

## **Evidence-Based Plan: Technical Adequacy of Assessments for Alternate Student Populations**

### A Technical Review of High Stakes Assessment Tests for English Language Learners and Students with Disabilities

#### Background and Goals

In recent years, several developments in education have converged to create a critical need for reliable, valid, unbiased methods for conducting high stakes assessment tests for students who are English language learners (ELL) or who have disabilities (Kubiszyn & Borich, 2003; Heubert & Hauser, 1999). One trend is that toward accountability in the school systems for student learning, and the path to accountability increasingly has been high stakes testing as a means to provide evidence to parents, policymakers, politicians and taxpayers that schools are performing. High stakes include more than decisions regarding the student (e.g., promotion or graduation) but also decisions regarding teachers, the district and the state. Accordingly, the interest in having assessment tests that meet standards requirements is keen. Another development was the Individuals with Disabilities Education Act (IDEA), which required that students with disabilities (SDs) are given full access to educational opportunities, and that access includes high stakes testing. The third development is the demographic change in the United States, such that a larger number and proportion of students in public schools live in families where languages other than English are spoken at home. These ELL students, sometimes referred to as Limited English Proficiency (LEP) students, also require fair access to the assessment tests. Finally, these three trends merged upon the signing of the No Child Left Behind (NCLB) in 2002; all states are now required to conduct assessment tests to ascertain whether all children are learning.

NCLB has focused attention on the performance of students who are traditionally underserved and underperforming. Assessing the achievement of two of these student groups – ELLs and SDs – has created a range of difficulties for states implementing the high expectations of NCLB. To assess these populations, states use three categories of approaches: (1) provide accommodations, such as increased time; (2) incorporate principles of Universal Design (UD) (Bowe, 2000, Rochester, 2004), and develop a portfolio of alternate assessments. Each of these approaches has faced considerable technical challenges, especially in regard to the critical requirement that the assessments be unbiased, reliable and valid.

Nevertheless, a good deal of work has been done in the area of providing ELL and SD students with access to assessment tests so that they can best tested fairly along with general education students. Studies have been conducted to evaluate the performance of these three core approaches. Other evidence, such as materials presented in conferences and workshops, or internal working papers and reports, also exists. Unfortunately, these various sources of evidence are not aggregated in any methodical fashion, so that a complete understanding of quality for these assessments is not available. Further, evidence supplied by some studies and papers has not been conducted in accordance with scientific methods of inquiry, so that their results may be inconclusive. This technical analysis proposes to rectify those problems by assembling available assessments for ELL and SD students, subjecting them to a review for

scientific quality using evidence from varied sources, and developing guidelines for using these techniques for inclusion from the results.

### Issues in researching assessment test quality for SDs and ELLs

The goal for assessment tests provided to specialized populations, such as ELL and students with disabilities is that the test remain the same test. When investigating validity in providing accommodations, for example, there should be no difference in the results for specialized students who received the accommodations compared to general education students who did not. The accommodation neither gives an advantage to the specialized students, nor does it modify the alignment of the material of the test with the curriculum or the state standards, alter any other components of validity, or introduce problems of bias or lack of reliability. A similar situation would prevail for studies that incorporated principles of Universal Design.

The assessment test should first be designed with Universal Design principles. Then, before the assessment test undergoes whatever accommodations necessary for specialized student populations, it should have undergone an audit for tests of validity, reliability, and bias. Subsequently, accommodations are applied and themselves would ideally undergo the study of possible threats to validity and reliability, or signs of bias. In addition to the threats to validity for assessment tests in general, there are threats that are specific to the impact of accommodations. It is important that any aspect of providing the assessment tests to the ELL and SD students does not increase the performance of both alternate and regular students. Test designers need to ask many questions when incorporating UD and accommodations of all forms. Does the accommodation affect alignment when the language is changed, or the form of administration is altered? Is the interpretation now altered by including scores of students with accommodations (especially with special education students) and those without? Haertel (1999) and Ryan (2002) encourage input from all relevant stakeholders to decrease the threats through increased understanding of the entire process.

