Validity Evidence in Accommodations for English Language Learners and Students with Disabilities

Wayne Camara

The College Board

Abstract

The five papers in this special issue of the *Journal of Applied Testing Technology* address fundamental issues of validity when tests are modified or accommodations are provided to English Language Learners (ELL) or students with disabilities. Three papers employed differential item functioning (DIF) and factor analysis and found the underlying constructs measured by tests do not change among these groups of students. Despite this strong finding, consistent and large score differences are present across groups. Such consistent and large score differentials among these groups on cognitive ability tests would be ideally contrasted with findings from alternative measures (e.g., portfolio's, performance assessments, and teachers' ratings). Two papers examine current methods used to identify and classify both ELL and students with disabilities, while other papers examine the performance of students with specific disabilities (e.g., deaf, mental retardation). The impact of modifications and accommodations on score comparability is discussed in relation to professional standards and current validity theory.

Validity Evidence in Accommodations for English Language Learners and Students with Disabilities

Introduction

Accommodations[1] are designed to minimize the impact of test taker attributes that are irrelevant to the construct. A standardized test that has been designed for 8th graders may be inappropriate for students with certain disabilities or students who are tested in their non-native language (AERA, APA, & NCME, 1999). A second purpose for such accommodations has been to make assessments more accessible to large numbers of students who have traditionally been excluded from accountability testing because of disabilities or language. The inclusion of these students in large-scale accountability testing also is relevant to the validity of inferences made from assessment results (Koretz & Hamilton, 2006).

The *Standards for Educational and Psychological Testing* (AERA et al., 1999) note that validity evidence pertains to the intended interpretation and uses of the test score. Threats to the internal validity of such interpretations stem from construct-irrelevant variance or construct under-representation. Messick (1989) noted that tests are not only "imprecise or fallible by virtue of random errors of measurement but also inevitably imperfect as exemplars of the construct they are purported to assess" (pp. 34). They either leave out something that should be included in the construct or measure something that should be excluded from the construct. For example, until 2005, writing was excluded from both major undergraduate admissions tests despite its centrality to college

[1] For purposes of this paper, accommodations are defined as changes made in the content, format, or administration procedure that makes a test more accessible for students with disabilities or limited language proficiency and does not change the intended construct. Modifications are defined as changes that will likely impact the construct. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) consider these two terms interchangeably.

success (Milewski, Johnsen, Glazer, & Kubota, 2005). Subsequent studies have demonstrated that writing is the single best predictor of freshmen grades (Kobrin, Patterson, Shaw, Mattern & Barbuti, 2008) in college and its exclusion is one example of such construct under-representation. Such exclusions are to be expected between what assessments can plausibly measure and the construct domain, and Kane (2006) notes that under-representation can range from a genuine threat to validity to exceptions for an individual or group.

If an assessment fails to measure a major ingredient within a construct, under-representation may occur to some degree. Much of the support for performance assessments emerged from a principled argument that constructs were often poorly represented when measured exclusively with objective items. Tests of writing that rely exclusively on multiple-choice items may result in strong relationships with criterion measures, but the lack of on-demand writing tasks can pose a threat to the construct domain. For English Language Learners (ELL), construct under-representation is of concern when students are not tested in their dominant language and their test score does not capture their true knowledge or ability in a subject domain such as algebra or geography. Similarly, it is difficult to argue that a reading comprehension test administered orally to a blind student or a listening comprehension test administered in text to a hearing impaired student does not suffer from some degree of construct-under-representation.

Construct-irrelevant variance is more often cited as a threat to the validity of score interpretations with ELL and students with disabilities (SWD) than under-representation. Limited language proficiency of students not tested in their dominant language can

interfere with demonstrated knowledge or skills. Similarly, SWD will often have impairments which can impact their performance on educational tests. Modifications or accommodations have been increasingly used to increase participation in large scale assessments, but several important issues emerge when evaluating the validity of score interpretations in such situations.

Separate and equal?

