Running Head: CRITERIA TO EVALUATE INTERPRETIVE GUIDES

Criteria to Evaluate Interpretive Guides for Criterion-Referenced Tests

William J. Trapp

Northern Illinois University

DeKalb, IL

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE

MASTER OF SCIENCE

DEPARTMENT OF EDUCATIONAL TECHNOLOGY,

RESEARCH & ASSESSMENT

Abstract

This project provides a list of criteria for which the contents of interpretive guides written for customized, criterion-referenced tests can be evaluated. The criteria are based on the *Standards for Educational and Psychological Testing* (1999) and examine the content breadth of interpretive guides. Interpretive guides written for state-level, Grade 5 mathematics tests are evaluated using these criteria. The data are then analyzed to determine which criteria are found least or most often in the interpretive guides. Criteria for which data are inconsistent across interpretive guides will be used to provide suggestions for improvements to interpretive guides, or suggestions for future research.

About the Author

William Trapp is a Test Developer for the Riverside Publishing Company and is pursuing a master's degree at Northern Illinois University in Educational Research and Evaluation. William taught high school for 3½ years in the Illinois public school system. He specializes in customized, criterion-referenced, standardized tests.

Overview

The *Standards for Educational and Psychological Testing* (*Standards*) (1999) was developed by a joint committee appointed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (NCME). "The intent of the *Standards* is to promote the sound and ethical use of tests and to provide a basis for evaluating the quality of testing practices" (p. 1). The *Standards* go on to state that "Evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every standard in this document, and acceptability cannot be determined by a checklist" (1999, p. 4). Nonetheless, the *Standards* clearly state expectations for test developers, test administrators, and test users. The research, on which the *Standards* are based, along with the method in which the *Standards* were developed, requires their use in this research as the guiding principles from which expectations of tests and test materials should follow.

Along with the *Standards*, the *Standards for Teacher Competence in Educational Assessment of Students* (1990), which were developed by a committee composed of representatives from the American Federation of Teachers, NCME, and the National Education Association. The *Standards for Teacher Competence in Educational Assessment of Students* state that, "Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement" (1990, Standard 4). Using results from classroom assessments and standardized tests are two completely different things because of the nature of their creation and administration. Yet, this standard requires mastery of both. A limited literature review on the use of teacher-developed classroom assessments yields a

wealth of information about past and current practices and teacher attitudes. However, no further standards were uncovered that would provide additional criteria that could be used for the purpose of this research.

Operationalization of Constructs

Interpretive guides are one of the ancillary test materials whose content is governed by the *Standards*. Documents published under the title "interpretive guide," or documents which intend to explain the content contained in individual score reports are termed "interpretive guides" for the purpose of this research. An interpretive guide may be a document that is printed, a website, or a combination of both. An interpretive guide for a test should contain the information or the location of the information necessary to interpret a test's scores.

The interpretive guides of interest are those published for tests that were administered during the 2005-06 academic year. These tests may have been given in the fall, winter, or spring of that academic year to be considered for this research. An interpretive guide may be applicable for several years, and these interpretive guides are also considered as long they apply to the 2005-06 academic year.

Many types of tests were published and administered during the 2005-06 academic year, but the current research is limited to interpretive guides developed for state-wide, customized, criterion-referenced tests. These tests must be operational, meaning that they contain items that have been field tested previously, and scores generated and reported are based upon a cut score identified through a standard setting procedure. Each test has been customized to assess a state's, or group of states', curricula (i.e., this may also be referred to as a state's standards). Also, these tests have been

created specifically to satisfy one of the requirements of the *No Child Left Behind* Act of 2002 (NCLB Act, 2002). NCLB requires each state to assess mathematics annually in grades 3 through 8 inclusive starting in 2005. Due to time constraints and the wide range of tests, this research will be limited to the review of a select group of interpretive guides. The list of the interpretive guides that were evaluated for this project can be found in Appendix B.

Statement of the Problem

There is a need to know if published interpretive guides contain the information that stakeholders need to make valid interpretations of test data. In order to determine whether interpretive guides are adequate, criteria to evaluate them needs to be created and approved by industry professionals.