There are several key forms of validity for which one needs to examine potential threats (Shadish et al., 2002; Ryan, 2002):

- |                |  |
|----------------|--|
| Construct:     | Does it measure what it is supposed to measure?  |
| External:      | Are the results generalizable to other groups, settings?   |
| Content:       | Does the test align with the measures and objectives?  |
| Internal:      | Does the score level possess an accurate correspondence to the outcome?  |
| Predictive:    | Do the results predict the future performance for which the score is applied?  |
| Consequential: | Does the use of the scores to make high stakes decisions cause other consequences that work against the goals of the assessment? |

Several researchers are particularly concerned about consequential validity: does using the resulting scores to make the high stakes decision begin to affect the test administration, pedagogy, or curricula.(Ryan, 2002; Lane & Stone, 2002)? The inclusion of accommodations has an impact on the administration, set, setting, instructions, and teacher interaction with examinees, all of which potentially introduce a good deal of threats to validity as well as reliability and bias. Add to this the possibility of teaching to

the test, narrowing of the curriculum, and increased drop-out rates, one can argue that the consequences of high stakes assessment tests may be the most important source of concern (Kane, 2002).

### Bias in Assessment Tests

Assessment tests always need to ensure that they are unbiased, that is, there is no difference in the way a subgroup of students answer the tests because they are a member of that subgroup. In general, this concern of bias applies to differences along lines of race, gender, religion or other protected subgroups in American society. With a disabled or ELL population, it is essential that the assessment test produces no differentials owing to one's particular disability or due to one's membership in a culture, language or dialect. In creating and analyzing assessment tests, sources of potential bias are considered by content experts and then undergo a statistical analysis for evidence of differential item functioning (DIF).

### Reliability in Assessment Tests

Consistency of results in different contexts is another hallmark of assessment tests. It should work not just in one school or classroom, but in all targeted classrooms, and it should yield more or less the same response for the same student at different periods of time (recognizing the existence of temporary personal variation in affect, health, etc.). While there is a need to update exams, an assessment test should work over time as well. There are specific longitudinal designs that assist in analyzing this characteristics of an assessment test. Another form is internal consistency, that students who get a certain question correct can be predictably estimated to get another question correct as well. This form, too has analytic techniques for diagnosing, such as split-half or Kuder-Richardson methods. Another component of reliability is alternate form, so that if there is more than one version, they too should work together consistently.

### Using Universal Design for Students with Disabilities English Language Learners

Tests including principles of Universal Design level the playing field and are an encouraging path. Incorporating these principles in educational assessment tests (Bowe, 2002; Rochester Institute of Technology, 2004) can avoid differences in testing between alternate and regular student populations in the first place.

### Accommodations for Students with Disabilities and English Language Learners

Researchers of accommodations frequently remind their audience that it is the accommodation, not the disability, that is under scrutiny. For that reason, Thurlow et al. (2000) advise focusing on the accommodations of greatest interest. Carrying out the technical analysis of these forms of inclusion requires that we assemble a 'database' of high stakes assessment tests that have been used for these specialized student populations and have been subject to the scrutiny of technical review. To that end, we will a number of requirements for acceptance of such studies as evidence for conclusions about the use of such assessment tests. For example, when examining the impact of accommodations, the method is to compare the specialized student group in an accommodated and non-accommodated test situation, and with a regular student

group, divided into one that experiences the accommodations in question and one that does not. In some cases, multiple accommodations are used concurrently, also complicating analysis (Koenig, 2002). Various other scientifically sound possibilities exist for designing the control and test groups (Thurlow et al., 2000; Ryan, 2000).

As noted earlier, strategies to include SDs and ELLs in high stakes assessment tests are varied. The accommodations for students with disabilities include the following:

- More time and breaks
- Paraphrasing, Dictation, Cueing
- ASL interpreter
- Instructions and/or test read out loud
- Modifications to the physical setting
- Braille tests
- Assisted technologies

Accommodations for English Language Learners include:

- More time and breaks
- Translations
- Dictionaries or glossaries.
- Modified language
- Tests given orally (administration effects)
- Assisted technologies

The difficulty in performing an analysis of technical adequacy in assessment tests for specialized student groups is that there are numerous issues that confound the understanding of accommodations and their success in providing access to SD (Elliot & Trimble, 1999; Koretz & Barton, 2003-4; Heubert & Hauser, 1999; and ELL populations (Abedi et al., 2004; Heubert & Hauser, 1999). If one changes the language in the instructions or test items it is possible that the construct or content validity is altered, or that new differential item functioning (DIF) bias is introduced. When allowing changes to the set and setting of the exam, or additional assistance in the form of oral instructions or physical surroundings, then it is also possible that reliability becomes suspect because of the variability in delivery.

This technical analysis will therefore consider the presence of these confounding factors in studies evaluating the accommodations. Indeed, for both populations there is the initial decision of placement: which students are designated to specialized populations.

