No More Excuses: New Research on Assessing Students with Disabilities

Stephen G. Sireci

University of Massachusetts Amherst

Correspondence concerning this article should be addressed to Stephen G. Sireci, Center for Educational Assessment, 156 Hills South, University of Massachusetts, Amherst, MA 01003. E-mail correspondence can be sent to Sireci@acad.umass.edu.

Abstract

The articles in this special issue of the *Journal of Applied Testing Technology* represent significant steps forward in the area of evaluating the validity of methods for assessing the educational achievement of students with disabilities. The studies address some of the most difficult student groups to assess—students with learning disabilities, students with severe cognitive disabilities, deaf/hearing impaired students, students with disabilities who are also English Language Learners, and students who are likely to be inaccurately measured on statewide reading tests. The authors use a variety of research designs and statistical methods to provide evidence for evaluating the validity of assessments of these students. This article highlights the novel contributions of these studies and raises questions for readers to consider as they read each study.

No More Excuses: New Research on Testing Students with Disabilities

Readers of this special issue of the *Journal of Applied Testing Technology* are in for a real treat. This set of five articles represents some of the most significant work in the area of evaluating assessments of students with disabilities (SWD) that I have seen to date, and I have spent considerable time over the past ten years reviewing research in this area. There are several laudable features of this special issue that are particularly significant.

First, the special SWD populations investigated represent some of the most difficult groups of students to study— students with learning disabilities, students with severe cognitive disabilities, deaf/hearing impaired students, students with disabilities who are also English Language Learners, and students who are likely to be inaccurately measured on statewide reading tests. Second, the research designs and statistical methodologies represent some of the most sophisticated designs and analyses applied in this area. Third, all five articles provide extensive and informative literature reviews.

I predict all articles in this special issue will be widely cited because they represent comprehensive study of groups of students that have often been dismissed for empirical study due to small sample sizes or difficulties in obtaining assessment data from them. These groups have been virtually ignored for so long in the psychometric literature that I decided to title my introductory remarks "No More Excuses." As the authors of these articles illustrate, it *is* possible to comprehensively study the validity of inferences derived from test scores of these important groups of SWD. My hope is that this special issue will not only inform readers, but will encourage us to do more to improve the assessment of SWD.

Four of the five studies in this issue not only investigate unique groups of students, they also illustrate new and sophisticated analyses of measurement invariance across these groups and

across test administration formats. A summary of the groups studied, the questions addressed, and the methodologies used in each study are presented in Table 1. The first four studies listed represent the studies that investigated at least one aspect of measurement invariance. The fifth study (Moen et al, this issue) is very different from the others because it focuses on the important pre-assessment activity of identifying students who are not likely to be accurately measured on a reading test. As is evident in Table 1, the questions addressed in these studies are diverse and important. All five articles address specific groups of students. This feature in itself separates the studies from previous research. Many previous studies have compared SWD to non-disabled students only after combining students with very different disabilities into a single group. This aggregation is often done to achieve sample sizes large enough for sufficient statistical power. However, given the increased inclusion of SWD in statewide educational assessments, the need to aggregate over these diverse groups of SWD is diminished. I believe disaggregation of SWD into more homogeneous categories is an important feature of 21st-century psychometrics that distinguishes it from most of the studies in this area conducted during the 20th century. Disaggregation is important because the heterogeneity within SWD is likely to obscure any effect that may be realized for groups of students who are more homogeneous with respect to their disabilities and assessment needs.

The four articles that investigate at least one aspect of measurement invariance represent new and important contributions to the analysis of data derived from assessing SWD. The degree to which the properties of an instrument are invariant across groups of examinees and test administration formats has long been an important area of study in educational testing. The specific aspects of invariance addressed in these four studies are invariance of test structure and invariance of items. Three of the five studies addressed invariance of factor structure across

groups or test administration formats, and one study investigated both item-level invariance (that is, differential item functioning) and factor structure. Three of the five studies represent analysis of operational test data, and one other (Abedi, this issue), presents summaries of previous analyses of operational test data, in addition to a comprehensive review of the issues involved in assessing English language learners (ELL) with disabilities. The fifth study focuses on the important issue of finding better means for identifying students who will benefit from test accommodations.