The use of principles of universal design in test development has greatly expanded since the advent of the Americans with Disabilities Act of 1990 and the mandates for inclusion of all students in No Child Left Behind (2002). Here the attempt is also focused on increasing the validity of inferences drawn from the test scores by reducing the impact of construct-irrelevant variance. Universal design proposes that test developers use the least restrictive environment or specialized requirements when designing assessments. Such practices may be ideal but not always feasible or attainable in reality. Ketterlin-Geller (2008) acknowledges this and argues that the goal should not be comparable forms of assessments that are appropriate for all students, but comparable interpretations. The assessment system maintains the integrity of the construct through flexibility in the format, presentation, delivery and administration.

Ultimately, the determination of whether or not different assessment forms or assessments that vary in format, presentation, delivery and administration are comparable is an empirical question as much as it is a theoretical argument. The same issues also arise when evaluating the validity of inferences based on assessment results with accommodations or modifications.

The first question concerns whether the impairment caused by the special needs is relevant or not relevant to the construct? For example, visual impairments would interfere with many items on geometry or statistics tests that ask students to interpret complex graphs and figures. Changes to such items or principles of universal design that measure the construct with different types of items would be justified in such examples because the impairment is not relevant to the construct. However, cognitive deficits that interfere with a student's performance on the same geometry or statistics test would generally be relevant to the cognitive constructs (Koretz & Hamilton, 2006). Language can also be relevant or irrelevant to the construct. When language proficiency is not a part of the construct, the linguistic or reading demands of the assessment should be kept to the minimum level necessary (AERA et. al., 1999). However, if the assessment is intended to measure oral comprehension in English a test administered in another language, dual languages, or with other modifications for ELL may actually introduce construct irrelevant variance. If the impairment caused by special needs or the differential language proficiency of the learner is irrelevant then attempts should be made to find accommodations that can minimize the impact on performance.

Once such accommodations or modifications have been investigated, evidence relating to the validity of inferences resulting from these test scores must be gathered. Depending on the purpose of the assessment different forms of evidence may be most persuasive. Koretz and Hamilton (2006) note that increased participation is clearly one of the major goals of NCLB and that when aggregate results are interpreted as reflecting all students (or 98% of all students) in a grade the systematic underrepresentation of ELL or students with disabilities is a threat to validity. They note the increased participation of

students in the National Assessment of Educational Progress (NAEP) when accommodations were provided. However, they also note that such inclusion requires evidence to support the validity of inferences for these populations where impairments or language differences are irrelevant to the construct.

A second question is whether the accommodation has introduced construct irrelevant variance. When attempting to minimize the impact of an impairment or differential language proficiency, does the accommodation introduce construct-irrelevant variance? For example, extended time has been a frequent accommodation for students with learning disabilities, yet there is conflicting evidence concerning whether results are comparable to those administered under standardized conditions (Cahalan, Mandinach & Camara, 2002; Sireci, Scarpati & Li, 2005). In college admissions testing, the predictive validity studies have been available that examine college performance of students with disabilities testing with and without accommodations and comparing findings to students without disabilities testing with and without accommodations. However, the absence of criterion measures in K-12 large scale testing has resulted in more emphasis on the internal psychometric properties of tests administered to different groups (Koretz & Hamilton, 2006).

The papers in this issue extend the literature in terms of examining these questions of validity and comparability of assessments with three primary focal groups – students with disabilities, English Language Learners, and English Language Learners with disabilities. They examine the psychometric characteristics of items and tests administered under various conditions with these groups in order to help testing

professionals gain greater insight into issues of validity as they relate to accommodations and special populations.

Classification

Limited English Proficient[2] students were estimated to comprise 9.6% of K-12 students in 2001, with nearly 80% of students speaking Spanish (Kindler, 2002). Students with disabilities comprised 13.8% of all students in pre-K programs through 12th grade in 2005-06 (U.S. Department of Education, 2007). Abedi (2009, this issue) cites estimates for the number of ELL students with disabilities (ELLWD) in K-12 at over 350, 000, or 9% of all ELL students and 8% of all children in special education.