With SD students, learning disabilities are especially difficult to categorize (McDonnell et al., 1997). The fact that the percentages vary from state to state suggest that there are inconsistencies in assignment to special education group (Koretz and Barton, 2003-2004), and this inconsistency threatens the validity and reliability of all accommodations for students with disabilities. ELL students exhibit tremendous variation in language competency (English and native), so this capability is the first to be assessed before being tracked into a set of accommodations or an assessment test otherwise provided to these students (Rivera, et al., 2000). How long the child has been in the school system is another issue that is not uniformly handled.

Because a good deal of the accommodations provided to students with disabilities are provided by the teacher (or proctor), the potential for threats to validity and reliability are substantial. In addition, the disability may be related to the construct being examined. A technique may also be shown to be ineffective. For example additional time alone does not appear to be as helpful as hoped because it does not change access, whereas providing physical accommodations for those with motor skills does provide promise.

For ELL students, the level of literacy in English does not correlate to a level of literacy in the native language. For example, the student may know 'street' native language, but the vocabulary for the school-taught subjects may be stronger in English. Cultural use of language, nuances, topics, values and more may inhibit simply translating words. Poverty and social class can be part of the dynamics of literacy as well, so that, for example, test scores can vary depending on whether a needs-based meal program is offered in the school system. Reducing the language issues per se improves the scores of LEP students, such that the gap between LEP and regular students for science and math is less than for tests where language requirements are more prominent, and gaps all but disappear in math computations.

In addition, merely translating to Spanish ignores the linguistic and cultural differences between, say, Mexicans, Dominicans and El Salvadorans. Translations are not cost-effective when there are numerous native languages represented in a school district, so making translations available for one or two languages but not others can be an inequity. Providing dictionaries and glossaries is another approach taken, but this too has been found to be problematic. Abedi et al. (2004) found that using published dictionaries was fraught with lack of comparability and reliability. Some published dictionaries include English-only, others are bilingual, and the content varies tremendously. Because these differences pose serious threats to validity, the authors recommend using tailored glossaries.

### Research Goal

The goal of this technical analysis is to address the following research question:

1. What is the technical adequacy of high stakes statewide assessments developed to assess the academic achievement of Special Education and English Learners student populations, when using UD and accommodations? That is, is there adequate supporting evidence that assessment tests were valid, reliable and unbiased?

However, in the process of addressing the first question we will be able to address the following two questions as well:

2. What is the "state-of-the-art" in developing, planning, and implementing bias, reliability and validity studies for assessments developed to assess the academic achievement of Special Education and English Learners student populations?

3. How has the technical adequacy of high stakes statewide assessments developed to assess the academic achievement of Special Education and English Learners student populations changed since the advent of NCLB?

### Selecting Assessments Tests for the Technical Analysis

We shall assemble assessment tests used by states for ELL and SD students, and associated evidence that can be used to indicate whether the assessment passes standards for avoiding bias, and threats to validity and reliability. This evidence used to assess test appropriateness are scattered, however, and much is known to suffer from lack of scientific rigor. In the course of the technical review, we will focus on the assessment tests where scientifically sound evidence exists, ascertained through subjecting them to an audit for scientific method or other criteria, and then using the results of satisfying evidence for our analysis and conclusions.

To ensure the rigor of this proposed technical analysis, WestEd will review and synthesize a wide range of available evidence including journal-published studies, technical reports and manuals, and conference presentations. Several recent publications provide initial starting places for identifying such studies, such as Rivera, et al., 2000; and Abedi, et al., 2004 for ELLs; and Koretz & Barton, 2003-4 and Thompson & Thurlow, 2001 for SDs. Additionally, the external consultants on this team bring a wealth of knowledge about research activities. Accordingly, our procedure to implement and develop the materials for review, we will use the following resources:

- Technical Analysis Advisory Panel consisting of nationally recognized experts in the areas of assessment/accountability (e.g., Dr. Ed Haertel, Stanford University), special education (e.g., Dr. Martha Thurlow, NCEO), English learners (e.g., Dr. Jamal Abedi, UCLA/CRESST), and innovative states (e.g., Scott Trimble, Kentucky Department of Education).
- The National Assessment and Accountability Work Group—this supported group has been working on NCLB technical issues for the past several years;
- Journal articles and conference presentations
- Technical Bulletins, Reports, Workshop proceedings, and Manuals from state assessment programs and national programs (e.g., NAEP, NRT's, SAT/ACT);
- Reliability and validity studies from relevant assessment;
- Independent experts for final review of research findings

