara, 2006). Not only have courts held that certain alterations to tests are part of the “reasonable accommodations” guaranteed by the 1990 Americans With Disabilities Act and Section 504 of the 1973 Rehabilitation Act, but special education legislation specifically requires the use of assessment accommodations. The 1997 reauthorization of the Individuals With Disabilities Education Act (IDEA) requires each state to demonstrate that its “children with disabilities are included in general state and district-wide assessment programs,” and to make wide participation in assessments more feasible, IDEA permits “appropriate accommodations, where necessary” (§ 612). More recent federal education legislation, including the No Child Left Behind Act of 2001 (P.L. 107–110), has maintained this provision, and assessment accommodations are again required whenever they are deemed necessary (Thurlow & Thompson, 2004).

Despite these aims, extended time accommodations are a hotly debated practice. At one extreme of this debate, certain critics (e.g., Lerner, 2004; Lichtenberg, 2004) have charged that extended time is deliberately sought and obtained by high-ability affluent students after receiving dubious disability diagnoses, and that the accommodation leads to an unfair advantage over these students’ classmates. More
moderate commentators (Koretz & Barton, 2003–2004; Pitoniak & Royer, 2001) have expressed the concern that extended time accommodations may be given too readily, without fully considering the effects of the extra time on how the resulting test scores should be interpreted. In addition, certain scholars (Earle & Sharp, 2000) have argued that extended time accommodations may ironically keep instructors, test administrators, and other professionals from considering whether timed exams are an appropriate measure of skills for examinees more generally; by proffering extended time accommodations, the appropriateness of the time limits for other examinees is never called into question.

Unfortunately, strong opinions for and against accommodations are rarely backed by empirical evidence, which sometimes fails to support intuitions about accommodation effects. This is not only in the case of students with disabilities; consider a recent meta-analysis of testing accommodations for English language learners. Kieffer et al. (2009) reviewed research on the effectiveness and appropriateness of seven different accommodations for English language learners. These investigators found that only one of the seven accommodations significantly benefited English language learners’ scores, despite the intuitive relevance of all seven of the accommodations to improving these students’ performance.

Validity and Extended Time Accommodations

Much of the controversy over extended time accommodations can be understood as consisting of competing sets of claims about the effect of extended time on validity—specifically, the validity of interpretations made on the basis of scores on tests taken under extended time conditions. The most recent Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), the consensus document designed to encourage valid test uses and score interpretations, acknowledges the cases made on both sides of the debate. On one hand, the Standards notes that “aspects of the testing process that pose no particular challenge for most examinees may prevent specific groups of individuals from accurately demonstrating their standing with respect to the construct of interest (e.g., due to disability or language background)” (p. 75). On the other hand, the Standards notes that although accommodations “are made with the intent of maintaining score comparability, the extent to which that is possible may not be known. Comparability of scores may be compromised, and the test may then not measure the same constructs for all test takers” (p. 61).¹

Because validity always concerns a specific use of a test or a specific interpretation made on the basis of a test score (Kane, 2006; Messick, 1995), the typical uses and interpretations of tests on which extended time accommodations are given must be considered. One common test type is postsecondary admissions testing, as seen in the Scholastic Aptitude Test (SAT), Graduate Record Examination, and related tests. The scores are typically interpreted as showing an examinee’s general academic skill or aptitude levels, and are used to predict his or her performance in a future educational setting (Zwick, 2006). Because these tests yield norm-referenced score interpretations, an examinee’s score only has meaning in comparison to the scores of other examinees. A second test type is certification and licensure testing, as seen in Praxis exams, the U. S. Medical Licensure Examination, and the Multistate Bar Examination for attorneys. Scores on these tests are interpreted as
showing an examinee’s skills and/or knowledge in a fairly narrow area of competence and are used to ensure that the examinee’s performance in real-world settings will be above a certain minimum standard. A third test type is K–12 statewide achievement testing. Interpretations and uses of scores from these tests vary, but interpretations may include inferences about students’ academic skills, the quality of instruction being provided to students, and the success of the academic program of a school or school district as a whole. A fourth and final major test type in educational measurement is made up of the classroom assessments designed and administered by K–12 teachers and instructors at postsecondary institutions. Here, the interpretations are primarily about students’ acquisition of the knowledge and skills taught in the classroom, and the tests are used to assign course grades to students (Airasian & Russell, 2008). Students with disabilities are frequently given extended time accommodations on all four of these types of tests, and the potential effects of extended time on validity are not uniform across these tests.

Two features of a testing procedure are especially relevant in determining the potential effects of extended time on validity. The first concerns the type of score interpretations typically used by a testing procedure. A distinction is frequently made between norm-referenced scores (e.g., percentiles), which locate an examinee’s performance relative to that of other examinees, and criterion-referenced scores, which do not depend directly on the scores of other examinees and are granted a direct interpretation (e.g., mastery). As Haertel (2003) noted, the distinction has substantial implications for accommodations. Norm-referenced score interpretations require direct comparisons between the scores of different examinees, and so the performance of examinees taking the test under different conditions (e.g., time limits) must be expressed using the same metric. For this reason, norm-referenced testing requires the strongest evidence of comparability between standard and accommodated tasks, and in this situation accommodations should be avoided unless absolutely necessary and supported by evidence of comparability. Of the types of major testing programs discussed above, postsecondary admissions testing typically involves norm-referenced score interpretations, whereas certification and licensure testing occasionally do, and K–12 accountability testing and classroom assessment rarely do.

Criterion-referenced score interpretations are actually of several types (Haertel, 2003; Haertel & Lorié, 2004), two of which are especially relevant here. In one type, the test is viewed as a direct sample of a larger body of test-like tasks. For instance, nursing students may be required to take an exam in which medication doses are calculated. If the items are a representative sample of the real-world dosage calculation tasks, the score need not be compared with other examinees’ performance to be interpreted; instead, a performance standard such as 95% correct can be used to give each examinee’s score a direct meaning. The appropriateness of an accommodation in this situation depends on the generalizability of the accommodation to the real-world tasks. If the nursing student will be able to take more time to calculate dosages in a real-world setting, the score can be interpreted with reference to the same standards as other examinees’ scores, and so the extended time would not lead to a loss of comparability.

In another type of criterion-referenced score interpretation, the score is interpreted as indicating the level of an underlying construct (e.g., eighth-grade mathematics knowledge). During development of the test, relevant experts characterize
various levels of the construct, and a cut score on the test is set for each level based on experts’ judgments of how examinees at different construct levels would perform on the test items (Cizek & Bunch, 2007). Most group-administered achievement tests used for school accountability purposes are developed this way, with each construct level (e.g., basic, proficient) tied to a different cut score. In this situation, scores obtained under extended time conditions may or may not be comparable to those of other examinees, in the sense that the same cut scores can be used. In the optimal case, the process of test development would be redone, with expert judges’ considering how examinees at various construct levels would respond to items under accommodated conditions, and their judgments would lead to the setting of cut scores that may or may not be identical to the cut scores already set (Haertel, 2003). Because this is typically not done, comparability remains an open question. Criterion-referenced score interpretations, then, do not eliminate comparability concerns entirely but may require different approaches to ensure that common cutoffs are appropriate for examinees who take tests under standard and accommodated conditions.