Purpose

The purpose of this research is to generate criteria that can be used to evaluate interpretive guides and to compare the content of interpretive guides. The criteria are designed specifically to evaluate interpretive guides for tests mandated by NCLB. By limiting the review of interpretive guides to this sample, the results of this research will be of interest to a larger population than if another type of test were used. Also, because these tests are mandated, there will be a greater likelihood that the interpretive guides are available for review.

Research Questions

The research questions for this study are:

What criteria are appropriate to use to evaluate interpretive guides published for state-level, grade 5, mathematics, customized, criterion-referenced tests?

Do interpretive guides published for state-level, grade 5, mathematics, customized, criterion-referenced tests that were administered during the 2005-06 academic year meet criteria based on the *Standards for Educational and Psychological Testing* (1999)?

Review of the Literature

There is a surprising lack of research written about interpretive guides. Not only is there an absence of research, there is a dearth of references pertaining to interpretive guides in the literature. Harris (2006) points out that states cannot share their tests used for NCLB because each state has a unique curriculum that is assessed. This necessitates that each unique test requires its own exclusive interpretive guide. The review of the literature was focused frequently on locating documents that could be used to guide the creation of criteria for evaluating interpretive guides. First, however, the availability of the desired interpretive guides was investigated.

Interpretive Guides

An initial review of the 37 interpretive guides found that they varied in title, length, format, breadth, and depth. Some interpretive guides provided only definitions of the information provided on a score report. Other interpretive guides reached far into tests and specified exactly what a test was intended for and what types of interpretations should be made at the student, classroom, school, district, and/or state level. Once the

general content of the interpretive guides was determined, the literature review continued with the intent of uncovering criteria that could be used to evaluate interpretive guides.

As stated earlier, the initial review of the literature found that little documentation existed that specified what information and level of detail should be contained in an interpretive guide. Articles found that focused specifically on test interpretation were either vague, as to cover all types of tests, (Amrein, 2002; Barton & Coley, 1994; Darling-Hammond, 1997; Kearney, 1983; Mertler, 2002), or were centered on psychological tests where the results were interpreted by counseling professionals (Tinsley & Bradley, 1986).

The tests emphasized in this project do not quite fit with either category. Several articles made reference to and encouraged the application of the *Standards* for such tests. Many such articles were found in the fall 2006 issue of *Educational Measurement: Issues and Practice,* which was devoted to articles related to the *Standards*. The *Standards* were referenced frequently in the literature, which supports their use for evaluating interpretive guides.

Standards for Educational and Psychological Testing

Camara (2006) provided a history of the development of the *Standards* and clearly documented how the *Standards* changed when they transitioned from the 1985 version to the 1999 version. He found that the number of standards related to the responsibility of test users went from 13 in the 1985 edition to 24 in the 1999 edition. Test users were the people who would be making interpretations based on test data. Therefore, the increase in the number of Standards related to test users' responsibility was important to the creation of criteria for evaluating interpretive guides. Camara states

that the 1999 version of the *Standards*, "went further in defining tests to include a broader range of instruments and assessments, as well as emphasizing their relevance to decision-making processes" (2006, p. 36). Appropriate interpretation of test data has played a key role in every prior version of the *Standards* and continues in the 1999 update.

The 1999 version of the *Standards* is now eight years old. However, NCLB was enacted three years after the current version of the *Standards* was published. Therefore, consequences of NCLB (such as assessing every student) are now highlighting certain areas that the *Standards* do not address. Harris (2006) looked at the difficulties of developing NCLB tests while looking to the *Standards* for guidance. He found that areas such as language translations and universal design principles could use refinement in the *Standards*. No literature was found that stated weaknesses in the interpretation guidelines in the *Standards*, nor was literature found that called for revisions to such areas. Although the current *Standards* date back to 1999, they still appear to be a solid foundation to refer to when determining appropriate ways to yield valid test interpretations.

Both the *Standards* and the literature state that the *Standards* are guiding principles. The *Standards* are not legally binding, nor is it required that every test meet every standard. Koretz (2006) found that, currently, self-regulation is the way the *Standards* are implemented, and other methods to enforce the *Standards* would be costly and time consuming. Also, Koretz states, "Many of the Standards are very general statements of principle, and it is not surprising that individuals in the field disagree about their implications when they are applied to controversial testing practices" (2006, p. 47). Based on this statement, the final criteria created to evaluate interpretive guides will certainly be both accepted and criticized based on this assertion. Therefore, the literature

was consulted about the current uses and interpretations of test data to help clarify what are realistic expectations of interpretive guide content.