In developing the review criteria, we will be able to eliminate those studies with insufficient technical merit to even consider in this analysis. Based on a preliminary review of published and internal studies, we anticipate most studies will not fall into the more rigorous ends of our review rubric, especially those “homegrown assessments” developed informally or with limited resources. Consequently, we will focus most of the project’s attention on those more formal studies with sufficient resources behind them to meet technical adequacy requirements. We anticipate identifying a few key states that are doing exceptional work with either EL or special education assessments; we will highlight the process that allowed this extraordinary work to take place. Our final report will identify those features of that top subset of research that permit generalizability of results to other states and programs.

## Technical Analysis Procedure

After selecting the assessments for which adequate scientific evidence exists and eliminating those without, we will then proceed to the rating process. We will evaluate the technical rigor of the methods used in these studies, up against standards in the field (AERA/APA/NCME joint standards, 1999). Based on a preliminary review of published and internal studies, we anticipate most studies will not fall into the more rigorous ends of our review rubric, especially those “homegrown assessments” developed informally or with limited resources. Consequently, we will focus most of the project’s attention on those more formal studies with sufficient resources behind them to meet technical adequacy requirements. We anticipate identifying a few key states, such as Kentucky, Minnesota, Maryland and Texas, that are doing exceptional work with either EL or special education assessments. We will highlight the process that allowed this extraordinary work to take place. Our final report will identify those features of that top subset of research that permit generalizability of results to other states and programs.

Ultimately, the assessment tests for all students, and for the alternate populations of ELLs and SDs must pass the criteria for validity, reliability and freedom of bias. There are several categories of validity that will be examined in the course of this project, notably: content, construct, criterion-related, predictive and consequential. And, for each of these threats to validity, several kinds of research queries exist for testing and identifying their presence (Ryan, 2002 and Thurlow, et al., 2000). Similarly, there are methods to identify whether a test is a reliable measure of student performance. Finally, the issue of bias, often discussed in terms of differential item response (DIF), will be considered, in order to verify that no groups of students (race, gender) respond differently to the same question such that the question is not a reflection of their knowledge or capability.

### *The rating system*

We will incorporate a rating system consistent with the principles underlying scientifically based research as exemplified by the What Works Clearinghouse (WWC). WWC established a set of standards around assessment validation research (WWC, 2004a). This rating system will be based on incorporating an adaptation of the joint AERA/APA/NCMA (1999) standards into a rubric by which the technical quality of the available evidence can be judged.

For example, an internally produced state specific report may receive less weight than a refereed journal article comparing results across several states based on specific technical criteria. Additionally, an assessment that reports multiple convergent validity evidence may be judged as more rigorous than one focusing only on content validity. We envision a rating system where each study is reviewed relative to each of the various standards included in the “joint standards,” organized by such super-ordinate headings as technical and legal requirements. Because standards are an ideal type to be upheld but often not found in their entirety in actuality, the consultants will provide assistance in deciding where the lack thereof is a serious threat to scientific evidence. We propose using the project’s technical advisory panels to finalize appropriate review criteria to

apply to the range of alternate assessment approaches and models being used to assess alternate student populations under NCLB.

Each criterion for quality in the technical review implies that correct procedure has been followed according to standards, and that the analysis to identify any problems has also been conducted in a credible fashion. For assessment-related studies, the AERA/APA/NCME joint standards (1999, especially chapters 9 and 10) as well as a century of research practice have led to the development of sophisticated research methods (e.g., item analyses, equating methodologies, DIF analyses, generalizability studies, various models and evidence for validity) and specific highly-technical criteria against which high-stakes assessments can be reviewed. The approach we will take to rate the assessment tests will be as follows: First, we will use the well-established methodology for evaluating assessment tests. Second, the AERA, et al. (1999) standards will form the basis of criteria. Finally, the team of experts will contribute their core of knowledge to refine the criteria. The categories of criteria for which we will examine are:

- Existence of appropriate test developmental process in concert with the AERA et al., 1999 standards and other techniques such as Universal Design. Standard setting (setting the cut scores). Acceptable procedures that have been used not just at proficiency but other performance.
- Reliability as examined using recommended analytic procedures.
- Validity as examined using recommended analytic procedures.
- Bias using a combination of committee and analytic review processes
- Item analysis using Item Response Theory (IRT)
- Specific issues for ELLs compared to SDs
- Administration, setting, scoring, reporting, high stakes decisions made

To allow comparisons of our review of assessment tests to other technical reviews, we shall adapt the What Works Clearinghouse (WWC 2004b) study rating system to present our findings, subject to our unique adjustments:

- ✓✓ "Meets Evidence Standards"—[
- ✓ "Meets Evidence Standards with Reservations"
- X "Does Not Meet Evidence Screens"--studies that provide insufficient evidence of causal validity or are not relevant to the topic being reviewed.

