Identifying Less Accurately Measured Students

Ross Moen, Kristi Liu, Martha Thurlow, Adam Lekwa, Sarah Scullin, and Kristin Hausmann

University of Minnesota, Twin Cities

Abstract

Some students are less accurately measured by typical reading tests than other students. By asking teachers to identify students whose performance on state reading tests would likely underestimate their reading skills, this study sought to learn about characteristics of less accurately measured students while also evaluating how well teachers can make such judgments. Twenty students identified by eight teachers participated in structured interviews and completed brief assessments matched to characteristics their teachers said impeded the students' test performance. Researchers found information from evidence provided by teachers, teacher and student interviews, and student assessments that confirmed teacher judgments for some students and information that failed to confirm or was at odds with teacher judgments for other students. Along with observations about student characteristics that affect assessment accuracy, recommendations from the study include suggestions for working with teachers who are asked to make judgments about test accuracy and procedures for confirming teacher judgments.

Identifying Less Accurately Measured Students

Test scores provide imperfect estimates of students' knowledge and skills. Statistics such as reliability coefficients, standard errors of measurement, and confidence intervals reflect random measurement error that affects all students. Any single test score may over- or underestimate a student's knowledge and skills. In a sense, then, most students are to some extent inaccurately measured. But some students are less accurately measured than others. In addition to random measurement error that clouds the picture for all students, some students' test scores also include systematic error that further distorts the picture of their knowledge and skills.

Some systematic error produces test results that give a misleadingly high impression of a student's knowledge and skills; some produces a misleadingly low impression. Some systematic error affects many test takers uniformly; some affects certain test takers differentially. For example, uniformly inflated results may be produced by a test that included poorly designed items that gave away the answers. Differentially inflated results may occur when some test takers cheat. Uniformly suppressed results can arise from scoring errors that mark correct responses as incorrect. Differentially suppressed results can be caused by personal characteristics that impede test performance more for some students than for others. Common examples of such characteristics are students who suffer from test anxiety or students who are penalized for weaker reading skills on tests that are not supposed to be measures of reading (AERA, APA, & NCME, 1999, p. 10; Haladyna & Downing, 2004, p. 19).

There are efforts to improve measurement accuracy targeting all of these kinds of random and systematic error. This paper focuses on systematic error that produces differentially low estimates of students' knowledge and skills. A huge amount of work already has been done in

this area. For example, all of the efforts to identify and eliminate judgmental and statistical bias from tests would fit here. Within the past two to three decades there has been a strong push to remove differential distortions in assessment results that are attributable to disabilities that some students have (McDonnell & McLaughlin, 1997; President's Commission on Excellence in Special Education, 2002). A lot of work on improving assessment accuracy for students with disabilities has looked at using accommodations to lessen the impact of a disability on test performance (Sireci, Scarpati, & Li, 2005; Thurlow, Thompson, & Lazarus, 2006; Thurlow, Ysseldyke, & Silverstein, 1995). More recent work introduced the notion of applying principles of universal design to assessments so there would be less need to use accommodations (Johnstone, Thompson, Bottsford-Miller, & Thurlow, 2008; Thompson, Thurlow, & Malouf, 2004; Thurlow, Johnstone, & Geller, 2008).

A popular theme in accommodations and principles for universally designed assessments is reducing the impact that reading ability has on non-reading tests (Abedi, 2006; Dolan & Hall, 2001). This theme is popular in part because many different kinds of disabilities are connected in one way or another to difficulty with reading (Thurlow, Moen, Liu, Scullin, Hausmann, & Shyyan, 2009). This creates a challenge for tests that are designed to measure reading. Fewer accommodations and universal design principles are available to be applied to the very test whose results seem most likely to be affected by many students' disabilities.

In response to this quandary, the federal government sponsored a set of national projects to investigate the development of accessible reading assessments that would permit more accurate estimates of the reading knowledge and skills of students who have disabilities. One of set of collaborators in this project, the Partnership for Accessible Reading Assessment (PARA), took as a starting point the agnostic stance that current knowledge about how disabilities hinder

reading test performance and about which assessment practices would best minimize these hindrances was insufficient to formulate a solution. We needed to learn more. Moving out from that starting point, several teams within PARA pursued different lines of investigation.

The investigation reported here sought to find a way to identify students who seemed likely to benefit most from accessible reading assessment. Our reasoning was that we cannot assume that all students with disabilities would benefit from accessible reading assessment. Students with disabilities are in the same boat as students without disabilities; tests should distinguish students who have the requisite skills from those who do not. Accessible reading assessments are supposed to help provide a clearer picture of the extent to which students have received and benefited from grade level reading instruction, not merely increase test scores for all students with disabilities. Accordingly, work on developing and implementing accessible reading assessment needs to focus on students who have attained grade level reading knowledge and skills but who have difficulty showing that on typical reading tests. A key challenge in this work, then, is being able to identify students who have more reading skills than the typical reading test would suggest.

We decided to see how feasible it was to use teachers to identify such students. The literature is replete with teachers' objections that large scale tests do not adequately measure what teachers teach (Abrams, Pedulla, & Maduas, 2003; Cizek, 2001; Popham, 2007; Prime Numbers, 2006). If these objections have any substance, presumably teachers see evidence of students' knowledge and skills that large scale tests miss. A series of informal conversations we initiated with teachers confirmed that each teacher was quickly able to generate examples of students who the teacher believed had reading abilities that the typical reading test would underestimate. If teachers could in fact identify likely candidates for accessible reading

assessment, using teacher judgment could help improve research on and implementation of accessible reading assessment, even if follow up individualized diagnostic assessments had to be added to bolster teacher judgments.

A review of work that others have done that can be related to teacher judgment offers mixed support for this enterprise. Many assessment specialists are likely familiar with the debate that has continued since Meehl's (1954) book pitted clinical versus statistical prediction in the field of counseling psychology. In a meta analysis of over half a century of research on this issue, Ægisdóttir and colleagues (2006) observed that "arguments in favor of the small, but reliable, edge of statistical prediction techniques are strong" (p. 373). This does not argue, as might be supposed, that clinical judgment is ineffective, merely that often statistical procedures can be developed that are more effective. When effective statistical procedures are not readily available, clinical judgment seems to be a reasonable option.