In addition to the type of score interpretations used, the speededness of a test affects the validity of inferences made on the basis of scores obtained under extended time conditions. Tests vary on a continuum from pure power tests, which are essentially untimed, and speed tests, which consist of items of trivial difficulty completed under severe time limits. Although the intentions of test developers are often investigated to gauge speededness, it is actually an independent characteristic of a set of test scores. Formally, a speeded test is one in which “examinees’ scores are determined by the amount of items attempted as well as the accuracy of responses” (Lu & Sireci, 2007, p. 30). The SAT, for instance, is partially speeded; although the developers describe the time limits as being for administrative convenience, not all examinees finish the exam in the allotted time (College Board, 2009; Mandinach, Bridgeman, Cahalan-Laitusis, & Trapani, 2005). K–12 achievement tests used for accountability purposes are generally untimed (within broad limits; typically, students must complete each exam section in a single day), and so they approach pure power tests (J. L. Elliott & Thurlow, 2006).

Both the actual and intended speededness of tests are relevant to making decisions about extended time accommodations. In the case that a test is designed to be partially speeded (to ensure, for instance, that examinees can function under time pressure or to examine the automaticity of their skills), extended time accommodations would generally be inappropriate because speed is one of the constructs of interest (Phillips, 2002). In the more common case that a test is unintentionally speeded, extended time accommodations still lead to problems because the accommodations are only given to students with disabilities, although nondisabled students may also be unable to complete the test in the allotted time.

Five Questions About Extended Time

Score interpretation and speededness are, then, the two main features of testing that affect the appropriateness of extended time accommodations. Before empirical evidence is examined, however, these two issues must be reformulated into more specific questions that can be asked about extended time accommodations. In an influential discussion of the general testing accommodations controversy, Phillips (1994) provided five questions that she suggested be asked before a test is
administered in any nonstandard way, and in this review I use these questions to organize the extant research on extended time accommodations.\(^2\)

First, does extended time alter the construct validity of inferences based on test scores? As was noted above, both norm- and criterion-referenced score interpretations assume that all examinees’ scores are indications of the construct(s) of interest rather than ancillary constructs. Even when examinees’ scores are not directly compared, there must be validity evidence for scores obtained under both standard and accommodated conditions. The joint testing Standards (American Educational Research Association et al., 1999) describes many sources of validity evidence, and in many cases this evidence can be compared for standard and accommodated conditions. If scores obtained under standard time and extended time conditions both have similar validity evidence, this would generally support the use of the accommodation. In addition, in certain instances where scores are compared directly across examinees, additional types of validity evidence would be helpful.

A second question concerns whether nondisabled examinees benefit from extended time. Extended time accommodations assume that a test would only be significantly speeded for the examinees who are receiving the extra time, so an examinee’s disability diagnosis is important insofar as it signals the need for extra time. Therefore, if at least some nondisabled examinees would also benefit from the accommodation (i.e., if the test is speeded more generally, even if unintentionally), it would be inappropriate to withhold the accommodation from them while allowing examinees with disabilities extra time.

A third question asks if the students with disabilities could adapt to the standard testing conditions. Given the psychometric and logistical issues that any accommodation raises, it would obviously be preferable if the accommodation were unnecessary. In the case of extended time, it would be helpful to know precisely why certain students with disabilities need extra time and whether those needs can be modified. Especially in K–12 education, students with disabilities receive individualized programming designed to remediate skill deficits and treat other problems associated with their disabilities. Presumably, the goals that they and their teachers and therapists are working toward could include the ability to take tests under standard conditions. Although the capacity for some examinees with disabilities to adapt to standard time allotments would not make extended time inappropriate in all cases, it would be helpful for test users to know whether the administrative burdens (to examinees as well as test administrators) associated with extended time could be reduced. Moreover, if there are any validity concerns at all regarding extended time, it would of course be preferable to avoid giving the accommodation unnecessarily.

A fourth question concerns the disability diagnoses that lead to eligibility for extended time accommodations. How reliable and accurate are the procedures for identifying examinees with disabilities? This identification issue is often neglected in discussions of accommodations, but as Koretz and Hamilton (2006) noted, “Inconsistencies in identification are problematic because they indicate that the benefits of statutes such as IDEA . . . are misallocated” (p. 563). In many settings, the documentation of a disability is necessary to obtain extended time, and so the appropriateness of the accommodation decision hinges in part on the degree to which the disability diagnosis can be trusted. If the diagnoses are unreliable and frequently inaccurate, errors in both directions can occur, with students who need
extra time being denied the accommodation while students who do not need the extra time are provided with it anyway.

Finally, a fifth question: beyond the issue of diagnosis, are the procedures used to make accommodation decisions of adequate technical quality? In postsecondary settings, not all extended time requests are granted, and in K–12 settings, extended time is made a part of some students’ individualized educational programs but not others’. To be defensible, these decisions must take account of relevant student variables while ignoring irrelevant variables. If data from diagnostic tests (used to make the disability identification) are used in the accommodation decision, research should support a relationship between those data and the examinees’ timing needs. If the decisions are not sufficiently based on relevant factors or if they are not replicable (i.e., the chance of being granted extended time depends on who is making the decision), then the accommodation may not be appropriate.

In the review that follows, the relevant evidence concerning each of these five questions is examined, often from diverse lines of research. Following analysis of the five questions, tentative answers are offered. The review and analysis are limited to extended time accommodations, and this decision was made for several reasons: First, extended time is one of the most common accommodations; second, extended time has received a greater share of criticism by skeptics of accommodations; and third, extended time is often provided to students with higher incidence disabilities, allowing for more research to be conducted and evidence to be reviewed. The conclusions drawn here cannot be generalized to other accommodations, but it is hoped that the types of evidence highlighted here might lead to similar reviews for other accommodations.

**Literature Search Strategy**

The literature search began with an electronic search (in February 2009, updated in November 2009) of both PsycINFO and ERIC with search terms *extended time* and *accommodations*. This search retrieved 66 items, and the abstract for each one was examined to determine its relevance. Of these 66 items, 17 appeared to be relevant, and copies of the full publications were procured; after more extensive examination, all 17 were retained for use in the review. An additional search in which *extended* was replaced by *extra* returned 3 additional relevant articles. Together, these 20 articles, along with additional research cited in the reference lists of these articles, were used to investigate the first two of the five questions covered in this article (construct validity and differential benefits from extended time). For these two questions, it was essential that each relevant study be located and included.

In contrast, the other three questions (concerning the ability to adjust to standard timing conditions and the technical adequacy of diagnostic and accommodation decisions) required investigation of broader areas of research literature and only representative use of specific empirical findings. For instance, a comprehensive analysis of the literature on the diagnosis of disabilities is far beyond the scope of this article; nevertheless, a brief and representative survey of the literature is crucial to an understanding of decisions regarding extended time accommodations.