#### *Analysis for the Technical Review*

During this analysis, we will use a methodology as outlined in Thurlow, et al., (2000) for assessment test research. We will categorize available accommodations (e.g., Setting, Timing, Scheduling, Presentation, Response, Other). Disabilities themselves will be categorized. While there are 13 identified disabilities, they can be

grouped into 4 or 5 categories (e.g., Motor, Sensory, Learning Disability, Emotional) and cover over 85% of students.

Auditing the analytic studies for methodological integrity is the primary focus of this effort. Item Response Theory (IRT) is necessary for evaluating “the extent to which the abilities measured by the individual items of a test are changed substantially as a result of an accommodation” (Thurlow, et al., 2000, p. 8). Factor analyses are necessary for evaluating the construct validity of tests. Criterion-related analysis varies, using methods such as correlation or multiple regression, and addresses the question: “Is the relationship between this score and other criteria the same?”

Experimental or quasi-experimental designs are required for true understanding of the accommodations. We will evaluate the methods used for sampling in particular, and definition of who is included in the group designated for accommodations. There are several appropriate research designs that set up test and control groups (see Thurlow, et al., 2000, pp. 15-27).

We will develop a rubric for evaluating studies, that is, a list of criteria to be considered when examining the research methodology and in creating the set of guidelines for use of assessment tests for the two specialized populations. One aspect is that of the quality of research design or other evidence, and the other is the inclusion in the research of relevant factors:

#### Research Design

- Presence of appropriate analytics
- Forms of validity investigated
- Control and comparison groups
- Sample selection (randomized, representative or convenience)
- Sample size
- Accommodation examined alone or in combination with other accommodations or inclusion strategies.
- Criteria for inclusion in specialized student group

#### Relevant Factors

- Type of test (norm-referenced or criterion-referenced)
- Grade
- Disability type
- Language
- Subject
- Contextual variables such as SES, population characteristics, school district characteristics, state policy requirements or orientations
- Year(s) developed for investigation of impact of NCLB.

After selecting the studies that meet the scientific quality criteria, the next step will be to identify under which circumstances the accommodation worked and in which situations it did not. Included in this portion of the analysis will be contextual issues, such as cost of translating into more than one other language, public policies around these specialized students and the accommodations themselves, history in the state with assessment tests and accommodations, and more.

Decisions to use an accommodation occur in a societal, economic and political context. Therefore, we will also consider whether the accommodations are feasible as well. For example, translating a test into one language may be affordable in a school system with only one language other than English, but prohibitive in a system where dozens of native languages exist. Similarly, giving one language group a translated test but not others would be viewed as unfair.

### Tasks to be Completed for the Technical Analysis

- Develop initial review criteria and rubrics both for including studies in this analysis and judging their technical adequacy;
- Convene advisory panels for initial discussion review and approval of technical analysis' focus, needs, and parameters, as well as review criteria and rubrics;
- Perform literature search based on advisory panels' recommendations and staff expertise and experience;
- Develop draft review criteria for technical documents, research syntheses, published studies (e.g., inclusion/exclusion criteria);
- Reconvene advisory panels to examine literature search's initial findings and analysis' application of review criteria and rubrics;
- Complete technical analysis and prepare draft preliminary report;
- Perform web-based and meeting-based review of draft preliminary report by advisory panels and other identified experts;
- Finalize technical analysis;
- Prepare additional final deliverables to ensure widespread use of findings of technical analysis (e.g., journal articles, knowledge briefs, conference presentations, guidelines for review of alternate assessments).

### Dissemination of Research Findings

In order to ensure that the findings of this technical analysis be used as widely as possible by a variety of audiences (of varying technical training and responsibility), we plan to develop the following range of final deliverables:

- Final report of technical analysis detailing method and findings;
- Journal articles for psychometric and policy journals [
- Knowledge briefs for less technical audiences;
- Conference presentations (AERA, NCME, CCSSO);
- Guidelines for development and review of alternate assessments for use by states and contractors charged with developing assessments for alternate populations.