Proper and consistent classification for students in these three groups continues to be a major concern that impacts the validity of research findings and estimates of the impact on aggregate test results. Abedi (2009, this issue) notes that less than 10% of the variance in ELL classification is explained by students' English proficiency. Kindler (2002) noted districts are responsible for identifying ELLs and that the most frequent methods of identifying are a home language survey, parental reporting, teacher observations, student records, teacher interviews and referrals. Forty-six states provided accommodations for ELL students on state assessments in 2000-01, but only 28 states reported data on accommodations for ELLs separate from those provided students with disabilities (Rivera, Collum, Schafer, & Sia, 2006).

Classification of students with disabilities is also inconsistent across teachers and schools according to the National Research Council (1997). There are inconsistencies in the processes used to identify students and criteria employed in classification. While

[2] For purposes of this paper, the terms limited English proficient (LEP) and English language learners are used interchangeably.

various criteria have been established to aid in diagnosis and classification, there is substantial heterogeneity among students classified in various special needs groups. The variety of disabilities, high prevalence of students with multiple disabilities and distinctions among the severity levels (and resulting impact on learning) produce many extremely small samples of special needs students that make meaningful research difficult to conduct (Koretz & Hamilton, 2006; Vacc & Tippins, 2002). Legal mandates and professional practice emphasize the need to conduct individualized assessments and tailor accommodations to the needs of the individual, which may be sound advice for instructional purposes, but also complicates research in the field.

ELL with Disabilities: Classification, Assessment and Accommodation Issues

Abedi (2009, this issue) begins his paper with a discussion of classification issues for ELLs with disabilities (ELLWD). He notes that misclassification may occur when the disability is hidden by an extreme lack of English proficiency or when the lack of language proficiency is mistaken for a disability. He argues that ELLWD students are more frequently misclassified than students in either single category and notes the need to develop and validate a classification system.

Abedi's paper is one of the few studies that examine the differential performance of three focal groups: (1) ELL students, (2) students with disabilities, and (3) ELLWD students. Group differences are transformed to a Disparity Index (DI) by subtracting the mean of the reference group from the mean of the focal group and dividing the difference by the mean of the focal group. This value is then multiplied by 100 to convert it to a percentage that distinguishes the performance disparity among the groups. A negative value results when the performance by the focal group is lower that of the referent group.

Two sets of data are used in his study. In the first site, the Stanford Achievement Test, version 9 (SAT9) is used with three focal groups. The referent group is students with no disabilities who are not ELL. Data are reported for SAT9 Math and Reading test scores for grades 3 and 8 prior to the implementation of NCLB accountability requirements. Data from a second site employed a state criterion-referenced test of math and reading taken by students in grades 5 and 8, post NCLB. Again, the DI is computed for the same three focal groups and referent group.

Results of the DI are somewhat difficult to interpret. For example, on the grade 3 SAT9 reading testing there was a DI of -53 between ELL students and the referent group and a DI of -208 between the ELLWD group and referent group. Abedi explains that the ELL students underperformed the referent group by 53.4% whereas the ELLWD students underperformed the referent group by over 200%. Computing effect sizes is an alternate method of examining the difference between group means. In this example, the effect size for the ELL group would have been 0.63 and the effect size for the ELLWD would have been 1.39 (Cohen, 1988). An effect size of 0.63 is moderate and an effect size greater than 0.80 is large. An effect size of 0.0 indicates that the mean of the focal group is at the 50th percentile of the referent group and vica versa. An effect size of 0.80 indicates that the mean of the one group is approximately at the 80th percentile of the second group. Finally, an effect size of 1.7 indicates that the mean of one group is at the 95th percentile of the second group. The effect size similarly provides an index of the percent that scores in the two groups overlap and is commonly used to interpret group differences in the social sciences.