Test Uses and Interpretations of Test Data

Several articles have been written about using test data. Some of these articles provide methods that can be implemented in a variety of testing situations. For example, Chester (2005), "explored the use of multiple measures to classify districts and schools" (p. 49). Also, Mertler (2002) looked at the teacher use of "empirical test data to assist in instructional decision-making" (p. 5). The methods described in these articles can be applied in other situations in order to make similar interpretations. Rudman (1989) provided a list of many general interpretations that can be made based on test data that are applicable in many testing situations. The literature also warns that using test data must be done with discretion. Rudman pointed out that for high stakes decisions, such as promotion and retention, there can be many factors that are part of that decision. Test data can be one of those factors, but it should not be the only factor. Chester states that "Conventional wisdom suggests that the use of multiple measures to reach high-stakes decisions improves the quality and fairness of the inferences" (p. 41). The *Standards* also suggest that having multiple measures is appropriate. Again, though, it depends on the severity of the decision being made.

Analyses of Test Uses

The literature cautions about incorrect uses of test data (Kearney, 1983; Nolen, 1992). Popham (1999) states that, "A standardized test provides a misleading estimate of a school staff's effectiveness" (¶ 2). Bauer (2000) found that, "it does not logically follow that [standardized achievement tests] should be useful as indicators of school

performance" (Summary and Conclusions, ¶ 2). These articles are difficult to generalize from because each state has a unique standardized achievement test. Still, the methods can be adapted to perform research on similar testing situations.

The approaches that articles take in the literature range from focusing on specific interpretations of certain tests to broad interpretations of tests in general. For example, Klein et al. (2000) investigated a specific interpretation made from data on the Texas Assessment of Academic Skills test. On the other hand, Koretz (2002) did not specify one particular test when describing ways to improve the use of accountability tests. Both types of articles provide useful information when determining appropriateness for making an identical interpretation on another test. Also, both types of articles can support what is called for in Standard 6.9 of the *Standards*, "Test documents should cite a representative set of the available studies pertaining to general and specific uses of the test" (1999, p. 69). The amount of literature on test use continues to grow, as does the amount of literature on the ability of teachers to interpret test data.

Teachers Lack the Training Needed to Interpret Test Data

The *Standards for Educational and Psychological Testing* and the *Standards for Teacher Competence in Educational Assessment of Students* both state that teachers need to be competent in interpreting standardized test results. However, research studies have shown concern that teachers receive less than adequate training in the areas of measurement and test interpretation (Gullickson, 1986; Mertler & Campbell, 2005; Quilter & Gallini, 2000; Stiggins, 2004). The assessment literacy of teachers is important for classroom assessments, as well as standardized assessments. Popham (2004) suggested several ways to increase one's assessment literacy, including reading books

and discussing them with colleagues. However, teachers may not be aware that their assessment literacy is poor. The Assessment Literacy Inventory, developed by Mertler and Campbell, can help with the problem of identifying who needs training in assessment topics. Professional development activities in the area of assessment can then be arranged by school administrators or provided at local colleges. If interpretive guides are the only source of guidance from which teachers have to make interpretations, then the clarity of the language in the guides must be a concern to the writers of the guides.

Unclear Evidence About the Effectiveness of Professional Development

Research studies have provided conflicting evidence about whether additional training for teachers yields growth in their competence. "Personal experiences with testing play an important role in understanding teachers' current attitudes toward assessment, whereas their professional training in educational measurement may play a negligible role" (Quilter & Gallini, 2000, p. 128). In contrast, O'Sullivan & Johnson (1993) found that students who completed a performance-based measurement course improved their assessment competency. In Brookhart's (2001) review of Plake, Impara, and Fager's (1993) study, she found that instruction caused a difference of less than 1 point on a 35-point test on assessment competency. While this difference was statistically significant, it appears to be a minor improvement in assessment competency. McCoy (2007) investigated the use of a pre-test to determine school staff's ability to interpret data and found that educators struggled with data interpretation. Additional research into the effectiveness of training (Bettesworth, 2007) found that some areas of teacher data interpretation improved after an intervention. However, none of the recent research indicated whether teachers can go back to their schools and apply their new

understanding to their school's data. Research has not determined whether additional education will improve competency in test interpretation, so there is a need to provide teachers with clear direction about interpreting customized tests. This is especially critical when dealing with new tests, or tests that may have been revised recently, due to legislation such as NCLB. Interpretative guides are the source for such information, but no research has been done to determine whether they are adequate or even being used.