A parallel in the field of industrial and organizational psychology might be the use of human judgment in assessment centers for the selection and development of managers and executives. An online publication by the American Psychological Association, *Psychology Matters*, said in 2008 that "standardized tests have not been widely accepted in selecting and evaluating managers and executives, in part because of the seeming gap between the simple skills measured by tests and the complex skills (especially people-oriented skills) believed to be critical for managers and executives" (paragraph 4). Consequently, the publication goes on to say, assessment centers using human evaluators are often seen as the method of choice for these kinds of tasks.

These examples from disciplines outside the field of education illustrate some of the complexities of determining how much reliance to place on human judgment. Within education,

there is a long history of using accumulated teacher judgment in the form of grade point average or high school rank to predict success in college (Willingham & Breland, 1982). Research on using tests for predicting college success has typically assumed such teacher-based indices as foundational and attempted to show that tests added a worthwhile incremental improvement in prediction over and above such indices (Noble & Sawyer, 2002; Pike & Saupe, 2002; Price & Kim, 1976). As with clinical prediction and assessment centers, a complex set of factors affect academic success (Willingham, 1985). An argument can be made that teachers who spend many hours during the course of a year with the same students have more opportunity to see students' skills than assessment center assessors or clinicians have to observe their clients' characteristics. On the other hand, from issues of grade inflation to concerns about subjectivity including outright favoritism, skepticism about the credibility of teacher-based evaluations abounds (Bradley & Calvin, 1998; Cizek, Fitzgerald, & Rachor, 1995; Ornstein, 1994).

Our examination of literature relating teacher judgment to test performance found that earlier research, much of it reviewed by Hoge and Coladarci (1989) and Perry and Meisels (1996), tended to support teacher judgments of student achievement. More recent studies seem to have highlighted questions about teacher judgment. For example, although several studies using curriculum-based measurement (CBM) found moderate to high correlations between teacher judgment and measures of reading fluency, Feinberg and Shapiro (2003) suggested that the correlational data may be masking some important issues such as a tendency for teachers to overestimate students' performance when the reading materials were below- or at-grade level (Eckert, Dunn, Codding, Begeny, & Kleinmann, 2006).

In considering issues that come closer to the ability of teachers to make judgments about the test performance of students with disabilities, Coladarci (1986) and Demaray and Elliott

(1998) report studies that show teachers were somewhat less accurate when judging the achievement level of lower-achieving students than when judging average- to high-achieving students. Coladarci (1986) worried that results pointed "tentatively to the disturbing implication that students who perhaps are in the greatest need of accurate appraisals made by the teacher in the interactive context are precisely those students whose cognition has a greater chance of being misjudged" (p. 145).

Moving beyond general estimates of student performance, we find greater difficulties when teachers are asked to make finer distinctions. Gresham, MacMillan, and Bocian (1997), for example, found that although teachers could identify which students were at risk of performing poorly on a test, they were not successful in distinguishing among three types of at-risk students: those who were considered to have a learning disability, those who were considered to have low cognitive ability, and those who were simply low achieving. Similarly, when Bailey and Drummond (2006) asked teachers to nominate kindergarten and first-grade students whom they believed to be struggling readers, the teachers succeeded in nominating students who scored below their norm-group on the standardized measures, but the teachers did not always capture the specific areas of weakness that many students showed on the standardized measures including comprehension, vocabulary, and phonological awareness deficits.

Studies of teachers' success in determining which students would benefit most from which accommodations also cast doubt on teachers' abilities to make distinctions finer than whether students' performance will be high or low. Although teachers should be knowledgeable about students, about their access to the curriculum, and about what accommodations may be most useful to them (DeStefano, Shriner, & Lloyd, 2001) and although teachers frequently play a central role in determining appropriate accommodations, Fuchs, Fuchs, and Capizzi (2005)

concluded that teacher decisions regarding accommodations are "often subjective and ineffective." (p. 7).

Part of the problem seems to be that teachers' knowledge of allowable accommodations has been questionable enough to put validity and reliability at risk. Hollenbeck, Tindal, and Almond (1998) found that large variability exists regarding what teachers perceive as being appropriate accommodations, that teachers have made inconsistent use of accommodations, and they have sometimes shown preference for particular accommodations regardless of state guidelines. Fuchs and Fuchs (2001) found that some teachers provided the same accommodations to most students regardless of students' individual needs and that other teachers sometimes grant accommodations to students who do not benefit from them. In one study where teachers were found not to be effective in recommending which students would benefit from having a read aloud accommodation for a math test, teachers' judgments were not more accurate than chance (Helwig & Tindal, 2003).

Looking more specifically at accommodations with reading tests, one study by Fuchs, Fuchs, Eaton, Hamlett, Binkley, and Crouch (2000) found that teacher judgments provided many more accommodations than did data-based standards and that the accommodations that teachers provided did not produce a greater differential boost for students who had the accommodations than for those students who did not have them. Effect sizes for the accommodations that teachers awarded were small, ranging from -.07 to .06.

Despite the doubts regarding teacher judgment that some of the studies raise, there are three main reasons we have persisted in examining whether teachers might be used to identify students who may be less accurately assessed. First, many of the studies that call into doubt teachers' judgment merely show some discrepancy between teacher judgment and some test

result. Rarely is evidence offered that the reason for the discrepancy is error in teacher judgment. Bailey and Drummond (2006), for example, explicitly point out that they did not seek to determine which measure was correct but merely observe that there was substantial misalignment. It could well be that discrepancies are sometimes due to limitations of the test and that the teacher judgment is taking into account information that the test lacks. Many writers have discussed differences between what teachers pick up on in classroom evaluations and what gets measured in large scale tests (Shepard, 2000; Brookhart, 2003; Moss, 2003; National Research Council, 2003). This is in fact a key premise of our study; in cases where the test would produce a misleading picture of certain students, teacher judgment should diverge from test results.