Results across both sites and all grades were very consistent in math and reading. The largest disparities (and effect sizes) were found for the ELLWD students. Disparities between the ELL students and students with disabilities were about half as large. Generally, the gaps were slightly larger for the students with disabilities than the ELLs and effect size for all comparisons at both sites were generally large. The exceptions were found primarily with moderate effect sizes among 3rd graders at site 1 using the SAT9 for the ELL only and students with disabilities only groups. DIs and effect sizes were noticeably larger at site 2 which employed criterion referenced tests post NCLB, but direct comparisons between sites should not be made because of likely differences in the populations and psychometric properties of the two assessments.

This study does demonstrate a significant gap between ELLWDs in relation to other groups. The reliability for this group was also consistently lower across both tests and subjects when compared to reliability with other focal groups and the referent group. In addition, the correlations between reading and math scores are consistently lower for the ELLWD group (.38-.52). Factor loadings were also generally lower for the ELLWD group which casts additional doubt on the validity and reliability of this classification and the psychometric properties of these assessments. Construct irrelevant variance is likely introduced as it relates to linguistic and cultural factors and are likely to have profound impact on the validity of score inferences with ELLWD students. Other approaches, such as the use of differential item functioning (DIF) should also be used in future research to examine such group differences.

Identifying less accurately measured students

Moen, Liu, Thurlow, Lekwa, Scullin and Hausmann (2009, this issue) conducted a preliminary study to determine if it is feasible to use teacher judgment to identify students at most risk of being misclassified by reading tests. The researchers note that test scores are comprised of random error that impacts all students, but seek to examine systematic error that they hypothesize exists among students with the greatest difference between predicted and actual performance. That is, they attempt to examine the validity of test scores for individual students and determine if teachers' judgment would be useful in those instances when test scores are poor measures of reading skills.

The rationale for the study appears similar to the rationale for previous efforts that have attempted to examine differential validity. Differences in validity coefficients on admissions and other educational tests have been frequently reported across ethnic groups with higher correlations between admissions test scores and college performance among females and whites (Mattern, Patterson, Shaw, Kobrin & Barbuti, 2008; Young, 2001). In a recent validity study between SAT scores and freshmen GPA across 110 institutions the differences in uncorrected correlations were largest between gender groups (0.07) and ranged from 0.01 to 0.05 among ethnic groups. The same study found even larger differences (0.10) between correlations of high school GPA and freshmen GPA. Differences in correlations are important to study in addressing issues of validity, but they have generally been associated with overprediction of minority performance. Such differences have less frequently been found in employment tests and when detected, they have often been attributed to methodological artifacts such as smaller sample sizes in the minority group (Hunter, Schmidt, & Hunter, 1979). In any event, differences in correlations among groups, when present, are not evidence of bias.

Rather than focus on established subgroups, the authors of this study are attempting to extend the differential validity or performance argument to individuals in order to identify some latent traits that could explain the underperformance. The exploratory study is designed to determine if there are underlying traits common among individuals for whom test scores are a less accurate measure of their reading performance. They note that "differential suppression of student performance is often due to a characteristic only some students have that interferes with successful performance on tests" (Moen, et al., 2009, this issue, p. 3).

The study focuses on whether teachers can successfully identify students whose reading skills would be underestimated by reading test results and provide evidence to support their assertions. A total of 77 students in 4th through 8th grade who would perform misleadingly poorly on the reading assessments were identified by 21 teachers across 10 sites. Only 20 of these students participated in the second phase of the study which involved comparisons of teachers' assertions with other evidence (e.g., student statements, observations, brief assessments). The researchers evaluated this evidence and agreed with teacher judgments in 14 of 20 instances. Teachers were able to identify discrepant performers when they focused on students' decoding difficulties, slow processing skills or exceptional difficulty staying on task.

The study provided some evidence that teachers may be successful in identifying some students who they believe will perform too poorly on standardized reading assessments and describing some of the reasons for the poor performance. However, the authors often found that differential performance existed on skills (e.g., decoding, comprehension) that are subsumed in the construct. If such skills are an essential

component of the construct then it may be inappropriate to attempt to modify assessments in order to minimize differences. Future studies might employ a standardized reading test as a dependent variable and examine the discrepancies between three groups of students: (a) those that perform consistently across both measures (teacher judgments and assessment results); (b) those that perform significantly higher on tests; and (c) those that receive significantly higher teacher ratings. This design would help to both examine the validity and consistency of judgments and provide a more objective criterion.