Barriers to Test Interpretation

Concerns about their ability and training aside, teachers and school administrators make decisions with the goal of improving test scores. Several research studies have documented the types of changes being made in schools (Brown, 1993; Cimbricz, 2002; Darling-Hammond & Wise, 1985; Ingram, Louis, & Schroeder, 2004; Koretz, 2002; Natriello & Pallas, 1999; Nolen, Haladyna, & Haas, 1992; Yeh, 2005). The foundation for these changes still remains in question because it is not clear whether such changes are appropriate based on test construction. Without being able to interpret the test data appropriately, making changes becomes a trial and error effort. Also, making decisions that are not appropriate based on the construction of the test undermines the criterion validity of the test. Ingram, Louis, and Schroeder (2004) listed seven barriers to data-driven decision making. Interestingly, none of the barriers are tied directly to teachers' lack of ability to read and apply data from score reports. Rather, the barriers listed describe how many teachers use intuition and other metrics for assessing students. This mindset prevents teachers from seeing how standardized test data can be useful.

Perceptions of Testing

Grant's (2000) study of teacher and administrator perceptions of the State of New York testing program, just prior to the signing of NCLB, found that mixed messages were being sent about the purpose(s) of the standardized tests by the state department. Brown's (1993) study found that teachers and principals across three states "reported that they were not aware of the purposes for the establishment of mandated testing" (p. 17). Also, Cimbricz (2002) noted that there are a limited number of studies that uncover the relationship between state testing and teaching. This disconnect between state testing and teacher practices makes it difficult to pinpoint where the transfer of information between state education departments and test users is breaking down.

Jones and Egley (2004) concluded that, "Until policy makers take teachers' concerns seriously and make an effort to address them, teachers will not likely support reform through high-stakes testing" (Conclusion, ¶ 1). Also, Zancanella (1992) found that, "understanding the specific dynamics of interactions among teachers, learners, content, and tests has become a necessity" (p. 294). Teachers feel stress over tests in many different ways. Mabry and Margolis (2006) studied the NCLB impact on classroom practice and school climate. They determined that test anxiety caused several of the 4th grade participants (teachers) in their study to leave. Also, Rudman (1989) stated that the literature shows teachers as less supportive of testing than in actuality. There are many different perceptions of testing, and many of these perceptions come from direct experience with standardized tests.

The Need for Clear Interpretive Guides

Clearly, research is needed about how students, parents, and teachers make interpretations of test data. The information needed to make interpretations should be stated in a test interpretation guide, and such information is unique for all customized, criterion-referenced tests.

Methods

Creation of the Criteria

1. The author created the first draft of the criteria.
2. The criteria were used to code a random sample of 10 interpretive guides.
3. The criteria were updated based on the results of rating the 10 interpretive guides.
4. The criteria were reviewed by 5 subject matter experts.
5. The criteria were updated based on the subject matter expert comments.

Data Collection

6. The interpretive guides were located.
7. The criteria were used to code all interpretive guides. Twenty of the interpretive guides were also coded by an additional coder.

Data Analysis

8. The data were analyzed using descriptive statistics and additional inter-rater reliability analyses.
9. Conclusions were made based on the analysis of the data.

Data Triangulation

10. The literature was revisited based on the findings.
11. The conclusions were updated based on the review of the literature.

12. The data analysis and the conclusions were reviewed by subject matter experts.
13. The conclusions were edited based on comments from the subject matter experts.

Final Report

14. The final report was written.

Creation of the Criteria

A review of the *Standards* yielded 15 criteria upon which to evaluate interpretive guides. The criteria were developed by the author based upon his expertise in criterion-referenced tests, experience in large scale testing, and insight from the literature. Each criterion consisted of three parts: the criterion, the standard(s) upon which the criterion is based, and a description that clarifies how the criterion is being applied. This first draft of the criteria was used to code a random sample of 10 interpretive guides. Edits were made to the description section of some of the criterion to clarify how the criterion should be applied. Then, the updated criteria were sent to five subject matter experts for their review.