The second reason for pursuing the potential use of teacher judgment is that some weaknesses in teacher judgment may be due to lack of information or misaligned perspectives that can be improved through training or support tools. DeStefano et al. (2001), for example, reported that after teachers went through systematic training, testing accommodations and instructional accommodations were more similar in number and type, students were more likely to receive accommodations on an individual basis, there was a reduction in accommodations for target skills (such as a reading accommodation on a reading test), and teachers felt greater confidence when selecting accommodations. Efforts currently underway by states to develop materials and provide training to help teachers make better accommodations decisions (for example, Minnesota Department of Education, 2008; Washington Office of Superintendent of Public Instruction, 2008) are premised on the assumption that teachers who are given the right training and tools can learn to make better accommodations decisions.

Finally, substantial benefits can accrue from working with teachers' judgments. If they provide good information, using teacher judgments could be less expensive, obtrusive, and time consuming (Perry & Meisels, 1996) than other methods of assessing student achievement levels. They could provide the deeper insight into student performance that some are concerned is missing from typical tests (e.g., Abrams, Pedulla, & Madaus, 2003). And teacher judgment already greatly affects students' lives through the feedback teachers give students (Black & Wiliam, 1998), and the impact that course grades have on students' options (Willingham & Breland, 1982), so any work that helps improve teacher judgment is likely to benefit students.

This paper describes a small-scale study that was designed to provide a preliminary look at the feasibility and advisability of using teacher judgment as part of the procedure for identifying students at most risk of being inaccurately measured by typical annual large-scale reading tests. The study used a close examination of a limited number of cases to shed light on several questions. The four main questions explored in this study were:

1. Are teachers able to distinguish students who are likely to be less accurately measured from other students?
2. Can teachers distinguish different reasons why a student is likely to be less accurately measured?
3. How well can the procedures employed in this study be used to evaluate teachers' judgments about likely assessment accuracy?
4. What are the distinguishing characteristics of students who are likely to be less accurately measured?

Throughout the rest of this paper, the acronym LAMS will stand for less accurately measured students.

Methods

We started by developing a questionnaire suitable for use in a large-scale study. We drew on what we learned from the literatures on reading, assessment, and disabilities and from literacy experts working with PARA to create a detailed questionnaire that was designed to permit asking many teachers to rate students on a host of variables thought to affect reading test performance. The questionnaire and plans for its use were distributed to 18 nationally known experts in reading, assessment, and disabilities. Their feedback provided support for the general goal of using teachers to identify LAMS. It also endorsed, sometimes enthusiastically, many of the details of the questionnaire. But some experts raised concerns that resonated with some of our own reservations. In particular, we came to agree with those experts who suggested that at this early stage in this investigation we might learn more by examining in depth what a few teachers and students thought than we would by having many teachers respond superficially to a long questionnaire. As a result, we developed the more open-ended questionnaire and procedures used in this study that are suitable for working with a small number of teachers and students.

Procedures

The present study had four main data collection steps:

1. Teachers completed a paper-and-pencil questionnaire nominating students they thought would be less accurately measured by typical large-scale annual state reading tests, and described why they thought a student would be less accurately measured.
2. Researchers interviewed teachers to clarify and confirm the information provided in the questionnaire and to review any evidence teachers could supply to support their assertions about the likelihood of measurement inaccuracy.

3. Researchers interviewed students to establish rapport and obtain students' attitudes and opinions about reading and assessments.
4. Researchers administered brief reading assessments to students that differed according to the explanation for why the student was thought to be less accurately measured.

Audio recordings were made of all interview and assessment sessions.

The study was run in two phases. The first phase was completed during the spring and summer of 2006 and the second phase in the spring and summer of 2007. A questionnaire that teachers completed remained consistent across both phases of the study, so questionnaire data from the two phases have been combined. Procedures for teacher and student interviews and student assessments changed enough from the first phase to the second by making the interviews more structured and refining assessment protocols so that only data from the second phase are reported here.

Tools

A paper-and-pencil questionnaire was used to ask teachers to nominate students they thought were inaccurately measured by large scale reading tests and to rate the degree of inaccuracy. The questionnaire described four reasons a student's reading test score might give an inaccurate picture of his or her reading skills. Using these four reasons or adding a reason of their own, teachers were asked to associate each nominated student with a reason that explained why the student would be inaccurately measured. Teachers were also directed to give a more complete description in their own words of why the reading test would misrepresent each student's reading skills.

The four reasons supplied on the questionnaire were:

1. Fluency limitations obscure comprehension skills.

2. Comprehension limitations obscure other reading skills.
3. Weakness in tested reading hides non-tested reading strengths
4. Responds poorly to standardized testing circumstances or materials.

A fifth option listed as "Other reasons" allowed teachers to add their own reasons.

The questionnaire gave explanations and examples showing how each reason might result in tests giving an inaccurate impression of students' reading skills. For example, the questionnaire suggested that there may be some students who have learned much about comprehending what they read but they have weak fluency skills that get in the way of showing their comprehension skills on typical reading tests. Disentangling reading skills so that weakness in one area does not prevent seeing strengths in another might be compared to the situation in mathematics tests where computation skills are sometimes separated from problem solving skills so that students with weak computational skills are still able to show their problem solving skills. Another example described aspects of reading that teachers may consider important and spend time working on that are not covered by typical annual tests. Such things could include developing positive attitudes toward reading, habits of independent reading, intentional choice making, a willingness to grapple with challenging materials, and skills in using the internet.

A structured oral interview for teachers had five main questions. The questions encouraged the teacher to: (1) provide a more in depth description of the student and why the reading test misrepresented the student's reading skills, (2) discuss and review evidence that could document the teacher's description, (3) rate the impact that several variables might have on the student's test performance, (4) describe his or her level of confidence in the description of the student and in the particular reason given for why a student would be misrepresented by reading

test scores, and (5) add any other comments he or she wanted to give about reading tests or related issues.

Evidence provided by teachers during the second interview question included materials such as test scores, classroom work, running records, or anecdotal observations by the teacher. During the third interview question, teachers were asked to use a five point scale to rate the impact of these seven variables: fluency limitations, comprehension limitations, low motivation for the test, keeping attention focused on the test, getting worn out by the test, anxiety, and other.