Using Factor Analysis and Differential Item Functioning to Investigate the Impact of Accommodations on the Scores of Students with Disabilities

A key issue in assessing the validity of accommodations and modifications for students with special needs is whether such changes are construct-relevant or construct-irrelevant. In admissions testing, comparisons of predictive validity among accommodated and standard administration conditions have been successfully employed to examine issues of validity. However, this approach does not lend itself to most K-12 tests which lack an agreed upon criterion. Studies of the internal psychometric properties of tests have utility and should be pursued (Koretz & Hamilton, 2006). The next three papers in this special issue employed differential item functioning and/or factor analysis to empirically test the comparability of test scores administered under modified conditions or with accommodations.

There remains substantial debate about whether delivering test content from a reading assessment by audio presentation (e.g., tape, reader) is an accommodation to an existing assessment or a modification that suggests scores may not be comparable. Cook,

Eignor, Steinberg, Sawaki, and Cline (2009, this issue) attempted to examine this issue by investigating the underlying constructs measured by the Gates-MacGinite Reading Tests (GMRT) for students with and without reading-based disabilities who took the GMRT under standard conditions or with a read-aloud change. This study employed a traditional 2 x 2 group design with exploratory and confirmatory factor analysis. Results demonstrate that the test measured a single factor for all four groups and the largest eigenvalue accounted for 59% to 58% of the variance. A single factor solution fit the data optimally and factor invariance held across all groups.

In reviewing previous research studies that presented reading content to students orally, the authors note that such changes in the mode of presentation resulted in no gains or comparable gains for students with and without disabilities, and few items exhibit any differential item functioning (Cook et al., 2009, this issue). The authors note inconsistent findings in two previous studies that employed factor analysis to examine comparability when read-aloud accommodations were provided. While differences in the population, disability, extent and nature of the oral accommodations, and the assessment employed vary across these types of studies, the present study does suggest important empirical evidence that read-aloud accommodations alone may not change the internal structure of the test. Equally important is that over 1,000 students were included in the sample of students with and without disabilities.

DIF has been increasingly useful to determine if an item functions differently for two or more groups in studies of accommodations and other administrative changes. For example, DIF has been used to examine the impact of calculator use and type in performance of math items on the SAT (Scheuneman, Camara, Cascallar, Wendler, &

Lawrence, 2002. Laitusis, Maneckshana, Monfils, & Ahlgrim-Delzell (2009, this issue) employed DIF to investigate performance based items on alternative assessments in English Language Arts (ELA) and math tests across three groups of students with cognitive disabilities. Typically, students without disabilities would serve as the focal group in such studies, but because items came from the alternative assessment, which is administered only to students with disabilities, the focal group in this study could not be students without disabilities.

Laitusis et al., (2009, this issue) sought to determine if specific item characteristics impact the performance of students with three types of cognitive disabilities (mental retardation, autism, and orthopedic impairments). Overall, items with the largest DIF were primarily found in the comparison between the mental retardation and autism groups and with more items identified in ELA than math. All items classified as decoding unfamiliar words had DIF and favored students with autism while about half of the items associated with rote learning also had DIF favoring this group of students. Studies such as this show promise in both assessing efforts to implement universal design and as a post hoc method that can inform future test development efforts. For example, items that required rote learning, with longer attention spans, were verbally administered, required a social exchange, and used first or second person pronouns, appeared to have DIF and may not have been construct relevant. Such items may not be required on an ELA test. In contrast to this conclusion, the authors noted that uniform DIF favoring students with autism was present in items requiring the decoding of unfamiliar words, but this skill appears construct relevant and a necessary component of ELA assessments (Laitusis et al., 2009, this issue).