Three of the subject matter experts are employees in state departments of education. They represent three different sections of the country and have varying years of experience. All three work directly with assessments that are mandated by NCLB.

The fourth subject matter expert is an employee for a test development company. They have vast experience in developing items, tests, and ancillary documents for tests. This expert has worked in the test development industry for over 10 years.

The fifth subject matter expert is currently retired and has worked on tests for governmental entities and for test publishing companies in two countries. This expert's most recent experience was managing state contracts for a testing company, while often being involved in reviewing and evaluating the tests developed.

Due to time constraints, the subject matter experts were a convenience sample selected from previous contacts with the author. However, the expertise of each subject matter expert allowed for a rigorous review of the criteria. Editorial comments and comments by the subject matter experts, which clarified the criteria, were implemented. Additional comments pertaining to the criteria, which suggested collapsing criteria; adding criteria that was not consistent with the *Standards*; or organizing the criteria, were not implemented. Such comments were taken under consideration and were revisited after the data were collected. Examples of comments that were not implemented are listed below:

> "Organize the criteria [under the following headings]:
>
> - Context of testing
> - What the scores mean
> - What the scores can be used for"
>
> "There should also be some explanation of each category of scores. For example scale scores versus percent correct, or scores for subdomains of a content area. How about examples of reports…?"

The subject matter expert review yielded minor changes to the wording of the criteria, to the organization of the standard(s) supporting each criterion, and to the wording of the descriptions. The final list of criteria, which was updated with the subject matter expert comments, can be found in Appendix A.

Data Collection

Interpretive guides written for state-level, grade 5, mathematics, customized, criterion-referenced tests that were administered during the 2005-06 academic year were located on the internet. Thirty-seven interpretive guides were located, but two of the interpretive guides located did not fit the initial requirements. The interpretive guide for the New England Common Assessment Program covers the test for New Hampshire, Rhode Island, and Vermont, and only the 2005 document was located. Also, the interpretive guide for the New Jersey grade 8 mathematics test was located, but the document for grade 5 could not be located. The author attempted to contact state department employees for states in which interpretive guides were not posted on the internet. Some states did not have an interpretive guide, and in other cases the author was unable to contact a state department employee. A list of all interpretive guides obtained, and their sources, can be found in Appendix B.

Each interpretive guide was coded by two independent raters using the final criteria from Appendix A. After beginning to collect data, it was noticed immediately that there was not enough information to code for criterion #1. Criterion #1 is met when the interpretive guide is made available at an appropriate time such as when the score reports are delivered. The coders were unable to determine when the interpretive guides were either posted, or delivered; therefore, criterion #1 was not coded.

To increase reliability, the two coders first coded an interpretive guide together. Then, they coded another interpretive guide independently. They met to compare their results and discussed differences until a consensus was obtained. Then, an additional 18 interpretive guides were coded individually. If an interpretive guide satisfied a criterion, it was marked as "yes" or "1." If an interpretive guide did not satisfy a criterion, it was marked as "no" or "0." Each time the coders disagreed, they discussed the difference in rating and reached a consensus. The original data from each coder were analyzed further, and the final data that the raters reached a consensus on were used for this study's data analysis. Due to time constraints, all of the interpretive guides were unable to be coded by a second rater.

To check the inter-rater reliability after coding was complete, the following table was used to record the results:

Table 1

Inter-rater Agreement and Disagreement Counts

	Coder 1 (1)	Coder 1 (0)	Sum
Coder 2 (1)	n_{11}	n_{01}	$n_{11}+ n_{01}$
Coder 2 (0)	n_{10}	n_{00}	$n_{10}+ n_{00}$
Sum	$n_{11}+ n_{10}$	$n_{01}+ n_{00}$	N

Cohen's kappa statistic (K) was calculated to determine the index of inter-rater reliability using the following equation.

$$K = (\Sigma a - \Sigma ef)/(N - \Sigma ef), \text{ where,}$$

Σa = the total number of agreements

ef = the sum of the expected frequencies of agreement by chance

N = number of cases

Lastly, a McNemar's test was also used to analyze the significance of the differences between coders. The value of n_{11} and n_{00} are ignored and the test determines whether n_{01} is as likely as n_{10}. McNemar's statistic is a chi-square-based statistic with 1 degree of freedom.

$$X^2 = (|n_{01} - n_{10}| - 1)^2 / (n_{01} + n_{10})$$

Data Analysis

The data were recorded in an Excel spreadsheet and in SPSS (Statistical Package for the Social Sciences) and analyzed using descriptive statistics and statistics to measure score reliability. Data of specific interest were the range of criteria met by interpretive guides and the average number of criteria met by the interpretive guides. Additional descriptive statistics revealed patterns that identified specific criterion that were met frequently or infrequently. Table 2 shows the number of criteria met by each state's interpretive guide.