A structured oral interview for students had six questions intended to establish rapport with the student, get a little better picture of who the student is as a person, and learn about his or her attitudes and experiences with regard to reading and reading tests. As part of this, students were asked to share their own opinions on the extent to which large scale reading tests show how well they can read. The last question measured students' opinions about how much certain changes to reading tests would affect their performance on the tests. Students used a five-point scale to rate the likely impact of these changes: (a) having shorter reading passages; (b) having more interesting passages; (c) taking the test on a computer instead of paper and pencil, (d) having the entire test read out loud by a tape, CD or MP3 player; (e) using a computer that let you choose words to have pronounced or explained while you read the printed text; and (f) other ideas you have.

Two assessment activities were used for each student. First, all students completed three curriculum-based measurement reading (CBM-R) probes. CBM-R is a quick assessment task targeting oral reading fluency in which students read grade-level narrative text for a duration of one minute (Shinn & Shinn, 2002). We followed typical CBM-R administration by marking the number of words read incorrectly, and subtracting that amount from the total number of words

read. The median score was recorded; this was selected to avoid outlier effects. Each median score was compared to AIMSWeb nationally-normed mean and standard deviation words per minute for the appropriate grade. AIMSWeb means were used because the large sample size of the norming population was unlikely to be significantly affected by outliers. The CBM-R probes were taken from released statewide reading tests. Information about normative data and the technical characteristics of this assessment can be found in Howe and Shinn (2002).

The second assessment activity varied depending on what the teacher had identified as the student's primary barrier to accurate test scores. For students placed by teachers in the first barrier category (having fluency limitations that obscure measurements of comprehension), the second activity comprised an approximately 250 word portion of a reading passage read aloud on tape to the student. Students were able to replay the selection as many times as needed until they thought that they understood the passage well. Then students orally retold as much of the passage as they could remember. Retellings were transcribed and then scored according to how many main ideas, sub-ideas, and details were recalled.

For students from the second barrier category (comprehension limitations obscure other reading skills), the second assessment activity involved having students read on their own an approximately 250 word portion of a reading passage. They then immediately orally retold as much of the passage (both main idea and details) as they remembered. Each retelling was transcribed and then scored according to how many main ideas, sub-ideas, and details were recalled.

For the students placed in the third barrier category (students who have strengths outside of what most reading tests cover), the second assessment activity entailed reading one grade

level reading passage and answering corresponding multiple choice questions. Students were offered a choice of reading the passage silently or hearing it read out loud on tape.

Students in the fourth barrier category (students who respond poorly to testing circumstances) read one grade level reading passage and answered the corresponding multiple choice questions. During this testing, students were encouraged to "think aloud" about difficulties experienced with the text and the items or suggestions for improvement.

Participants

We recruited participants from ten elementary and middle schools in urban, suburban, and rural locations in two states. Thirteen teachers completed questionnaires during the first phase of the study and eight during the second phase for a combined total of twenty-one teachers. The teachers taught grades ranging from 4 through 8 in both general and special education. Teachers in the first phase identified 57 students as less accurately measured and the teachers in the second phase identified 20 such students. We met with two teachers and six students in the first phase. During the second phase, we met with eight teachers and twenty students. All of the teachers and students who were interviewed were from a single midwestern state.

Analysis

Quantitative data were tabulated from the nomination questionnaire and from the teacher and student structured interviews. Results from these tabulations are presented as descriptive statistics with cautions about over-interpretation because of the small number of cases. Qualitative analyses integrated observational information gathered during the interviews and the assessments with data obtained from the questionnaire, the brief assessments and teacher-provided evidence. In a series of weekly meetings that spanned three months, four of the authors

met to review this information. We worked to reach consensus on the extent to which information from separate sources converged to support conclusions. When consensus was not easily reached from the summary information, more detailed examination was undertaken of original source materials, including transcripts of interview and assessment sessions. Situations where we could not eventually reach consensus led us to conclude that a determination could not be made. The primary determinations sought were whether evidence supported: (1) the teacher's assertion that a student is likely to be less accurately measured, and (2) the teacher's assertions about why the student is likely to be less accurately measured.

Results

Teacher Judgments

During both phases of the study combined, 21 teachers from 10 sites submitted questionnaire responses nominating a total of 77 students as less accurately measured students (LAMS). During the second phase of the study, data presented in this paper were obtained from 8 teachers and 20 students who participated in the structured interviews and brief assessment sessions.

Data from the nomination questionnaire are shown in Table 1. Note that the number of designations adds up to more than the 77 nominated students. This indicates that some teachers placed some students under more than one explanation for why reading test scores would be misleading. The questionnaire instructions had intentionally been left ambiguous on this point so that teachers could use more than one explanation per student if they chose to do that. Teachers made greatest use of the explanations "fluency limitations obscure comprehension skills" (30%) and "responds poorly to standardized testing circumstances or materials" (29%). They made less use of "some comprehension limitations obscure other skills" (20%) and "has strengths outside

of what most reading tests cover" (17%). Relatively little use was made of "other reasons" (5%). The mean rating of how much a reading test would distort the picture of these students' reading skills was 3.89 with a standard deviation of .89. This was on a 5 point rating scale in which 1 signified that a test would be *a little off* and 5 that it would be *way off*.

Table 1

Reasons for Less Accurate Measurement

Reasons for Identifying Students as LAMS	Count	Percent
Fluency limitations obscure comprehension skills	32	30%
Some comprehension limitations obscure other skills	22	20%
Has strengths outside of what most reading tests cover	18	17%
Responds poorly to testing circumstances or materials	31	29%
Other	5	5%

Note. The entries total greater than 77 students because students could be assigned to more than one reason category. Percentages are based on the total counts (n=108) rather than the total number of students.

Data from one of the questions in the teacher interview explicitly invited teachers to apply more than one explanation to each student by asking them to use a 1 to 5 rating scale to indicate how much impact several variables had on each student's test performance. Table 2 shows results from this question. Bear in mind that these interview data are based on only eight teachers rating only twenty students. The results for this group of teachers and students suggest patterns worth discussing that would be good to confirm with a larger sample.