Steinberg, Cline, Ling, Cook, and Tognatta (2009, this issue) also employed these methods in evaluating ELA assessments for 4th and 8th grade students who were deaf and hard of hearing and non ELL. Specifically, they examined the internal structure of the ELA assessments for consistency across non-disabled and disabled groups, and each group was further split in terms of their ELL status. As expected, the performance of students with disabilities was significantly below that of non-disabled students, with a difference of nearly one standard deviation between the mean performances of the two groups. On average, non-disabled students who were ELL performed slightly below that of deaf and hard of hearing students who were not ELL but significantly below non-disabled non ELL students. Finally, deaf and hard of hearing students who were ELL performed more than 1.5 standard deviations below non-disabled, non ELL students and significantly below the deaf and hard of hearing students who were ELL. Results across 4th and 8th graders were consistent with only one demonstrated substantial levels of DIF (out of 75 items) between the students with disabilities and non-disabled students who were ELL (Dorans & Holland, 1993). Additional comparisons of non-ELL students in these two groups of students revealed no C DIF items. Factor invariance was largely supported across all four groups and a one-factor solution was the best fit for data for all groups. Collectively, results of these three studies suggest that the underlying constructs measured by tests do not change among ELL and disability groups using traditional methods to detect differential item performance or construct invariance. However, consistent and large score differences are present across groups which should be compared with other measures and indicators to determine if other factors (e.g., testing mode) are suppressing the scores of ELL students and students with disabilities.

Conclusion

The Americans with Disabilities Act (1990) and best practices in special education advise that accommodations or modifications to standardized testing practices should measure the necessary skills, without reflecting the individual's impairment. The selection of the appropriate instrument (or assessment) and necessary accommodations should be based on the individual's needs. NCLB mandates greater inclusion of students with disabilities or limited language proficiency, and authorizes the use of alternative assessments that measure the same construct, but may differ substantially in all other surface features.

These legal and regulatory provisions, as well as the genuine well meaning of educators have pushed the profession toward greater variances and exceptions to standard administrative and responding requirements. There is also a tension between legal mandates and professional standards, the latter of which call for large samples to evaluate comparability and provide "normative data from the population of individuals with the same level or degree of disability" to facilitate individualized interpretation of assessment results (AERA, APA, & NCME, 1999, p. 107).

Each of the five papers in this special issue contributes to professional efforts to expand our research designs beyond classical comparability studies. Collectively, they illustrate how research on the internal psychometric properties of tests can be evaluated through DIF and factor analytical approaches, or how item characteristics can be evaluated to detect features that may be construct-irrelevant and ultimately improve test design.

Research on test accommodations is incredibly difficult to conduct because samples available for study are quite small once you consider the type of disabilities, the combination of disabilities, the severity of disabilities, and other relevant individual characteristics and experiences that can impact performance on assessments. Professional standards and practices encourage us to continue to conduct rigorous research to demonstrate comparability, but increasingly we must explore new methods of establishing the comparability of assessments through construct representation rather than simple crossover designs that employ groups of disabled and non-disabled students taking tests that have and have not been changed.

Clearly, there is evidence that impairments as well as accommodated assessments can introduce construct-irrelevance. Traditional empirical approaches to establishing comparability by minimizing departures from standardization and then demonstrating scores do not change or change in the same magnitude across groups of students is not feasible in an environment where alternative assessments may differ in many forms from standardized assessments. Thompson and Way (2007) proposed alternative models of demonstrating comparability between paper and computer-based tests that do not attempt to capitalize only on consistent features but focus on alternative approaches in measuring the same construct. The practioner is most concerned with threats to the validity of inferences made about assessment results and somewhat less concerned with strict comparability. These five papers remind us of the central issues that must be addressed in determining how to fairly assess students to get the most reliable, valid and fair measures.

References

Abedi, J. (2009, this issue). English Language Learners with disabilities: Classification, assessment and accommodation issues. *Journal of Applied Testing Technology.*

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

Americans with Disabilities Act of 1990, Pub. L. No. 101-336, 2, 104, Stat. 328 (1991).