With respect to the evaluations of measurement invariance in four of the studies in this special issue, the methods applied represent some of the most sophisticated techniques available. The analysis of differential item functioning (DIF) is conducted using the Mantel-Haenszel procedure, which has a large body of research supporting its use in identifying items that "function differentially" across different groups of students (Holland & Wainer, 1993). In the present case, groups are defined by disability status or ELL status. DIF analyses flag items as "functioning differentially" when students who are considered to be approximately equal with respect to the skills being measured on a test respond differently to the item. An item could be flagged for DIF, for example, if it is more difficult for deaf students than it is for hearing students of comparable proficiency. Items flagged for DIF are not automatically considered to be biased, but the DIF flag allows researchers and test developers to scrutinize the item more closely to determine if biasing factors are present and to eliminate such factors from the current and future tests, where warranted.

Laitusis, Maneckshana, Monfils, and Ahlgrim-Delzell (2009, this issue) go beyond flagging items for DIF by using the results to evaluate a priori hypotheses regarding how students with different disabilities will perform on items containing features that are likely to

Download table on page 6 separately

interact with their disabilities or with a specific test accommodation. By linking DIF results to the specific features of items, tests administration conditions, and students, they extend our understanding of how students perceive and respond to specific tasks under standard and accommodated testing conditions. These results are likely to inform and improve future test development and administration efforts. For example, features of items that may interfere with a disability and are not particularly relevant to the skill being measured (e.g., testing vocabulary by selecting the "underlined word" in a passage) may be eliminated from future tests, or modified instructions could be provided to certain groups of students.

The statistical procedures applied to the evaluation of invariance of test structure are also sophisticated and well supported in the literature. Statistical options include exploratory and confirmatory factor analysis and multidimensional scaling. Analysis options include performing separate analyses on the data for each group and then comparing the results, or conducting multi-group analysis simultaneously across all groups. In the Cook et al. (2009, this issue) and Steinberg et al. (2009, this issue) studies, both separate group and multi-group strategies are used. The authors first evaluate the data separately for each group to determine the number of factors underlying the data, and then use multi-group analyses to evaluate whether the *same* factors underlie the data for all studied groups. By employing both procedures, they avoid the possibility of conducting a confirmatory analysis on a model that does not adequately represent the data. Furthermore, by testing nested confirmatory factor analysis models, they are able to test whether the specific aspects of test structure (i.e., number of factors, factor loadings, errors associated with those loadings, and factor intercorrelations) are invariant across groups. The application of both exploratory and confirmatory analyses to the same data set is an important new development in the area of evaluation of factorial invariance.

Statistical Issues to Consider in Evaluating Invariance

Up to this point, I have applauded the authors for their ability to gather and analyze data from unique groups of students (and teachers) and for the sophisticated techniques they used to evaluate their research questions. In this section, I raise a few issues that could affect the results of an analysis of item-level or test-level invariance, and discuss how the authors addressed these issues.

There are at least three issues that may affect the results of invariance analyses—small sample sizes, non-overlapping proficiency distributions, and unreliability of item-level data. These issues are important to consider at the outset because they will inform research design and data analysis choices.

The issue of sample size is important because of its relation to statistical power. In many cases involving SWD, the SWD group is small, particularly in comparison to the non-disabled group. Sufficient sample sizes are needed to identify a lack of measurement invariance when it exists—whether it is at the item level as in DIF analyses or at the test level as in analysis of test structure. Small numbers of SWD have limited research on the invariance of educational tests across SWD and non-SWD populations, and in many cases, analyses of DIF and test structure are simply not conducted. The authors of the current studies solved the small sample size problem by using data from large statewide assessment programs and limiting their analyses to subgroups of sufficient size so that a statistical difference could be discovered, if such a difference existed. The authors also used descriptive statistics and effect sizes to gauge the magnitude of any lack of test structure. In so doing, they minimized the effects of sample size in interpreting the results.