Table 2

Number of Criterion Met

State	Number of criteria met (14 possible)	State	Number of criteria met (14 possible)
New York	3	Tennessee	6
Michigan	4	Alaska	7
Mississippi	4	Arkansas	7
New Mexico	4	Connecticut	7
West Virginia	4	Kansas	7
Alabama	5	Minnesota	7
Arizona	5	Missouri	7
Delaware	5	Oklahoma	7
Illinois	5	Washington	7
Indiana	5	Georgia	8
Maryland	5	Texas	8
Montana	5	Wisconsin	8
North Dakota	5	Kentucky	9
Virginia	5	NECAP	9
California	6	New Jersey	9
Florida	6	South Carolina	9
Louisiana	6	Colorado	10
Massachusetts	6	South Dakota	10
Ohio	6		

Each interpretive guide reviewed met at least three of the criteria. No interpretive guide met more than 10 criteria. The average number of criteria met was just above 6, with a standard deviation of 1.80. Twenty-four percent of the interpretive guides met more than half of the criteria reviewed. One guide received the minimum score (3) and two guides received the highest score (10). The median of the data was 7, with 8 of the interpretive guides receiving said score. Table 3 lists additional statistics about the number of criteria met by each state.

Table 3

Statistical Summary of the Number of Criterion Met

Statistic	Value
Minimum	3
Maximum	10
Range	7
Sum	236
Mean	6.38
Standard Deviation	1.80

The percent of interpretive guides meeting each criterion ranged from 0 to 100. Criterion 7 was not met by any interpretive guide, and 2 additional criteria were met by 2 or fewer interpretive guides. Six of the criteria were met by 62% or more of the interpretive guides. Table 4 shows the percent of interpretive guides that met each criterion.

Table 4

Number of Interpretive Guides Meeting Each Criterion

Criterion	Number	Percent
1	N/A	N/A
2	37	100%
3	15	41%
4	34	92%
5	32	86%
6	2	5%
7	0	0%
8	4	11%
9	24	65%
10	2	5%
11	34	92%
12	11	30%
13	23	62%
14	14	38%
15	4	11%

To measure inter-rater reliability, the inter-rater agreement and disagreement was recorded in the table 5 and analyzed with Cohen's kappa statistic and also with a McNemar's test.

Table 5

Inter-rater Agreement and Disagreement Counts

	Coder 1 (1)	Coder 1 (0)	Sum
Coder 2 (1)	105	34	139
Coder 2 (0)	16	125	141
Sum	121	159	280

The inter-rater score reliability was K = 0.64 (standard error = .06; 95% confidence intervals = (.53, .76)). Landis and Koch (1977) provide a guide concerning general interpretations of kappa values, where inter-rater agreement values are seen as:

< .00 = Poor

.00 to .20 = Slight

.21 to .40 = Fair

.41 to .60 = Moderate

.61 to .80 = Substantial

.81 to 1.00 = Almost Perfect

These guidelines indicate a substantial level of agreement between the coders.

The McNemar's test yielded X^2 value = 5.78 ((1), $p < 0.025$). This result indicates that the disagreement was not evenly spread. The effect size as determined by Φ was equal to 0.82. Values of Φ greater than 0.50 indicate a large effect size. The amount of variance in the ratings accounted for is measured by Φ^2 and was equal to 67%.

Data Triangulation

Subject matter experts were asked to review the final data analysis and the initial conclusions. Each subject matter expert has direct experience with the type of test for which the interpretive guides were written. A total of six subject matter experts reviewed the findings. The combined experience of the subject matter experts covered four testing companies and two state departments. Currently, one subject matter expert is working for a state department, one is retired, three are working for testing companies, and one is a consultant.