To locate research on the other three questions, a variety of different searches were performed at various times during this project. For example, once test anxiety was identified as a possible factor leading to requests for extended time accommodations, a search of the test anxiety literature was performed. Wherever
possible, review articles and related treatments in books and book chapters were used to help ensure that conclusions from wide bases of research literature would be warranted. In addition, representative empirical studies were examined for possible inclusion. This process of broader searches was performed for the first two questions covered as well to supplement the studies identified through the initial searches. For instance, once reading fluency was identified as a factor affecting performance on timed tests, a search of the reading fluency literature was conducted for use in the construct validity section. At the end of the search process, a wide variety of publications had been found from many different areas of education and psychology.

**Construct Validity**

Do test scores obtained under extended time have adequate construct validity evidence? This issue can be investigated in two ways, broadly speaking. First, the psychometric characteristics of scores obtained under extended time can be compared with the characteristics of scores obtained under standard time conditions. This will be important information regardless of whether examinees’ scores will be directly compared; even if they will not, it is imperative to demonstrate that similar validity evidence (signaling psychometric adequacy) exists for both kinds of scores. A second type of investigation involves conceptual analysis of the test tasks under different timing conditions to determine whether extended time accommodations would affect the construct being measured by the test; this is especially important in the case where examinees’ scores are directly compared, although again, even if they will not, test users should understand how accommodations may affect the measurement of underlying constructs.

**Comparable Psychometric Characteristics**

In 1988, Willingham and colleagues published a comprehensive review of the psychometric characteristics of SAT and Graduate Record Examination scores obtained under various administration conditions. Although their analyses sometimes combined extended time with other accommodations and similarly combined various disability groups, their findings remain pertinent. They found the effects of accommodations on test reliability, factor structure, predictive validity, and item functioning to be generally quite small, and they concluded that so long as scores on these admissions tests are interpreted in the context of other information (e.g., examinees’ grade point averages), there were few concerns about validity. Willingham et al. were tentative and humble in their conclusions, calling for more focused research, but although issues of score comparability have only grown more controversial in the intervening years, surprisingly little work has been done. However, seven recent studies have looked specifically at these issues with regard to extended time accommodations, examining three psychometric issues: predictive validity, factor structure, and item functioning. Although different tests are designed with different purposes in mind, these three issues are almost always relevant to determining the construct validity of a test.

**Predictive validity.** Cahalan, Mandinach, and Camara (2002) used the SAT scores of examinees with learning disabilities who received extended time accommodations \((n = 241)\) to predict the examinees’ 1st-year college grade point averages.
(GPAs). When compared to nondisabled students \((n = 33,771)\), the correlation was lower \((r = .35\) for students with learning disabilities using extended time but \(r = .48\) for nondisabled students\), which suggests less predictive validity evidence. Moreover, there was a consistent bias in prediction for male examinees with a learning disability such that their college GPA was overpredicted by .21 GPA points (no bias was found for female examinees).

Thornton, Reese, Pashley, and Dalessandro (2001) conducted a similar study using scores on the Law School Admissions Test (LSAT) to predict 1st-year law school GPA. Again, the correlation was somewhat lower for students who used extended time accommodations \((n = 1,249; r = .34)\) than for other students \((n = 121,607; r = .40)\). Thornton et al. provided separate results for each disability group, and the two largest groups, students with learning disabilities and attention deficit hyperactivity disorder (ADHD), who received extended time accommodations \((ns = 787\) and 172, respectively) each showed validity coefficients of \(r = .34\) as well. In addition to lowered predictive validity, there was overprediction of GPA again for students with extended time accommodations (approximately 5.7 percentage points); although Thornton et al. did not provide separate analyses by gender, they reported that for students with learning disabilities and ADHD, GPA was overpredicted by more than one half of a standard deviation.

These two predictive validity studies suggest that extended time accommodations may inflate test scores unduly, such that students’ later performance in the “real world” (in fact, performance on papers, projects, and classroom tests in college or law school) is less than what would be expected given their admissions test scores. However, because neither of these studies examined students with disabilities who did not receive any accommodations, it is difficult to say whether the disability conditions themselves, rather than the extended time, played a role in the overprediction of performance.

Internal factor structure. Huesman and Frisbie (2000) conducted a principal components analysis of the reading test of the Iowa Tests of Basic Skills for groups of sixth-grade students with and without learning disabilities \((ns = 129\) and 409, respectively) at both standard and extended time. The factors extracted in each group were generally quite similar, and although quantitative indexes were not used to compare factor structures, the data for students with learning disabilities had two significant factors (factors with eigenvalues > 1.0), whereas the data for nondisabled students had one significant factor. The extended time accommodations appeared to reduce the presence of the second factor for students with learning disabilities, making the factor structure of the test more comparable to that of nondisabled students.

More recently, Lindstrom and Gregg (2007) used confirmatory factor analysis (CFA) to examine the factor structures of the sections of the new SAT for nondisabled students who took the test at standard time and students with disabilities (learning disabilities, ADHD, or both) who took the test under extended time conditions.\(^3\) To make the two groups equal in size, a random sample of nondisabled examinees was taken, leading to 2,476 students in each group. These investigators generally found invariant factor structures across the two groups for each section (critical reading, mathematics, and writing), concluding that the resulting test scores “reflect the same underlying constructs across both samples” (p. 92).
Lovett

These studies suggest that extended time accommodations may not change the internal structure of the tests, except to make them more similar to those of nondisabled students. However, comparable factor structures are not sufficient to demonstrate the appropriateness of extended time accommodations. These studies’ analyses focused on the relationships between sets of items, but extended time could inappropriately inflate scores on all item sets while maintaining the relationships between the item sets; if extended time accommodations were tantamount to adding a constant to each item set score, the relationships between item sets may not change but the scores would still not be reflective of the individual’s true skills when compared to scores of other individuals who did not receive the accommodation.

**Differential item functioning (DIF).** Bolt (2004) used item response theory (IRT) models to examine the effects of several accommodations on item functioning for elementary and high school students with various disabilities. Bolt modeled the relationship between individual item responses and estimated ability levels for students who took a test under different conditions, to detect and quantify DIF across conditions. Any DIF would suggest that performance on the item showed a different relationship with students’ ability levels across conditions. Bolt found that between 16 and 44% of test items showed significant DIF between students receiving extended time (ns ranging from 442 to 824, depending on sample) and nondisabled students (all ns were 1,000), depending on the dataset and subject area examined. A small number of items (between 4 and 8%) exhibited large effect sizes for the DIF. Moreover, students with disabilities who completed the tests without accommodations exhibited less DIF than did those who used accommodations, which suggests that DIF was not due solely to students’ disability status.

Cohen, Gregg, and Deng (2005) also used IRT models, comparing students with learning disabilities who were given extended time on a high-stakes mathematics test to nondisabled students who took the test under standard time conditions (n = 1,250 in each group). These investigators found that more than 75% of the multiple-choice items showed DIF between the two conditions, but by using more complex IRT models that located groups of students whose item responses had similar IRT model parameters and then comparing the students in those groups, Cohen et al. concluded that the DIF was only in part due to the accommodations per se and that differences in profiles of mathematics competence was the main cause of the DIF.