A second important issue when evaluating invariance issues in assessing SWD is the degree to which the proficiency distributions overlap across the groups studied. In many cases, a focal group such as SWD or ELL have test score distributions that are positively skewed and centered far below the distributions of the reference groups. Such differences may affect the invariance comparisons in predictable ways (e.g., very easy items are flagged for DIF in favor of the focal group, the first factor is less "salient" for reference group, etc.) and could be due to overall proficiency differences, rather than differences due to disability status or test administration format.

Abedi (2009, this issue) illustrated that large achievement gaps exist between SWD, ELL, and other groups of students and he pointed out "…assessments that are developed and field tested for the mainstream student population may not provide valid outcomes for these students." If the assessments are not appropriate for a group of students, it is likely their responses to the items contain more guessing, and hence more error, which will obscure the general factor being measured by the test. A better comparison group might be a subgroup of students from the reference group who have a similar distribution of total test scores.

Interestingly, Cook et al. (2009, this issue) use an experimental design (randomly assigning standard or read-aloud conditions) to evaluate the invariance issue *within* SWD and non-SWD groups. In so doing, the proficiency distributions for the accommodated and non-accommodated students in each of the SWD and non-SWD are randomly equivalent. This research design solves the problem of obtaining statistical significance due simply to wide differences in the distributions of proficiency between the groups studied.

A third issue pertains to the evaluation of test structure, and the issue is whether to conduct the analyses at the item level or at some more aggregate level based on subtest scores or

"parcels" of items. The problem with analyses at the item level is the reliability of a response to a single item is low. Thus, factor analyses may fit factors to model the "noise" in item-level data. Earlier in my career, when I had more hair, I spent a lot of time trying to interpret the minor factors that appear in item-level factor analyses. Now I realize they are idiosyncratic and not substantive, and so there is no use in trying to interpret them. For this reason, Cattell strongly advocated for combining items into parcels before conducting factor analyses (Cattell, 1956; Cattell & Burdsal, 1975).

Cook et al. (2009, this issue) and Steinberg et al. (2009, this issue) both use a parceling strategy to address this problem, and as Cook et al. illustrate, even the reliability of parcel-level data can be low (see their Table 2). I believe parceling is an effective strategy for distinguishing noise factors from substantive factors, but there is a danger in that if multidimensionality does exist in the item-level data, parceling the items may wash it out. One way to avoid such danger is to do the analyses on both item and parcel-level data, which I think was done in an earlier study by Steinberg, Cline, and Sawaki (2008), but is not reported in this set of studies.

Theoretical Issues

In addition to statistical issues, there are also theoretical issues to consider when designing studies to evaluate the validity issues in assessing SWD. The most significant issue is what types of evidence are needed to support the inferences derived from the performance of a SWD on a test. Different types of evidence come from different hypotheses related to validity. One validity hypothesis that has been proposed is the evaluation of differential boost, which states that an accommodation improves the performance of SWD on a test more than it improves the performance of students without disabilities (Fuchs & Fuchs, 2001). This hypothesis can be challenged however, because better performance for SWD does not necessarily mean more valid

assessment. Nevertheless, the finding of differential boost or disability group-by-accommodation interaction can support the use of accommodations for specific groups of students (Sireci, Scarpati, & Li, 2005).

Other validity hypotheses relate to invariance of a test across SWD and non-SWD groups or across accommodated and non-accommodated test administrations. These are the hypotheses addressed in the articles in this special issue. Invariance hypotheses take many forms. When the invariance of factor structure is evaluated it is often described as evaluating "construct equivalence;" however, factorial invariance is not the same as construct equivalence. Rather, it refers to the situation where the same factors can be used to describe students' performance on a set of items. Factorial or structural invariance is one aspect of construct equivalence, and an important aspect, but it does not represent a complete evaluation of whether the constructs assessed are equivalent across groups or administration formats. Other aspects of construct equivalence would involve comparison of correlations of test scores with other relevant criteria (e.g., statistical comparison of nomological networks), analyses of the cognitive processes students use to respond to items, and the degree to which the actions taken based on interpretations of test scores are consistent. Using the language of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), the evaluations of construct equivalence conducted by the studies presented in this special issue have focused on validity evidence based on internal test structure, and possibly also test content, but not on the other three "sources of validity evidence" (relations to external variables, response processes, and testing consequences). Thus, laudable though these studies are, they represent

only one approach for evaluating the validity of interpretations derived from the scores SWD achieve on educational tests.