Table 2

Teacher Ratings of Barriers to Students' Performance

Barrier	Hardly At All	A Little	Some	Quite a Bit	A Lot	Blank	Mean
Fluency limitations	3	2	4	7	4	0	3.35
	15%	10%	20%	35%	20%	0%	
Comprehension limitations	0	2	7	7	4	0	3.65
	0%	10%	35%	35%	20%	0%	
Low motivation for the test	8	3	4	1	4	0	2.50
	40%	15%	20%	5%	20%	0%	
Keeping attention focused on the test	4	7	5	2	2	0	2.55
	20%	35%	25%	10%	10%	0%	
Getting worn out by the test	5	6	4	3	2	0	2.55
	25%	30%	20%	15%	10%	0%	
Anxiety	6	6	5	0	3	0	2.40
	30%	30%	25%	0%	15%	0%	
Other	0	0	0	2	8	10	4.80

The two factors rated as having the largest impact on a student's reading test performance, aside from the teachers' "other" explanations to be discussed below, are comprehension limitations and fluency limitations. The means on a 5-point scale for these two factors are 3.65 and 3.35 respectively. For both of these factors, over half of the students were rated in the top two categories indicating that these factors affected them *quite a bit* or *a lot*. All

of the students were described as being at least *a little* affected by comprehension limitations. But for fluency limitations, three students were rated in the lowest category as being *hardly at all* affected. For the rest of the provided explanations, over half of the students were rated in the lowest two categories as *hardly at all* or only *a little* affected. Yet there were some students for each of these variables that received the highest possible rating indicating that some students were affected *a lot* by these variables. This pattern of ratings indicates some commonality in that all of the nominated students' reading test scores are affected by multiple factors and in particular all are affected by comprehension limitations. At the same time, there is considerable diversity in that each of the listed factors affects some students only *a little* and other students *a lot*. The diversity found among this small number of students is perhaps best seen by looking at the descriptions of individual students referenced in the integrative analysis section found later in this document.

When missing values are left out, the highest mean rating (4.80) was for the "other" explanations that teachers supplied for ten students. Some of the explanations that teachers added here seemed to us closely related to explanations we had offered such as motivation and anxiety. Several other explanations could have fit under the "testing circumstances or materials" used in the nominating questionnaire but that explanation had not be repeated in the interview as an option. In particular, teachers said for several students that test materials that relied on multiple choice tests or other written responses disadvantaged these students who performed better with oral responding. A couple of other teacher generated explanations delved into issues such as background and family expectations that we judged had more to do with why a student might not have developed effective reading skills than with why a test might obscure effective reading skills.

Teachers provided a variety of evidence for their descriptions of the students they nominated as LAMS. They shared samples of class work, recent standardized test scores, reports of students' participation in class literature conversations, and informal reading assessments. The strength of the evidence that students would be inaccurately assessed varied by teacher and student. In some cases, the nature of the student's characteristics limited the potential evidence for the teacher to provide. For example, it was easier for a teacher to provide tangible evidence of low fluency than evidence of responding poorly to testing situations. Some teachers were more thorough than others, providing a greater quantity of evidence or evidence with greater depth in detail. Additionally, some teachers provided evidence that was specific to each student, whereas other teachers provided the same evidence (same work samples or test information) for all students nominated as LAMS.

Student Opinions

Of the 77 students nominated as LAMS in the first phase of this study, 20 participated in the second phase, which included structured interviews and brief assessment sessions. Interviews with students identified as LAMS provided information on student attitudes towards reading, opinions about traditional reading tests, and ideas about assessment practices that might improve their performance. Students used the 5-point rating scale shown in Table 3 to answer all of these questions. Only the results for ideas about helpful assessment practices are presented in Table 3.

This paragraph describes results for student attitudes and opinions that are not in Table 3. All students reported enjoying reading at least *some*, and about a third of the group indicated enjoying reading *quite a bit* to *a lot*. When asked about how difficult reading is, responses were more varied and were distributed relatively evenly from *hardly at all* to *quite a bit*; no students rated the difficulty of reading with a 5 or *a lot*. When asked their opinions about how much

reading tests show their reading skills, 10% of the students said *a lot*, 70% responded with *some* to *quite a bit*, 5% said *a little*, and the remaining portion did not have a specific response. Most students seemed to have difficulty thinking about the role and quality of tests from a psychometric viewpoint, and some likely did not understand the question very well.

Table 3

Student Ratings of Barriers to Students' Performance

Assessment Practice	Hardly At All	A Little	Some	Quite a Bit	A Lot	Blank	Mean
Shorter reading passages	0	2	5	8	3	2	3.67
	0%	10%	25%	40%	10%	15%	
More interesting reading passages	0	3	1	5	9	2	4.11
	0%	15%	5%	25%	45%	10%	
Computer instead of paper & pencil	2	3	2	4	6	3	3.53
	10%	15%	10%	20%	30%	15%	
Entire test read aloud by CD, etc.	4	1	7	3	3	2	3.00
	20%	5%	35%	15%	15%	10%	
Computer pronounces/explains	0	0	3	6	9	2	4.33
words you pick	0%	0%	15%	30%	45%	10%	
Other ideas you have	1	1	0	2	6	10	4.43
	5%	5%	0%	10%	30%	50%	

Table 3 shows student attitudes toward assessment practices that might help them perform better on a reading test. Note that we decided we could not ask students to distinguish

general performance improvement from increasing measurement accuracy, so we merely asked what they thought would improve their test performance. Also, bear in mind that these tabulation results need to be taken cautiously because they are based on the answers of only 20 students. The methods students were asked to consider included: shorter reading passages, more interesting passages, computerized test administration, test read out loud electronically, and assistive technology to aid decoding. The highest rated practice aside from "other" was "computer pronounces or explains words that you pick." This practice had a mean rating of 4.33; all students who answered this question rated it as likely helping their performance from *some* to *a lot*. All students rated more interesting or shorter passages as being likely to help them at least *a little*. Both the options to use a computer and to have the entire test read aloud had some students say that it would help *hardly at all*. Several students commented with regard to having the test read aloud to them that they would prefer to have control over the pace of reading.

Integrative Analyses

To evaluate the assertions teachers made in the paper-and-pencil questionnaire and in the structured teacher interviews about whether students were likely to be inaccurately assessed and why, we compared those assertions with other information. The comparison information came from the evidence teachers provided during interviews to support their judgments, students' statements during structured student interviews, the results of the brief student assessments, and observations we made during the interviews and assessments.