Finally, Finch, Barton, and Meyer (2009) disentangled issues of disability status and accommodation use directly by examining IRT models of DIF for third-through eighth-grade students with disabilities, only some of whom received accommodations (cell ns ranging from 92 to 346). Because different students received different sets of accommodations, separate IRT models were calculated for each accommodation to determine if it contributed to DIF. Across all grades and two subject areas (language arts and mathematics), extended time accommodations were found to contribute to significant DIF for 3 items (out of 343 different test items across all tests): two fifth-grade language arts items and one seventh-grade mathematics item (with a large effect size). Generally, then, extended time did not lead to different relationships between students’ skill levels and their performance on individual test items.
Unfortunately, like the internal factor structure research reviewed earlier, DIF studies may be more applicable to other accommodations than to extended time. If extended time, on average, increases performance on all test items, the relationship between performance on individual items and performance on the test as a whole (which is examined in most DIF research) may not change while still leading to inappropriate interpretations regarding the skill levels of examinees with disabilities.

To summarize the research on psychometric characteristics, although the factor structures of tests appear unlikely to be affected by extended time, the relationships between test scores and external criterion measures may be affected, and the relationships between individual items and total test scores may also be changed significantly. (Studies disagree on this last point, suggesting that tests may differ.) Additionally, there are few studies that are able to separate out disability status from accommodation status. Finally, the differences in validity evidence may be statistically significant because of the very large sample sizes (e.g., in studies of the SAT and similar tests), but the practical significance of these differences is unknown because the field has no consensus on how large of a difference is viewed as problematic.

Task Comparability

Investigating task comparability is as important as assessing the psychometric characteristics of tests administered under different conditions, but it has been discussed far less. Extended testing time reduces time pressure, which in turn reduces the test’s speededness. If speed is unimportant, as it often is, reducing speededness leads to an improved test (Lu & Sireci, 2007). However, if this is the case, extended time accommodations should be given to all students who can benefit from them rather than making accommodations contingent on a disability diagnosis.

Moreover, speed is not necessarily unimportant. In the past 30 years, a substantial amount of research evidence has accumulated for the utility of rate-based measures of academic skills (often as part of curriculum-based measurement; Marston, 1989). In these measures, frequency counts of academic behaviors (such as reading a word aloud) become the numerator of a ratio where a standardized amount of time is the denominator. For instance, if a child is asked to read a story aloud, the measure of interest will be how many words are read correctly in 1 minute. In the area of oral reading, the subject area that has received the most attention, rate-based measures predict performance on standardized comprehension tests as well as other comprehension tests themselves do and better than norm-referenced measures of word decoding and word identification do (Marston, 1989). More recently, oral reading rate has been found to predict unique variance on high-stakes state reading test scores (Wood, 2006), and its predictive validity has been established across elementary and secondary levels, although the effects are somewhat larger in earlier grades (e.g., Yovanoff, Duesbury, Alonzo, & Tindal, 2005). Recent work on the design of reading assessments that are accessible for all students (Thurlow et al., 2009) has emphasized the need for assessment tools that differentiate components of effective reading, including fluency; these tools will likely aid future research on students’ timing needs.

In addition to predicting performance well, K. R. Johnson and Layng (1992, 1996) have argued that fluency—the ability to perform a skill both accurately and
rapidly—leads to other, more important performance goals that slower (but equally accurate) performance does not. These goals include retention of skills over time, even when performances are separated by long periods without practice; endurance in performing skills for long periods of time without a decline in rate; stability in skill performance despite environmental distractions; and application or generalization of skills to more complex problems of which the learned skill is a component.

Because many testing accommodations are provided in postsecondary contexts, it might be asked whether fluency is worth measuring outside of K–12 settings. This depends on the ecological validity of the testing procedures—that is, the degree to which they mirror tasks performed outside of testing. For instance, when discussing the appropriateness of extended time accommodations on law school exams, Kelman and Lester (1997, p. 180) noted that “lawyers are frequently asked to write a great deal about topics given very short turnaround times” and “some [lawyers] work in settings in which they are expected to produce such a large quantity of work that it is not realistically possible for them to expand work hours to accomplish all that their colleagues do.” Of course, this kind of informal analysis is insufficient for making a decision about whether speed is a desired aspect of examinees’ performance, and it may be the case that empirical investigations would suggest that speed is not necessary in most areas of legal practice. The point is simply that test users must consider whether speed is important rather than assuming that it is not.

Differential Benefits From Accommodations

Much empirical research on testing accommodations has focused on a single issue: Do students with disabilities benefit more from accommodations than non-disabled students do? Many scholars have accepted it as the most important question to ask, and advocates of extended time (e.g., Runyan & Smith, 1991; Shaywitz, 2003) have claimed that this accommodation only affects the performance of students with disabilities. Before reviewing the empirical evidence for this claim, two different evaluative criteria should be distinguished. Some scholars who examine differential benefits have argued that, for extended time to be appropriate, non-disabled individuals (i.e., those without access to extended time) must not be able to benefit from it at all. Zuriff (2000) called this the “maximum potential thesis” because the claim presumes that nondisabled participants are able to work up to their maximum potential during the standard time limits, making any time extension useless. Alternatively, other scholars (e.g., Fuchs & Fuchs, 2001) hold that as long as students with disabilities benefit from extended time more than nondisabled students do, the accommodation is appropriate. Sireci, Scarpati, and Li (2005) called this the “differential boost hypothesis,” because it predicts that extended time will provide a larger boost to scores of students with disabilities than to those of nondisabled students.

Most of the relevant research up to 2004 was discussed in Sireci et al.’s (2005) definitive review. These scholars reviewed several studies that examined the effects of extended time accommodations on students with and without disabilities (mainly learning disabilities), only one of which found that students with disabilities benefited from extended time while nondisabled students did not. Even this one study, conducted by Huesman and Frisbie (2000), actually found a beneficial
effect of extended time for all groups of students, but the effect did not attain statistical significance in one of the groups of nondisabled students. If, then, the maximum potential thesis is used as the criterion, extended time would not be an appropriate testing accommodation because students without disabilities typically benefit from extended time.

However, several of the studies that Sireci et al. (2005) reviewed found that students with disabilities benefited more from extended time than nondisabled students did. For instance, Alster (1997) gave two equivalent forms of a standardized algebra test to 88 college students, half of whom had learning disabilities. Each student took one form of the test at standard time and the other at extended time, and although both groups performed significantly better under extended time conditions, the students with learning disabilities benefited significantly more. Importantly, other studies that Sireci et al. reviewed found that neither group benefited at all or that both groups benefited equally. Sireci et al. concluded that, overall, “extended time tended to improve the performance of all students, although students with disabilities tended to exhibit relatively greater score gains” (p. 457).