A second theoretical issue pertains to the interpretations to be validated. I describe the studies in this special volume as validity studies, and since validation involves gathering evidence to evaluate interpretations of test scores (Kane, 1992, 2006; Messick, 1989) many interpretations may be of interest. For example, we may want to evaluate whether scores for SWD can be interpreted in the same way as scores for students without disabilities. Alternatively, we may want to evaluate whether scores from accommodated test administrations can be interpreted in the same way as scores from standard administrations. Both cases refer to analyses at the group level, but would require different populations and research designs.

A third validity evaluation may investigate the validity of a test score for specific types of SWD or for an individual SWD. These investigations would also require different research designs and probably different types of data. For example, validation of a specific test score for a specific student would require gathering additional achievement data on this same student. Since validation activities must always focus on the intended and unintended uses of tests and on testing purpose, different uses and purposes will require different types of studies[1]. The Moen et al. (2009, this issue) study illustrates important new work in helping improve the validity of interpretations at the individual student level by exploring new means for determining the best assessment options for specific students.

One other theoretical issue to consider in assessing SWD is the heterogeneity within the SWD population. As I mentioned earlier, one of the features I really like about this special issue is that the studies focus on important subgroups of SWD rather than collapsing very different

[1] For a discussion of validity issues related to testing SWD for the purposes of college admissions, see (Sireci, 2005).

groups into a single SWD category. However, within subgroups heterogeneity will always exist to some extent, and one thing I have learned in reviewing the literature in this area is that *the composition of SWD* subgroups may change across grade levels. For example, in a recent study I noticed that the proportion of students with speech-language disabilities decreased markedly across grade levels (Engelhard, Fincher, & Domaleski, in press). Such observations regarding the composition of SWD subgroups across grades may help us better understand why the effects of test accommodations sometimes differ across grade levels.

Questions for Readers to Consider

I can talk and write about validity issues in assessing SWD all day, but I do not want to delay your reading of the articles in this special issue any further. I would, however, like to raise a few questions I think might be helpful to consider as you read these important articles.

1) *How can analyses of DIF and test structure inform*
    a. *future test development?*
    b. *our interpretations of test scores?*
    c. *decisions about test accommodations?*

As Laitusis et al. (2009, this issue) and Stone et al. (in press) illustrate, analyses of DIF can be used to identify aspects of items that interact with students' disabilities. Similarly, Cook et al. (2009, this issue) and Steinberg et al. (2009, this issue) illustrate methods that can be used to identify potentially problematic sections of an assessment for SWD. What work needs to be done to decide if such aspects are construct-relevant or construct-irrelevant? Should we avoid certain item formats or test instructions? If so, do we improve or diminish validity? How does this research help us interpret test scores of SWD?

2) *What types of analyses should state assessment programs be conducting to inform test accommodation and interpretation decisions?*

The No Child Left Behind (NCLB) legislation has had an important and positive impact on the education of SWD because it forces states to provide data on the educational achievement of these students. Consequently, states are required to include SWD in their assessment program and are sanctioned if such students do not participate. However, standard test administration formats may present barriers for SWD and so accommodations to standard test administration formats may be needed (see Sireci, Scarpati, & Li, 2005 and Zenisky & Sireci, 2007 for reviews of accommodation formats and their effects). The studies in this special volume help us evaluate the degree to which scores from certain test accommodations granted to certain groups of students are similar to scores from standard test administrations. However, the results may not generalize beyond the subject areas, grade levels, and state assessment programs studied. Each state needs to know when to provide an accommodation to a student and when such accommodations might change the skills measured on a test. The research reported in this special issue should help inform educators when scores from accommodated tests can be used to confirm students have obtained certain skills (e.g., when tests are used as part of a high school graduation requirement) and when it is appropriate to include such scores in aggregate measures of performance, such as those required for determining adequate yearly progress under NCLB. All states should be encouraged to conduct research to evaluate the validity of the use of SWD's test scores.