## Project Team

A project such as this requires the input of those with vast experience in the fields of expertise required, that is, knowledge about assessment tests in general, applications for the ELL and SD populations, and hands-on experience in conducting high-stakes assessment tests for specialized populations at the State level. To that end WestEd has assembled an admirable team of leaders in their respective fields.

<b>Name</b>	<b>Affiliation</b>	<b>Role in Project</b>
		Project director Lead research: Special Education
		Lead Researcher: English Language Learners
		Special Education consultant
		English Language Learner consultant
		Consultant regarding state behavior and experience
		Consultant on validity of assessment tests

## Project Schedule

<b>Milestone</b>	<b>Completion Date</b>
Develop preliminary Technical analysis Plan	August 2004
Finalize Technical analysis Plan	September 2004
Develop draft review criteria and rubrics	October 2004
Convene Advisory Panels	November 2004, ongoing
Perform Technical Analysis	November 2004, ongoing
Develop Draft Final Report	September 2005
Complete Final Report	October 2005
Complete additional deliverables	November 2005

## Budget Estimate

\$250,000

## References

- Abedi, J., Hofstetter, C.H. and Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, 74(1): 1-28.
- AERA/APA/NCME (1999). *Standards for Educational and Psychological Testing* (3<sup>rd</sup> edition). Washington, DC: Author.
- Bowe, F. G. (2000). Universal design in education: teaching non-traditional students. Westport, CT: Bergen & Garvey.
- Elliot, J. and S. Trimble (1999). Performance trends and use of accommodations in statewide assessment: Students with disabilities in the KIRIS On-Demand Assessments from 1992-93 through 1995-96. Presented at the 1999 NATD Symposium: Issues and Trends in Inclusive Assessment Practices, Montreal, Canada.
- Haertel, E.H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, Winter: 5-9.
- Haertel, E.H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice*, Spring: 16-21.
- Heubert, J. and Hauser, R., editors, (1999) *High Stakes: Testing for Tracking, Promotion, and Graduation*. National Research Council. National Academic Press: Washington, DC.
- Kane, M. (2002) Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, Spring: 31-41.
- Koenig, J. A., editor (2002). *Reporting Test Results for Students with Disabilities and English Language Learners: Summary of a Workshop*. National Research Council, National Academy Press: Washington, D.C.
- Koretz, D. and Barton, K. (2003,2004). Assessing students with disabilities: Issues and evidence. *Educational Assessment* 9(1&2): 29-60.
- Kubiszyn, T. and Borich, G. (2003). *Educational Testing and Measurement: Classroom Application and Practice*. Wiley/Jossey-Bass Education (Seventh Edition), John Wiley and Sons: New York.
- Lane, S. and Stone, C.A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice*, Spring: 23-30.
- McDonnell, L.M., McLaughlin, M.J., and Morison, P., editors (1997). *Educating One and All: Students with Disabilities and Standards-Based Reform*. National Research Council. National Academy Press: Washington, DC.

Rivera, C., Stansfield, C.W., Scialdone, L. and Sharkey, M. (2000). An analysis of state policies for the inclusion and accommodation of English language learners in state assessment programs, 1998-99. George Washington University, Center for Equity and Excellence in Education: Arlington, VA.

Rochester Institute of Technology (2004). Class Act: Access for Deaf and Hard-of-Hearing Students. Online source: <http://www.rit.edu/~classact/side/universaldesign.html>

Ryan, K. (2002). Assessment validation in the context of high-stakes assessment. *Educational Measurement: Issues and Practice*, Spring: 7-15.

Shadish, W., Cook, T. & D. Campbell (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton-Mifflin: New York.

Thurlow, M. L., McGrew, K.S., Tindal, G., Thompson, S. L., Ysseldyke, J. E., & Elliott, J. L. (2000). *Assessment accommodations research: Considerations for design and analysis* (Technical Report 26). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved November 28, 2004, from the World Wide Web: <http://education.umn.edu/NCEO/OnlinePubs/Technical26.htm>

What Works Clearinghouse (2004a). *WWC Study Review Standards*. Online publication: [http://www.whatworks.ed.gov/reviewprocess/study\\_standards\\_final.pdf](http://www.whatworks.ed.gov/reviewprocess/study_standards_final.pdf).

What Works Clearinghouse (2004b). *WWC Evidence Standards*. Online publication: <http://www.whatworks.ed.gov/reviewprocess/standards.html>.