Since 2004, several more studies have been conducted, and although they do not lead to any substantial modifications of Sireci et al.’s (2005) conclusions, they provide for a more nuanced understanding of the effects of extended time. Mandinach et al. (2005) found that students with \( n = 264 \) and without \( n = 1,665 \) disabilities (in this study, learning disabilities and ADHD) both benefited from extended time on the SAT but that ability level (operationalized as students’ PSAT scores) predicted degree of benefit, with low-ability students failing to benefit at all from extended time and high-ability students benefiting the most. These investigators also found that although almost 75% of students with disabilities reported, when taking the test under standard time conditions, that they needed extra time to finish, so did almost 50% of the nondisabled students.

In another recent study, Lesaux, Pearson, and Siegel (2006) grouped examinees (adults between the ages of 17 and 60) by word-identification ability, classifying individuals with scores at the 25th percentile or below \( n = 22 \) as having a “reading disability” and also creating groups of average and above average readers \( n = 22 \) and 20, respectively) using the same scores. Lesaux et al. then gave examinees the reading comprehension sections of two parallel forms of the Nelson-Denny Reading Test, one at standard time conditions and the other entirely untimed. Unlike in the Mandinach et al. (2005) study, examinees with “reading disabilities” benefited significantly more from the untimed condition than did either of the other two groups, which were not significantly different from each other in their degree of benefit. The effect sizes (Cohen’s \( d \)) for the extended time (compared to standard time) for the above average, average, and “reading disability” groups were 0.73, 0.96, and 1.68, which suggests that all three groups benefited substantially from the extended time.

Most recently, Lewandowski and colleagues (Lewandowski, Lovett, Parolin, Gordon, & Codding, 2007; Lewandowski, Lovett, & Rogers, 2008) conducted two studies with the intention of eliminating ceiling effects that may have kept nondisabled students from benefiting from extended time in prior research. Lewandowski et al. (2007) administered a highly speeded mathematics computation test to 54 middle school students, half of whom had ADHD. After working on the test for 12 minutes, students were told to circle the problem that they had just completed and
then were given an extra 6 minutes to work (i.e., 50% extended time). In this study, the nondisabled students benefited more from the extended time than the students with ADHD did, answering more items correctly during the extended time period. The effect sizes ($d$) for the nondisabled and ADHD groups were 1.16 and 0.98, respectively. In a follow-up study, Lewandowski et al. (2008) administered the Nelson-Denny Reading Test comprehension subtest to 64 high school students, half of whom had reading disabilities. To help avert ceiling effects, the standard time length for the Nelson-Denny Reading Test was reduced from 20 to 13 minutes, making the (50%) extended time amount 6.5 minutes. Again, the nondisabled students benefited significantly more from the extended time ($d = 1.94$) than did the students with disabilities ($d = 1.55$).

To summarize the differential benefits literature, the effects of extended time are certainly not specific to examinees with disabilities—that is, nondisabled examinees usually benefit from extended testing time. The benefits are usually larger for examinees with disabilities, but this is not a consistent finding and it depends on examinees’ ability levels and test characteristics (e.g., speededness). Thus, the evidence fails to meet the standard set by the maximum potential thesis, which forbids nondisabled students from benefiting, but there is at least some evidence supporting the differential boost hypothesis, for tests that are not highly speeded.

Examinees’ Ability to Adapt to Standard Conditions

The ability of examinees with disabilities to adapt to standard timing conditions has not been directly addressed in the empirical literature, in that no researchers have developed interventions for the purpose of reducing students’ need for extended time testing accommodations. Advocates for increased availability of extended time argue that it is necessary; Shaywitz (2003) even makes an analogy between extended time for students with dyslexia and insulin for individuals with diabetes. However, there is empirical evidence that bears on the issue indirectly, and it challenges the presumption of extended time’s necessity. Examinees with disabilities are granted extended time accommodations when they have certain relevant personal attributes (e.g., slow test-taking speed), and to the degree that those attributes can be changed, their need for extended time accommodations can be changed.

One set of personal attributes that leads to accommodations is suggested by a pioneering qualitative study by Rickey (2005), who made detailed observations at three middle schools and conducted intensive interviews with various stakeholders in the testing accommodations decision-making process: teachers, parents, students, and school administrators (total $n = 28$). These individuals were asked about the place of high-stakes tests in the culture of the school, the reasons why accommodations were given, and the effects of those accommodations. At each of the three schools, Rickey found that the primary reason accommodations were provided was to “reduce the stress and frustration experienced by the student during testing” (p. 124). At one of the three schools, this was the only reason given. Parents spoke about how the experience of testing could lower their children’s self-esteem and how accommodations could protect it; at one school, a teacher spoke about how anxious the students might get if they were not given accommodations. If frustration with testing is a primary reason for accommodations, reducing frustration would, then, help students adapt to standardized test administration conditions.
Admittedly, Rickey’s (2005) sample size was small, and her findings may not generalize beyond the schools in her study. However, more generally, research has shown that frustration, stress, and test anxiety are frequent experiences reported by students with high-incidence disabilities (e.g., Lackaye, Margalit, Ziv, & Ziman, 2006; Wachelka & Katz, 1999) at many levels of education as well as sources of extended time accommodations requests among adults taking standardized tests (Zuriff, 1999). Further supporting the importance of affective factors in accommodation use, other researchers have given students with disabilities tests with and without accommodations, then asking the students to reflect on their experiences. S. N. Elliott and Marquart (2004) gave two alternate forms of a mathematics test to 97 eighth-grade students, one at standard time and one at extended time; 23 of the students had high-incidence disabilities. Of the students with disabilities, 78% reported feeling more relaxed and 52% reported feeling more motivated during the double time administration. In a similar study, Lang et al. (2005) gave students with disabilities (152 fourth graders and 142 eighth graders) math and reading tests under standard and accommodated conditions (here, the accommodations included extended time and other accommodations) and found that 46% of the students felt more comfortable when they took the tests with accommodations, whereas fewer than 10% felt more comfortable without the accommodations; and 62% of the students found the tests easier with accommodations, whereas fewer than 10% felt the opposite to be the case.

It might be argued that students’ test anxiety and discomfort are caused by cognitive and academic skill deficits, and these deficits are the ultimate reason for accommodations. However, the available evidence suggests otherwise. S. N. Elliott and Marquart (2004) and Lang et al. (2005) generally found the same effects of accommodations on motivation and comfort in students without disabilities. Lovett (2007) also found evidence for a unique, direct influence of test anxiety. Lovett assessed 127 nondisabled college students’ perceptions of their test timing needs using a questionnaire with items such as “I could do better on my exams if I had additional time.” He also administered a test anxiety questionnaire and a reading fluency test. Students’ test anxiety scores predicted their perceptions of timing needs even after controlling for their reading fluency scores, suggesting that their test anxiety was not itself solely due to skill deficits. Reading fluency predicted perceptions at \( r = .51 \), but adding students’ test anxiety raised the predictive power to a multiple correlation of \( R = .63 \).

If test anxiety and other emotional issues such as frustration are leading many students, parents, and school personnel to consider extended time accommodations, it is unclear why counseling and other therapeutic interventions are not tried first to see if the examinee can adapt to standard testing conditions. There are validated interventions for test anxiety (e.g., Wachelka & Katz, 1999), and stress and frustration regarding high-stakes testing can often be effectively combated using test-preparation strategies that are increasingly provided to schools, teachers, and to examinees directly at upper levels of education.