3) *How much confidence can we have in the degree to which students are correctly classified into SWD and ELL categories?*

Abedi (2009, this issue) raises this important question and provides some data to suggest students may not be classified appropriately. Such misclassifications are troubling because it directly affects the quality of students' education. We know it is hard to correctly identify students' disabilities and this is where assessments can help, if they are used and interpreted appropriately. Abedi reminds us that disability status and language proficiency are separate characteristics, and when students are English Language Learners with disabilities that may inhibit their learning, accommodations for both instruction and assessment are likely to be needed. Moen et al. (2009, this issue) illustrates research that can help improve the classification of students into appropriate test administration conditions. Clearly, more work in this area is needed.

Closing Remarks

I appreciate the opportunity to provide some introductory comments on the problems addressed by the articles in this special issue of the *Journal of Applied Testing Technology*. These articles represent cutting edge research in the area of assessing SWD. Of course, there is always more research to be done, particularly in an area as difficult as assessing SWD. Future research directions could include enhanced methods for coding items, analyzing test structure, and synthesizing results. There are certainly more state testing programs, other important subgroups of SWD, more subject areas, and more grade levels to examine. Thankfully, the authors of these studies have provided us with stellar examples of how to conduct research in these areas, and have given us concrete results we can learn from to improve future test development and test administration efforts to facilitate more valid measurement of SWD. I am encouraged by the accomplishments of the authors of these studies, and I am inspired to do more work in this area. Data for evaluating validity hypotheses pertaining to SWD are becoming more

available, and as the authors of these studies have illustrated, there are new and interesting statistical and qualitative methods available to assist us.

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

Cattell, R. B. (1956). Validation and intensification of the Sixteen Personality Factor Questionnaire. *Journal of Clinical Psychology, 12*, 205-214.

Cattell, R. B., & Burdsal, C. A. (1975). The radial parcel double factoring design: A solution to the item-vs.-parcel controversy. *Multivariate Behavioral Research, 10*, 165-179.

Cook, L., Eignor, D., Steinberg, J., Sawaki, Y., & 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*.

Engelhard, G., Fincher, M., & Domaleski, D.S. (in press). Mathematics Performance of Students with and without Disabilities under Accommodated Conditions using Resource Guides and Calculators on High-Stakes Tests. *Applied Measurement in Education*.

Fuchs, L. S., & Fuchs, D. (2001). Helping teachers formulate sound test accommodation decisions for students with learning disabilities. *Learning Disabilities Research & Practice, 16,* 174–181.

Holland, P.W., & Wainer, H. (Eds.). (1993*). Differential item functioning*. Hillsdale, New Jersey: Lawrence Erlbaum.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*,527-535.

Kane, M. (2006). Validation. In R. L. Brennan (Ed). *Educational measurement (4th edition*). Washington, DC: American Council on Education/Praeger.

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

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

Moen, R., Liu, K., Thurlow, M., Lekwa, A., Scullin, S., & Hausmann, K. (2009, this issue). Identifying less accurately measured students. *Journal of Applied Testing Technology*.

Sireci, S. G. (2005). Unlabeling the disabled: A perspective on flagging scores from accommodated test administrations. *Educational Researcher, 34(1)*, 3-12.

Sireci, S. G., Scarpati, S., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*, 457-490.

Steinberg, J., Cline, F., Ling, G., Cook, L., & Tognatta, N. (2009, this issue). Examining the 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*.

Steinberg, J., Cline, F., & Sawaki, Y. (2008, March). *Examining the internal validity of a state standards-based assessment of science*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Stone, E., Cook, L., Laitusis, C. C., & Cline, F. (in press). Using differential item functioning to investigate the impact of testing accommodations on an English language arts assessment for students who are blind or visually impaired. *Applied Measurement in Education.*

Zenisky, A. L., & Sireci, S. G. (2007). *A summary of the research on the effects of test accommodations: 2005-2006 (Technical Report 47)*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Available at http://cehd.umn.edu/nceo/OnlinePubs/Tech47/default.html.