Generalizing information from these analyses is challenging because there were a small number of cases and each teacher-student pair we examined was unique. There were variations in student reading strengths and weaknesses, the ways teachers perceived students' strengths and weaknesses, and the ways student's skills relate to standardized reading tests. To get more in-

depth information about individual students and teachers than is reported here, readers can consult a more anecdotally oriented paper by Moen, Liu, Thurlow, Lekwa, & Scullin (in process).

For 3 of the 20 students, all of the available information clearly converged to support the teachers' assertions about whether a student would likely be inaccurately assessed and why. One of these students, as reported by the teacher and supported by assessment data, had extraordinarily low decoding and fluency that was out of proportion to his other skills and abilities. For a second student, observations and assessments confirmed the teacher's description of very slow processing that extended beyond reading decoding and fluency. For a third student, although he had below average reading skills, those skills that he did have would be nearly invisible in a typical annual reading test because the student so readily went off task. Information from the student's interview, the brief assessment session, and the researcher's observation of the student all converged to support the teacher's description of this student.

In contrast to this, we found that 3 of the 20 students that teachers had nominated as LAMS clearly did not fit the model used in this research. In one case the teacher indicated that the student was performing below grade level so the typical grade-level test would not be able to measure the student's level of achievement. In another case the teacher talked about a student who had more potential to learn than the student was using. A third student was nominated for low comprehension that the teacher suspected stemmed from an under-stimulating environment during early development. In all of these situations, it appeared that the teachers were describing learning problems rather than assessment problems. For these students, low performance on a test would accurately communicate the message that they could not do what the standards required.

Results for the remaining students fell somewhere between these two extremes. For 11 students, we were inclined to agree with the teachers' assertions that the student would likely be less accurately measured than other students, but the evidence to support teachers' assertions about the reasons for measurement inaccuracy was weak, ambiguous, or missing. For example, anxiety was identified as a factor that affects the performance of several students; in a few cases it was cited as the primary inhibiting factor. Combining our observations of student demeanor with statements students made about anxiety and evidence teachers presented of student performance under various circumstances let us evaluate whether we had at least some support for some teachers' assertions about anxiety. But because our student interview and assessment procedures were designed to minimize student stress, we lacked the ability to observe a high level of anxiety during testing that would have provide more solid confirmation of teacher assertions.

Similarly, teacher nominations of students as LAMS because they are hampered by particular assessment methodologies that require writing and multiple choice instead of oral examination and open ended responses and that prohibit assistive devices such as highlighters and placeholders, seemed compatible with most of the information we obtained, but our short assessments did not have the power to confirm or disconfirm these teacher assertions as clearly as we would have liked. For students who were nominated because some reading weaknesses hid other reading skills but who did not have the glaring disparities in abilities shown in the three most clear cut cases, there were less sharp differences between, for example, measures of reading fluency and reading comprehension.

Given the limitations of our confirmatory tools, we were disposed to give teachers' judgments the benefit of the doubt when we could produce neither confirmatory nor dis-

confirmatory evidence. For three students, we did question the teachers' assertions because something in the comparison data seemed to be in conflict with the assertions. For example, one teacher described a student as responding poorly to standardized reading tests but the student showed good test-taking skills during the study. Another teacher described a student as having good oral comprehension that does not show up when the student takes a written test, but the assessment evidence we collected did not support an assertion that the student comprehends well.

In sum then, we tended to agree with teachers' basic assertions that identified students who would likely be less accurately assessed for 14 out of the 20 students. For 3 of these we judged that there was strong supporting evidence and for the other 11 we judged that the evidence tilted in support of teachers' enough that we would be reluctant to challenge their assertion. We rejected teachers' judgments about three students and we were skeptical of their assertions for three other students.

Discussion

The clearest instances of LAMS were a few students that teachers identified who have one notable characteristic that radically undermines performance on a reading test that can be documented by classroom evidence and confirmed by a brief student interview and assessment. In this study, teachers were able to provide consistent descriptions of the assessment challenges for showing the reading skills of students who process everything very slowly, have extraordinary decoding difficulties, or have exceptional difficulty staying on task. Information we obtained by interviewing and assessing the students corroborated teacher-provided evidence to confirm these characteristics and their impact on assessment performance.

The classic example of test anxiety was cited for several students but most often the test anxiety was mixed with other characteristics that affect reading test performance. This made it harder for teachers to describe and for us to confirm the nature of any measurement inaccuracy.

We learned during the first phase of the study about one of the challenges in describing students who have multiple factors in play when we saw that teachers sometimes wound up emphasizing different characteristics by the end of the teacher interview than they had initially given on the paper-and-pencil questionnaire. We considered whether this change from initial rating to statements during the interview should be treated as contamination by the researcher and whether we should stay with a student assessment protocol based on the teacher's initial description.

Instead, we decided to view the teacher interview as a legitimate part of teacher training that gave the teacher more insight into the research questions and in how to more effectively describe a LAMS. Accordingly, we made the explicit decision that the teacher's final description of the student would be used in determining the student assessment protocol. This experience has lead us to wonder to what extent the results of studies that find weaknesses in teacher judgment are attributable to requiring teachers to work within a framework that is initially alien to them and whether more time spent working with teachers to arrive at a shared framework before requiring them to make judgments might produce a more positive view of teacher judgment.

In documenting test anxiety, we had to lean heavily on the evidence teachers provided, students' answers during interviews, and our observations of students' demeanor because our short assessment protocols did not induce the kind of stress that would have helped us directly observe the impact of test anxiety on test performance. Although not as strong a form of confirmation as we would have liked, we concluded that our observations provided sufficient basis to judge whether our data were at least consistent with the teacher's assertions of high test anxiety. If additional research is done that focuses more extensively on test anxiety, a transactional process model such as that described by Spielberger and Vagg (1995) might be adopted that would lead to the use of tools such as Spielberger's state-trait and test anxiety

inventories to get a better understanding of the nature of the impact of anxiety on students' reading test performance. Spielberger counts emotionality that impairs functioning on complex tasks and worry that has unproductive cognitions intruding and distracting from the task at hand as two distinct ways that anxiety can affect test performance.