Of course, whatever emotional issues individuals who receive extended time are facing, many will certainly have deficits in cognitive and academic skills that are related to their accommodations (specifically reading fluency deficits; see Burns, 1998; Lindstrom, 2007). However, one should not assume that these are immutable features either; in fact, for K–12 settings, a large literature on fluency interventions
exists that is very relevant to extended time. In their recent review of this literature, Fletcher, Lyon, Fuchs, and Barnes (2007) reported on many intervention programs with moderate to large effect sizes on fluency \((d_s = 0.50–0.68)\), and even more recently, Wexler, Vaughn, Edmonds, and Reutebuch (2008) reviewed programs designed specifically for high school students struggling with reading (although effect sizes were more variable; \(d_s = 0.23–1.02)\). Unfortunately, although students receiving extended time accommodations are typically also receiving other special education services, those services may not include training in reading fluency.

To summarize the literature on adaptation to standard testing conditions, the empirical research consistently suggests that (a) much of the interest in extended time accommodations may be due to emotional factors such as test anxiety, which can be addressed more directly, and (b) examinees who have lower levels of reading fluency may benefit from a number of validated intervention programs. Admittedly, test developers and administrators are not generally in an appropriate position to suggest interventions and strategies such as these. Moreover, the implementation of interventions can be costly and time consuming. However, members of multidisciplinary special education evaluation teams in K–12 schools (the teams who make testing accommodations decisions) can include interventions as part of the services to students with disabilities who receive extended time accommodations, and disability services office staff in postsecondary institutions can refer students who receive extended time to appropriate resources. Overall, the literature is optimistic, suggesting that at least some examinees can adapt to standard timing conditions.

**Technical Adequacy of Diagnostic Procedures**

Most discussions of testing accommodations for students with disabilities have not addressed the issues involved in determining whether an individual student has a disability. Instead, it appears to be taken for granted that when researchers discuss, for instance, the effects of extended time accommodations on students with learning disabilities, the students with the learning disability label actually have the same type of disabling condition that separates them from their peers in a fundamental way. When considering certain low-incidence disabilities, this may be a reasonable assumption. However, the high-incidence disabilities that comprise many special education classifications and lead to the greatest number of extended time accommodations are different. A comprehensive review of the literature on disability diagnosis is beyond the scope of this article, but the following section briefly reviews selected issues in the diagnosis of learning disabilities and ADHD, two conditions often associated with requests for extended time as the sole accommodation (Ranseen & Parks, 2005). The diagnostic issues associated with these disabilities extend to high-incidence disabilities generally.

**Learning Disabilities**

Learning disabilities are defined by their specificity; that is, when a student has trouble learning to read, write, or do arithmetic but has no general deficit in the ability to learn (i.e., low intelligence, as seen in intellectual disability), a learning disability is suspected (Kavale & Forness, 2000). Therefore, the most popular diagnostic criterion for learning disabilities over the past few decades has been a “severe discrepancy” (usually about 1.5 standard deviation units) between a
student’s intelligence (IQ) and achievement in a subject area (Kavale, 2002). The IQ-achievement discrepancy has long been required in many states to make a learning disability diagnosis, although, as Reschly and Hosp (2004) noted, the size and nature of the required discrepancy is much larger in some locales than in others, meaning that whether a student has a learning disability depends in part on where he or she lives. This contributes to the variable rates of learning disabilities across different states; in 2001, for instance, fewer than 3% of students in Kentucky had a learning disability diagnosis, whereas more than 9% of students in Rhode Island did (Reschly & Hosp, 2004).

Even if the same IQ-achievement discrepancy were used everywhere, the reliability and validity of the diagnosis would be problematic. Among the many statistical problems with difference scores, they generally have poor reliability, especially when the two tests used to calculate the difference have a high correlation between them, as intelligence and achievement tests do (Sternberg & Grigorenko, 2002). In addition, the discrepancy approach fails to identify students who require special help early enough (Fletcher et al., 2007). Finally, many students who are performing at grade level in the classroom are identified because their high IQs yield a discrepancy, even though they do not appear to have any unique educational needs (Lovett & Lewandowski, 2006).

These problems have led many states and school districts to adopt a very different approach, the response-to-intervention (RTI) model (e.g., Gresham, 2002). Under the RTI model, all students in a school are provided research-supported instruction in basic academic skills, and their skill development is closely monitored. Students whose skills are not developing adequately are given increasingly intensive academic interventions, and if even the most intensive interventions fail to work, a learning disability is diagnosed. Although the RTI model addresses several of the problems associated with IQ-achievement discrepancies, it has its own limitations. Different indexes of the adequacy of students’ responsiveness to interventions have low rates of agreement (Barth et al., 2008), and RTI systems differ greatly across states (Berkeley, Bender, Peaster, & Saunders, 2009), again making diagnosis a function of where a student lives.

As basic to diagnostic reliability as the choice of a diagnostic algorithm is the need to follow whatever algorithm is chosen. However, MacMillan, Gresham, and Bocian (1998) found that schools often ignored their own diagnostic decision rules when classifying children as having a learning disability. At the postsecondary level, Sparks and Lovett (2009) found that fewer than half of their university students with learning disability diagnoses (n = 357) met any of the standard criteria for diagnosis based on test scores, and Harrison, Nichols, and Laroche (2008) found that a quarter of their students applying for learning disability accommodations at Canadian colleges (n = 247) submitted no documentation at all, whereas others submitted only a physician’s note without test scores or other more detailed information.

Many of these practices may be reasonable within an educational context. The decision rule for classifying a student as having any problem that exists on a continuum (as learning disabilities do; Fletcher et al., 2007) is at least somewhat arbitrary, and so states and school districts might use financial and logistical factors to determine it. Similarly, depending on which remedial services are provided to students with learning disabilities, following official decision rules and requiring
strict documentation might not be very important. However, higher diagnostic reliability and validity are needed if a diagnostic label is taken as evidence supporting the provision of extended time accommodations, as it often is.

**ADHD**

Like learning disabilities, ADHD exists on a continuum (Haslam et al., 2006), and so some of the diagnostic issues revolve around the quantity and intensity of symptoms of ADHD that someone should have before a diagnosis is made. The current edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV-TR; American Psychiatric Association, 2000) lists 18 symptoms of the disorder, grouped into symptoms related to inattention (e.g., making careless mistakes), impulsiveness (e.g., impatience), and overactivity (e.g., talking excessively). Perhaps unsurprisingly, many typical children and adults will endorse individual symptoms. The DSM-IV-TR also requires that the individual has experienced “clinically significant impairment” in real-world settings, but many diagnosticians do not assess this in a standardized way, and far fewer diagnoses are made when the impairment criterion is applied (Gordon et al., 2006).