Due to time constraints, the subject matter experts were selected by the same manner as the subject matter experts who reviewed the initial draft of the criteria (i.e., a convenience sample selected from previous contacts with the author). The subject matter experts shared their comments through a survey that contained five Likert-type questions with room under each question to provide additional clarification. The four options for each question were: strongly disagree, disagree, agree, and strongly agree. A copy of the protocol can be found in Appendix C.

Prior to sending the survey to the subject matter experts, it was directed through the proper Institutional Review Board (IRB) steps. First, a professor reviewed all of the materials being shared with the survey participants, along with the IRB application. Then, the department chair reviewed the application and materials. Lastly, the application and materials were sent to the IRB for approval. Once approved, the subject matter experts were sent the materials and were given two weeks to provide responses. Initially, nine subject matter experts were contacted, but only six replied by the deadline, yielding a return rate of 67%.

The survey was scored as follows:

- 1 = Strongly Disagree
- 2 = Disagree
- 3 = Agree
- 4 = Strongly Agree

Table 6 shows the results of the survey.

Table 6

Average Response to Survey Questions

Question	Average
1	3.50
2	3.50
3	1.67
4	1.67
5	2.00

The strong, positive response to questions 1 and 2 indicate that subject matter experts felt that some criterion should be satisfied by all interpretive guides. To elaborate further, one subject matter expert listed the following applicable criterion: "Purpose of test, how results will be reported, uses and possible misuses, limitations, and error (SEM)." Another subject matter expert wrote, "It seems that each of the criteria created based on the *Standards for Educational and Psychological Testing* are important. It could be argued that some are more important or even prerequisites to others." Beyond these comments, the subject matter experts provided no additional information about which specific criterion or criteria should be satisfied by all interpretive guides.

The negative response to question 3 confirms what was found in the literature; the *Standards*, though 8 years old, are still appropriate to use as a building block for the criteria. No further clarification was given by the subject matter experts.

Question 4 asked participants whether interpretive guides were the only method by which state departments shared information about interpreting test results. The response by the subject matter experts indicated that state departments used several media to share information about interpreting results. All but one of the subject matter experts provided clarifying information concerning their response:

> "Technical reports, information bulletins, the tests themselves (or sample items with responses), and presentations are also used."
>
> "I think that some of the information listed in the 15 criteria would be found in other places such as literature (Coordinator Manuals, Directions for Administration, etc....) sent to districts, school, and/or parents before the assessment, technical reports, and test/item specifications."
>
> "I think they're the most formal vehicle for publicizing certain information, but I also see a lot of information that goes out informally—statements to the press, teacher committees, etc. And sometimes the reports themselves have interpretive information on the back or in small print."
>
> "Based on my experiences with state departments of education from content reviews, in general, info comes out in different bits

> and from different directions. It is likely the states do address some of the criteria here but not in a collective interpretive guide."
>
> "Brochures, The Web, On Score Reports, Technical Reports, Media, and Training (of District Administrators ala train-the-trainer model) are other methods."

The variety of media listed by the subject matter experts brings about a new concern. If these materials are critical to interpreting the test results, why aren't they at least referenced in the interpretive guide? Three of the subject matter experts referenced technical reports in their clarifications. However, technical reports are, by nature, difficult to understand unless having been trained in some level of psychometrics. Possibly the subject matter experts are suggesting that some criterion are too technical to be put into an interpretive guide.

Question 5 of the survey summed up the purpose of the research. The subject matter experts were divided on the question. Two strongly disagreed that interpretive guides met the criteria, two disagreed, and two agreed. The two subject matter experts who had the most experience, in terms of years and in types of occupations marked, "agree" on the survey. The two subject matter experts who had the least experience, from the point of view of a state department marked, "strongly disagree." Therefore, this question seems to be related to the essence of the current research. Perceptions of what is adequate and inadequate appear to shift based on an understanding of the point of view of a state department employee. While the extremely small sample size does not allow for anything beyond a conjecture, a larger question emerges. Input from parents and teachers may be needed to clarify this apparent disagreement pertaining to the adequacy of

interpretive guides. The subject matter experts were relied upon at this point to bring experience to the data and offer values of how many criteria are needed to be adequate. Rather than draw a dividing line, many lines have been drawn and it appears up to each individual (person and/or state) to determine how many criteria are needed to be adequate. This is a reflection of the current educational system in the United States.