Another kind of student characteristic that teachers repeatedly cited fit Campbell and Fiske's (1959) description of another source of measurement inaccuracy known as "method variance." Campbell and Fiske argued that convergent and discriminate evidence for validity should be examined by considering how strongly measures of different traits that used the same methods correlated compared to how strongly measures of the same trait obtained with different methods correlated. In this study, teachers asserted that the method of using multiple choice tests gave a distorted picture of students' comprehension skills. They said that a very different picture of some students' comprehension skills would emerge than that portrayed by multiple choice tests if these students were allowed to respond using constructed responses in writing or orally, especially in the context of a group discussion. Examining these students' performance on multiple choice tests in other content areas such as math and more extensive examination of their written and oral performance than our brief assessment protocols permitted could help confirm or disconfirm these teacher claims.

A particular challenge for us in evaluating claims regarding test anxiety and method variance was that sometimes teachers identified students who already performed well on tests but the teacher asserted that the skills they exhibited in non-test situations would be substantially higher. Although misleadingly low test results for higher performing students was not part of our original LAMS target, we concluded that for purposes of evaluating teachers' success in identifying LAMS these cases could be considered as possible LAMS. Especially in those cases

where students were able to perform well on our short assessments, our ability to confirm teachers' assertions was limited to evidence teachers provided, students' interview statements, and our observations of student behavior. Further research designed to look more closely at higher performing students could strengthen observations about how well teachers can identify more students whose test performance might be raised.

Test anxiety and method variance are sources of measurement inaccuracy that fit the classic model of construct irrelevant variance presented by authors such as Haladyna and Downing (2004). The ideal circumstance is when a characteristic can be isolated that impairs test performance that is completely unrelated to the construct being measured. One of our three clearest cases of LAMS, a student with great difficulty with staying on task, probably fits this model the best. The main issues to be resolved for this kind of characteristic are finding assessment practices that will reduce its impact and determining that the benefits of implementing those practices outweigh the costs. The situation becomes more complex when the characteristics that hinder test performance are entangled with or are themselves characteristics that do have some relevance to the construct that is being measured. It is hard to argue that the characteristics of two of the three students we counted as a clear instance of a LAMS, very slow information processing and exceptionally low decoding skills, are unrelated to the construct of reading. And the test anxiety and method variance characteristics we observed were usually mixed with some weaknesses in decoding, fluency, and comprehension. A key question is whether the notion of construct relevance will be treated as an absolute or a matter of degree. People who want to evaluate the reading skills of students who have characteristics such as these need to decide whether they want to be able to see what reading skills these students can demonstrate when these impediments to reading are discounted or removed.

Conclusion

We found evidence that there are some less accurately measured students and that teachers can be used to identify and describe at least some of these students. We also saw evidence that teachers can probably do this task better when they have more background information or training. If the findings from this study could be extended and confirmed with a larger number of teachers and students, perhaps a simple explanatory framework could be developed along the lines of the Valencia and Buly (2004) typology that would include our emphases on assessment and disability factors. A simple framework clearly explained could help teachers see more fully from the outset how to do this novel and complex task.

Procedures for confirming teacher judgments could be enhanced substantially if they could be connected to opportunities to observe students taking full-fledged reading tests. An example of something along these lines is research completed by the New England Compact (NECAP) that used teacher judgment in conjunction with a state administered math test to identify what they described as gap I and gap II students (New England Compact, 2007). If the NECAP research design were adjusted so that teacher judgments about students and explanations for anticipated discrepancies between teacher evaluations of students and test scores were examined before rather than after the test, researchers could observe the activity of students during the test who had been identified as being a gap or less accurately measured student. On an even larger scale, if performance assessments that rely on teacher judgment such as those recently advocated by Wood, Darling-Hammond, Neill, and Roschewski (2007) were regularly incorporated into assessment systems to complement the information from standardized tests, data from these two multiple measures of student achievement could be compared to evaluate and improve both teacher judgment and the accuracy of standardized tests.

Finally, the pay off for identifying less accurately measured students will come from developing and implementing assessment practices that improve reading test results for these students. Our colleagues in the PARA team and other colleagues in similar projects are working hard to develop such practices. This work involves complex issues such as determining the boundaries of construct irrelevant variance and the impact of practical cost/benefit calculations that the field of assessment needs to resolve so innovative solutions to making accessible reading assessments can be developed and implemented.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., Rush, J. D.(2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*, 341-382.

Abedi, J. (2006). Language issues in item-development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377-398). Mahwah, NJ: Erlbaum.

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory Into Practice, 42*(1), 18-29.

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

American Psychological Association. (2008). Assessment centers help companies identify future managers. *Psychology matters*. Retrieved August 29, 2008, from http://psychologymatters.apa.org/

Bailey, A., & Drummond, K. (2006). Who is at risk and why? Teachers' reasons for concern and their understanding and assessment of early literacy. *Educational Assessment, 11*(3 & 4), 149-178.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5* (1), 7–74.

Bradley, D. F., & Calvin, M. B. (1998). Grading modified assignments: Equity or compromise? *Teaching Exceptional Children*, *31*(2), 24-29.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*. *22*(4)*,* 5-12.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*. *56,* 81-105.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*. *20*(4)*,* 19-27.

Cizek, G. J., Fitzgerald, S. M., & Rachor, R. A. (1995). Teachers' assessment practices: preparation, isolation, and kitchen sink. *Educational Assessment*, *3*(2), 159-179.

Cleveland, L. (2007). Surviving the reading assessment paradox. *Teacher Librarian. 35*(2), 23-27.

Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology, 78*, 141-146.

Demaray, M. K., & Elliott, S. N. (1998). Teachers' judgments of students' academic functioning: a comparison of actual and predicted performances. *School Psychology Quarterly, 13*(1), 8-24.

DeStefano, L., Shriner, J. G., & Lloyd, C. A. (2001). Teacher decision making in participation of students with disabilities in large-scale assessments. *Exceptional Children, 68*, 7 – 22.

Dolan, R. P. & Hall, T. E. (2001). "Universal Design for Learning: Implications for Large-Scale Assessment." *IDA Perspectives 27*(4), 22-25.

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C., & Kleinmann, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. *Psychology in the Schools, 43*, 247-265.

Feinberg, A. B., & Shapiro, E. S. (2003). Accuracy of teacher judgments in predicting oral reading fluency. *School Psychology Quarterly, 18*, 52-65.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C., Binkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children, 67*, 67-81.