School psychologists, clinical psychologists, psychiatrists, and pediatricians, among others, can all diagnose ADHD, and each specialty uses somewhat different assessment tools and criteria to make a judgment about the disorder’s presence (Handler & DuPaul, 2005). Even within a single specialty, school psychologists have been found to vary greatly in their assessment methods; Koonce (2007) found that 23% of 246 school psychologists surveyed said that an intelligence test was the only standardized test needed to make an ADHD diagnosis, whereas 19% said that they would never or only occasionally include an intelligence test when assessing for the presence of ADHD. Like learning disabilities, then, the probability of being diagnosed with ADHD is highly dependent on the diagnostician.

To summarize the diagnostic literature, the reliability and validity of diagnoses of learning disabilities and ADHD are in serious question. As is typical of high-incidence disabilities, these problems exist as segments of continua, but consistent cut points are not used in diagnosis. Moreover, the assessment measures used to assess these disabilities are not consistent across diagnosticians, with some insisting on using measures that others hold to be of no value, and with clinical judgment taking precedence over test score data. Debate over how to properly diagnose these conditions means that although certain individual diagnostic decisions may be obviously correct (in that they are clear cases that all diagnosticians would agree on) the diagnostic categories are too broad and the individuals within them too heterogeneous for the diagnoses to be useful in making accommodation decisions.

**Technical Adequacy of Accommodation Decision Procedures**

It is generally held that testing accommodations are inappropriate if the decisions regarding them cannot be defended as reliable and valid; certainly, if the decision to provide an accommodation is made unsystematically, it would seem preferable to allow all examinees access to the accommodation. To increase the reliability and validity of accommodation decisions, several decision models have been proposed, and most of these can be applied to extended time accommodations. Still, one should begin by noting the impact of the literature already reviewed on this question. If the diagnosis of high-incidence disabilities has dubious reliability
and validity, then looking at an examinee’s disability status alone would be an unreliable and invalid decision procedure. Additionally, given that school personnel may base accommodation decisions around the desire to increase the comfort of students during tests (Rickey, 2005), the opinions of teachers and school administrators, even when formalized as “special education team” or “individualized education plan team” decisions, should be supplemented with objective data.

Standardized decision procedures, then, provide the most hope for reliability and validity. One such procedure is the Dynamic Assessment of Test Accommodations (DATA; Fuchs, Fuchs, Eaton, & Hamlett, 2003), which assesses whether a student benefits significantly more from an accommodation than typical nondisabled students do (as in the differential boost hypothesis). The DATA contains brief forms of reading and mathematics tests that can be given under various conditions; to determine whether extended time is appropriate, two forms of the same type of test are given, one at standard time and one at extended time. A student’s benefit from the extended time is calculated as a gain (difference) score, and that score is compared to normative data on the gain scores obtained by nondisabled students. If the target student’s gain score is substantially greater than that of the average nondisabled student (at or above the 84th percentile of the nondisabled gain score distribution), extended time is recommended.

The tests used in the DATA have adequate reliability and predict scores on longer, standardized measures such as the Stanford Achievement Test (Fuchs et al., 2003). In addition, two studies have found that the DATA leads to notably different accommodation decisions than unaided teacher judgments do, with the DATA recommending fewer accommodations and more beneficial ones (in terms of increasing students’ test scores). Fuchs, Fuchs, Eaton, Hamlett, and Karns (2000) used the DATA to make decisions about accommodations on mathematics tests for 192 fourth-grade students with learning disabilities. In three domains of skills (computation, concepts/applications, and problem solving) covered by the mathematics test, teachers recommended accommodations for 65%, 93%, and 93% of the students, respectively, whereas the DATA recommended accommodations for 10%, 32%, and 43% of the students in these domains. Fuchs, Fuchs, Eaton, Hamlett, Binkley, and Crouch (2000) compared teacher and DATA judgments of accommodations on a reading test using the same sample of students; teachers recommended accommodations for 72% of the students on this test, whereas the DATA only recommended accommodations for 41% of the students. In both of these studies, the DATA was also superior at predicting students’ gains from accommodations on large-scale assessments.

These findings are impressive, but the DATA has limitations as well. Among these, it takes differential benefit as the sole consideration in determining the appropriateness of extended time, it recommends extended time even when up to 16% of nondisabled students would benefit more from the accommodation than the target student would, and it does not consider the absolute amount gained by nondisabled students—if nondisabled students would benefit substantially from extended time, it is unclear whether the accommodation should be limited to students with disabilities just because they would show even greater benefits (cf. Zuriff, 2000).

Ofiesh, Hughes, and Scott (2004) proposed another approach to decisions; they focused on postsecondary students with learning disabilities. These scholars began by noting that a disability diagnosis is insufficient to make decisions about
extended time and recommended that disability services providers use a diagnosis as a starting point to begin asking other questions such as the following: What academic areas, if any, are affected by the disability? What is the impact of the disability on the completion of timed tests? Are the tests that the individual seeks accommodations for designed to measure speed of response? Only if the disability is found to affect the completion of timed tests in academic areas in which the student is taking classes, but in which the course instructors are uninterested in response speed, would extended time be recommended.

Ofiesh et al.’s (2004) model spurs the asking of key questions and provides a sound logical argument for or against extended time accommodations in individual cases. However, their recommended operationalization of the disability’s impact on test taking involves the use of norm-referenced tests of cognitive and academic skills. For instance, if the student’s score on a standardized processing speed task is very low (e.g., below the 10th percentile), this might be interpreted as affecting the student’s ability to complete timed tests. Unfortunately, a study by Ofiesh, Mather, and Russell (2005) suggests that these tests may not be appropriately interpreted with regard to extended time accommodations. These investigators used a battery of norm-referenced tests to predict whether each of 43 university students with learning disabilities benefited from extended time accommodations and found that only 2 of the 10 speeded scores derived from the battery predicted which students benefited, and neither of the 2 scores (Reading Fluency and Academic Fluency indexes from the Woodcock-Johnson Battery) explained much variability in who benefited (rs = .38 and .36, respectively).

Other, somewhat less standardized decision models have been proposed by Hollenbeck (2002) and Braden and Joyce (2008), who each provided flowcharts that they recommended test users follow before making an accommodations decision. These flowcharts each address many of the issues that Ofiesh et al. (2004) raised: whether the accommodation will affect the construct that the test is designed to measure, whether the student’s capabilities actually make standard testing conditions inappropriate, and so forth. Although no research has been done on how use of these flowcharts affects decision outcomes, merely requiring decision makers to consider the relevant questions might make decisions more thoughtful.

Newer decision models and other, less formal guidance on making testing accommodation decisions continues to be made available. For instance, the National Center on Educational Outcomes at the University of Minnesota publishes new materials each year on testing accommodations—original descriptive research on accommodation use as well as reports offering advice to K–12 school district personnel on how to use accommodations effectively (e.g., Christensen, Thurlow, & Wang, 2009). In addition, testing companies are partnering with government agencies to develop examinations appropriate for students with disabilities. For instance, the Designing Accessible Reading Assessments project (King, 2008) has already produced several publicly available reports, and its research work continues.