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cook. L., Eignor, D., Steinberg, J., Sawaki, Y., and Cline, F. (2009, this issue). Using factor analysis to investigate the impact of accommodations on the scores of students with disabilities on a reading comprehension assessment. *Journal of Applied Testing Technology.*

Dorans, NJ, & Holland, PW. (1993). *Differential item functioning* (p. 35-66). In Holland, PW, & Wainer, H (Eds.). DIF detection and description: Mantel-Haenszel and standardization. Hillsdale, NJ: Lawrence Erlbaum.

Hunter, J.E., Schmidt, F.L., Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86,* 721-735.

Kane, M.T., 2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64) Washington, DC: American Council on Education and Praeger.

Ketterlin-Geller, L.R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice, 27*(3), 3-16.

Kindler, A.L. (2002). *Survey of the state's limited English proficiency students and available educational programs and services 2000-2001 summary report*. Washington, DC: National Clearinghouse for English Language Acquisition & Language Instruction Educational Programs.

Kobrin, J.L., Patterson, B.F., Shaw, E.J., Mattern, K.D. & Barbuti, S.M. (2008). *Validity of the SAT for predicting first-year college grade point average* (College Board Research Report No. 2008-5). New York: The College Board.

Koretz, D.M. & Hamilton, L.S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 531-621).Westport, CT: American Council on Education and Praeger

Laitusis, C.C., Maneckshana, B., Monfils, L., and Ahlgrim-Delzell, L. (2009, this issue). Differential item functioning comparisons on a performance-based alternative assessment for students with severe cognitive impairments, autism and orthopedic impairments. *Journal of Applied Testing Technology.*

Mattern, K. D., Patterson, B.F., Shaw, E.J., Kobrin, J.L., and Barbuti, S.M. (2008). *Differential validity and prediction of the SAT* (College Board Research Report 2008-4). New York: The College Board.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-100). Washington, DC: American Council on Education.

Milewski, G., Johnsen, D., Glazer, N, & Kubota, M. (2005). *A survey to evaluate the alignment of the SAT writing and critical reading sections to the curricular and instructional practices.* (College Board Research Report 2005-1) New York: College Board.

Moen, R., Liu, K., Thurlow, M., Lekwa, A., Scullin, S., and Hausmann, K. (200). Identifying less accurately measured students. *Journal of Applied Testing Technology.*

National Research Council (1997). *Educating one and all: Students with disabilities and standards-based reform.* Washington, DC: National Academy Press.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 U.S.C. § 1425 (2002).

Rivera, C., Collum, E., Shafer, W.L. & Sia, J.K. (2006). Analysis of state assessment policies regarding the accommodation of English Language Learners. In C. Rivera & E. Collum (Eds.), *State Assessment Policy and Practice for English Language Learners* (pp. 1-174). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Sireci, S.G., Scarpati, S., & Li, S. (2003). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*(4), 457-490.

Steinberg, J., Cline, F., Ling, G., Cook, L., & Tognatta, N. (2009, this issue). Examining validity and fairness of a state standards-based assessment of English-Language Arts for Deaf and Hard of Hearing Students. *Journal of Applied Testing Technology*.

Thompson, T. & Way, D. (2007). *Investigating CAT Designs to achieve comparability with a paper test.* Paper presented at the Applications and Issues Conference of the Graduate Management Admissions Council, Minneapolis, MN.

U.S. Department of Education (2007). *Digest of Educational Statistics*. Retrieved October 19, 2008 from http://nces.ed.gov/programs/digest/d07/.

Vacc, N. A. & Tippins, N. (2002). Documentation. In R.B. Ekstrom and D.K. Smith (Eds.), *Assessing individuals with disabilities in educational, employment and counseling settings* (pp. 59-70). Washington, DC: American Psychological Association.

Young, J. (2001). *Differential validity, differential prediction and college admissions testing: A comprehensive review and analysis* (College Board Research Report 2001-6). New York: The College Board.