Fuchs, L. S., Fuchs, D., & Capizzi, A. M. (2005). Identifying appropriate test accommodations for students with learning disabilities. *Focus on Exceptional Children, 37*, 1-8.

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.

Gresham, F. M., MacMillan, D. L. & Bocian, K. M. (1997). Teachers as "tests": differential validity of teacher judgments in identifying students at-risk for learning difficulties. *The School Psychology Review, 26*(1), 47–60.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27.

Helwig, R., & Tindal, G. (2003). An experimental analysis of accommodation decisions on large-scale mathematics tests. *Exceptional Children, 69*, 211-225.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*, 297-313.

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

Howe, K. B., & Shinn, M. M. (2002). Standard reading assessment passages (RAPs) for use in general outcome measurement: A manual describing development and technical features. Accessed November 1, 2008 at: http://www.aimsweb.com.

Johnstone, C. J., Thompson, S. J., Bottsford-Miller, N. A., & Thurlow, M. L. (2008). Universal design and multi-method approaches to item review. *Educational Measurement: Issues and Practice*, *27*(1), 25-36.

McDonnell, L. M., & McLaughlin, M. J. (1997). *Educating one & all: Students with disabilities and standards-based reform.* Washington, DC: National Academy of Sciences, National Research Council.

Meehl, P. E. *Clinical versus statistical predication: A theoretical analysis and review of the evidence.* Minneapolis: University of Minnesota Press, 1954.

Minnesota Department of Education (2008). *Minnesota manual of accommodations for students with disabilities in instruction and assessment – A guide to selecting, administering, and evaluating the use of accommodations: Training guide.* Roseville: Author.

Moen, R. E., Liu, K. K., Thurlow, M. L., Lekwa, A., Scullin, S., & Hausmann, K. (in process). Studying less accurately measured students. Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment.

Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice*. *22*(4)*,* 13-25.

New England Compact (2007). Reaching students in the gaps: A study of assessment gaps, students, and alternatives (Grant CFDA #84.368 of the U.S. Department of Education, Office of Elementary and Secondary Education, awarded to the Rhode Island Department of Education). Newton, MA: Education Development Center, Inc.

Perry, N. E., & Meisels, S. J. (1996). *How accurate are teacher judgments of students' academic performance?* National Center for Education Statistics Working Paper Series [No. 96-08]. U.S. Department of Education, Office of Educational Research and Improvement.

National Research Council. (2003). *Assessment in support of instruction and learning: Bridging the gap between large-scale and classroom assessment. Workshop report.* Committee on Assessment in Support of Instruction and Learning. Board on Testing and Assessment, Committee on Science Education K-12, Mathematical Sciences Education Board. Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Noble, J., & Sawyer, R. (2002). Predicting different levels of academic success in college using high school GPA and ACT composite score. ACT Research Report Series, 2002-4.

Ornstein, A. C. (1994). Grading practices and policies: An overview and some suggestions. *NASSP Bulletin, 78*(561), 55-64.

Perry, N. E., & Meisels, S. J. (1996). *How accurate are teacher judgments of students' academic performance?* National Center for Education Statistics Working Paper Series [No. 96-08]. U.S. Department of Education, Office of Educational Research and Improvement.

Pike, G. R., & Saupe, J. L. (2002). Does high school matter? An analysis of three methods of predicting first year grades. *Research in Higher Education, 43*(2), 187-207.

Popham, W. J. (2007). Instructional Insensitivity of Tests: Accountability's Dire Drawback. *Phi Delta Kappan, 89*(2), 146-150.

Price, F. W. & Kim, S. H. (1976). The association of college performance with high school grades and college entrance test scores. *Educational and Psychological Measurement, 36,* 965 – 970.

President's Commission on Excellence in Special Education. (2002). *A new era: Revitalizing special education for children and their families.* Washington, DC: U.S. Department of Education, Office of Special Education and Rehabilitative Services.

Prime Numbers. (2006). *Teacher Magazine, 17*(5), 5-10.

Shepard, L.A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Shinn, M. R. & Shinn, M. M. (2002). *AIMSWeb training workbook.* Eden Prairie, MN: Edformation, Inc.

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75* (4), 457-490.

Thompson, S. J., Thurlow, M. L., & Malouf, D. (2004). Creating better tests for everyone through universally designed assessments. *Journal of Applied Testing Technology, 6* (1). Accessed November 1, 2008 at: http://www.testpublishers.org/jattmain.htm.

Thurlow, M. L., Johnstone, C., & Ketterlin Geller, L. (2008). Universal design of assessment. In S. Burgstahler & R. Cory (Eds.), *Universal design in post-secondary education: From principles to practice* (pp. 73-81). Cambridge, MA: Harvard Education Press.

Thurlow, M. L., Moen, R. E., Liu, K. K., Scullin, S., Hausmann, K., & Shyyan, V. (2009). *Disabilities and reading: Understanding the effects of disabilities and their relationship to reading.* Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment.

Thurlow, M. L., Thompson, S. J., & Lazarus, S. S. (2006). Considerations for the administration of tests to special needs students: Accommodations, modifications, and more. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 653-673). Mahwah, NJ: Lawrence Erlbaum.

Thurlow, M. L., Ysseldyke, J. E., & Silverstein, B. (1995). Testing accommodations for students with disabilities. *Remedial and Special Education, 16* (5), 260-270.

Valencia, S. W., & Buly, M. R. (2004). Behind test scores: What struggling readers really need. *The Reading Teacher, 57* (6), 520-531.

Washington Office of Superintendent of Public Instruction. (2008). *Washington state's accommodations guidelines for students with disabilities*. Olympia, WA: Author.

Willingham, W. W., & Breland, H. M. (1982). *Personal qualities and college admissions.* New York: College Entrance Examination Board.

Willingham, W. W. (1985). *Success in college: The role of personal qualities and academic ability.* New York: College Entrance Examination Board.

Wood, G. H., Darling-Hammond, L., Neill, M., & Roschewski, P. (2007). *Refocusing accountability: Using local performance assessments to enhance teaching and learning for higher order skills.* Briefing paper prepared for members of the Congress of the United States. Accessed November 1, 2008 at: http://www.forumforeducation.org.