To summarize the decision-making literature, there are several different standardized decision procedures, and each has the potential to improve decision quality substantially. However, there is very little research on how these procedures are implemented in practice. Given that the standardized decision procedures do not always agree with each other or with the opinions of school personnel, and given
that no decision procedure is used consistently across schools, school districts, colleges, or states, the reliability of decisions appears inadequate. In addition, the available research on the DATA as well as Ofiesh et al.'s (2004) model suggests that both can lead to inappropriate accommodations, either because of poor predictive power or narrow considerations in decision making. Very recent accommodation research projects hold out the promise that in the next several years more will be known about how to develop assessment techniques that are appropriate for students with disabilities. However, in the meantime, far more work is needed examining how accommodation decisions are currently made—including the degree to which standardized decision models are used and the degree to which they help improve decision quality.

**Conclusions**

**Limitations of Research Base**

As other scholars who have reviewed the testing accommodations literature (e.g., Pitoniak & Royer, 2001; Sireci et al., 2005) have noted, studies in this area tend to have a variety of expected methodological limitations, including small samples of students with disabilities and heterogeneous disability groups. Additional limitations found in this review include the inability of many studies to distinguish between accommodation effects and disability effects (when only students with disabilities are given the test with accommodations) and the tendency for extended time to be provided, especially in K–12 settings, as part of a more comprehensive testing accommodations package, making it difficult to separate the effects of different accommodations. The wide variety of research designs across studies also limits the degree to which studies can be directly compared and integrated.

In addition to these limitations, there are a number of issues that have been inadequately addressed and that relate to extended time accommodations. First, although the effects of extended time on test performance depend on the speededness of the test, few researchers attempt to calculate the speededness of the measures used in their studies, nor have they tried to design measures with speededness equivalent to that of actual high-stakes tests. Second, test-taking speed depends in part on motivation and effort, just as speed in all tasks depends in part on these factors, but researchers rarely attend to examinee effort when studying extended time accommodations. Finally, although mental speed is a component of most models and standardized assessments of cognitive ability, and experimental psychologists have studied response speed for more than a century, there is little work on the utility of speed in educational settings, outside the recent research on academic fluency. Because the validity of speeded tests—and thus the appropriateness of extended time accommodations—depends on this utility, more empirical investigation is warranted.

**Tentative Answers About Extended Time**

Despite the limitations of this research base and the areas yet to be examined, there is sufficient research to provide tentative answers to the five questions posed in this review when considered with the caveats just discussed. First, extended time appears to have small effects on the psychometric characteristics of test scores—predictive relationships with external variables may be somewhat weaker, and
some items may show different item-test relationships. There is also evidence that in certain settings (e.g., assessment of reading skills) speed may be important to evaluate, and so test users must always consider the importance of speed before choosing a test and designing administration conditions.

Second, nondisabled students almost always benefit from extended time accommodations on timed tests, although students with disabilities typically show somewhat greater benefits. On highly speeded tests, nondisabled students may actually benefit more from extended time. This suggests that test users should consider making extended time available to all examinees if speed is of no interest to the users.

Third, the potential exists for some students with high-incidence disabilities to adapt to standard timing conditions. Interventions for reading fluency as well as other issues that can lead to extended time accommodations (e.g., test anxiety) are available, and although they will not benefit all examinees equally or address all problems that lead to extended time, they should be considered when providing sets of services for students with disabilities in the hopes of eliminating the need for extended time for at least some examinees.

Fourth, the reliability and validity of diagnoses of high-incidence disabilities is insufficient to support accommodation decisions made by referencing diagnostic status. Learning disabilities and ADHD were examined in this review as examples of disability conditions that frequently lead to extended time accommodations, and experts disagree about how these conditions should be assessed. At times, the diagnostic models vary widely, and examinees with these diagnoses frequently do not meet official criteria for the disorders.

Finally, the reliability and validity of accommodation decisions are currently inadequate. Most decisions are made without any standardized procedures, instead relying on teachers’ judgments or students’ disability labels. Moreover, even the most standardized procedures (e.g., the DATA) have limitations that have yet to be addressed. Much work on the topic is currently underway both to expose decision makers to standardized procedures and to improve those procedures, but at present some of the more common decision making processes do not appear to be defensible.

Recommendations

In highlighting the hurdles that an extended time accommodation must overcome to be appropriate in a specific situation, this review suggests the importance of developing tests such that few extended time accommodations ever need to be considered. The push for universally designed assessments (tests made so that almost all students can take them in the same format; Thompson, Johnstone, Anderson, & Miller, 2005) may be helpful in regard to extended time issues. Using principles of universal design, test developers would carefully consider the relevance of speed to the test, and test administration would be designed so that either all students could take as long as they wished or so that (if speed was deemed important) time limits would be empirically determined and inflexible.

To conclude, the ultimate question concerning what test administrators and test users should do regarding extended time, requires far more than a decision about whether to provide an accommodation. It may involve, for example, the development and refinement of a comprehensive extended time accommodation decision procedure tailored to the particular test and setting, the redesign of a test to ensure that the time limit is appropriate for all examinees, or even in certain settings the
referral of examinees with disabilities to various intervention programs to reduce test anxiety and improve fluency skills. In short, there are a variety of potential improvements to assessment policy that are inspired by Phillips’s (1994) questions and supported by extant research evidence. The challenge now is to implement these improvements, simultaneously reducing the need for extended time accommodations and enhancing the utility of assessment programs.

Notes

For helpful comments on earlier drafts of this article, I thank Tanya Eckert, Matthew Eggert, Michael Gordon, Linda Johnson, Alexander H. Jordan, Lawrence Lewandowski, Joyce Smith, Nicole Talge, and Michael Vance.

1 Comparability concerns have led some scholars (e.g., Hollenbeck, 2002) to make a distinction between an accommodation and a modification; an accommodation comes with evidence that test scores obtained under the alteration are comparable to scores obtained under standard conditions, whereas a modification does not. Although this recognition of comparability concerns is admirable, the term accommodation is used throughout this article because, in practice, it is often difficult to say whether an alteration to test administration conditions yields comparable scores.

2 Phillips’s (1994) five questions stated the five issues in a slightly different way, devoting two questions to issues of comparability and combining the issues of diagnostic decision making and accommodation decision making into a single question.

3 Interpretation of confirmatory factor analysis studies and, to a lesser degree, differential item functioning studies such as those presented in the next section, is controversial—in part because the critical values for the statistical tests and various indexes of model fit are debated and in part because these techniques generally require larger sample sizes and additional statistical assumptions, which may not always be met. Although a full treatment of these issues is beyond the scope of this article, the confirmatory factor analysis and differential item functioning studies reviewed in this article were examined to ensure that they were typical in their use of these procedures. Readers can find fuller treatments of the relevant issues in texts by Brown (2006) and Osterlind and Everson (2009).

References


Lovett


Lovett


Author

BENJAMIN J. LOVETT, Ph.D., is an assistant professor of psychology at Elmira College, Department of Psychology, Elmira College, 1 Park Place, Elmira, NY 14901, e-mail: blovett@elmira.edu. His research focuses on the conceptual and psychometric foundations of psychoeducational assessment and psychiatric diagnosis. He earned his Ph.D. in school psychology from Syracuse University.