Discussion

The rigor in which the criteria were applied is immediately in question. All interpretive guides met criterion #2, which states that the interpretive guide must be written clearly for the intended reader. The judgment of the coders determined that all guides were appropriate, but no specific measurement tool, such as Lexiles, was used to confirm this. It was clear that an attempt was made by the author of each interpretive guide to write often in non-technical language. However, criterion #2 was not used to measure how clear the language was; rather criterion #2 was marked as being met if it appeared that the author made an attempt to write clearly for the intended reader. The results for each criterion should be reviewed in the same manner.

As indicated by both the range of criteria that the interpretive guides met (i.e., minimum = 3 and maximum = 10) and the range in percent of the number of interpretive guides meeting each criterion (i.e., minimum = 0% and maximum = 100%), there was much diversity among interpretive guides. For the two guides that met 10 of the criteria, they received the same score in only 8 of the criteria. The reason for this variation is unclear. One subject matter expert pointed out that the criteria do not address score reports. Nearly every one of the interpretive guides contained references and/or examples of student score reports. Many of the interpretive guides appear to have been written with

the purpose of explaining the score reports, rather than explaining what interpretations can be made based on the data in the score reports.

The mean number of criteria met by each interpretive guide was 6.38 out of a possible 14 criteria. If we assume that each interpretive guide met criterion #1 (i.e., that it was available at an appropriate time such as when the score reports were delivered), then the average would be 7.38 out of a possible 15 criteria. Twenty of the 37 guides met less than 50% of the criteria. This is disconcerting if interpretive guides are the only source of information that readers have to help them make decisions about their child's education. However, there is a possibility that states are supplementing this information with either a presentation at a conference or through other documents.

If interpretive guides are being supplemented by other information, it was unclear where this information could be located. The interpretive guides themselves were often difficult to locate. A search on each state department of education's website for "interpretive guide" often yielded results other than the desired document. No records were kept on the amount of time it took to locate each interpretive guide, nor were records kept about what navigation on websites was necessary to get to each interpretive guide. This information would be interesting to analyze in a different study if it had been documented.

When an interpretive guide could not be found on a state's department of education website, contact information for someone responsible for the assessment was searched. This search was necessary for 12 states. The attempt to locate the interpretive guide through the state department employee was successful for 1 of the 12 states. For the other 11 states, either the interpretive guide did not exist or the state department

employee could not be contacted. To reiterate the previously stated concern, it is disconcerting if interpretive guides are the only source of information available to help make decisions. If they are the only source of information, then 22% of the states have no communication to parents and/or teachers about how to make valid interpretations.

Each criterion is unique and independent. The remainder of the discussion is devoted to further clarification about each criterion and comments that were provided by either the coders or the subject matter experts.

Criterion #1 was not coded because there was not enough information to determine if the interpretive guides were available at an appropriate time. This criterion is important, but difficult to determine whether it was met or not.

Criterion #2 was met by 100% of the interpretive guides reviewed. In general, the interpretive guides were written in clear, concise language. The technical nature of some of the topics in the interpretive guides caused the complexity of the language to fluctuate. For example, in the interpretive guide for Kansas it states that, “Total test equated scores are reported for each content area” (p. 5). While the guide goes on to define some of the terms, the language will likely be difficult for a large number of readers (whether teachers or parents). Overall, the interpretive guide for Kansas, and the other states, was written in clear, concise language which is why they were coded as meeting the criterion. The coders gave the same rating for this criterion to all interpretive guides. However, having a formal analysis, such as Lexiles, would help ensure reliability when determining if the interpretive guides met the criterion.

An example of an interpretive guide meeting criterion #3 is Colorado. An entire chapter of the Colorado interpretive guide is devoted to listing proper preparation

activities, appropriate interpretation of test results, and inappropriate interpretations of test results. One example provided in the Colorado Student Assessment Program (CSAP) interpretive guide of an inappropriate interpretation is, "Basing student retention or promotion decisions on CSAP results alone" (p. 23). The Connecticut interpretive guide cautions, "While scale scores are comparable across forms in a given subject area within the same grade, they are not comparable across subject areas or grades" (p. 15). This is the only misinterpretation found that the Connecticut interpretive guide cautions against. The coders did not determine, or judge, whether stated misinterpretations were, in fact, misinterpretations. Rather, if the interpretive guide cautioned against any specific misinterpretation, the guide was marked as meeting the criterion. The coders had different ratings for 30% of the interpretive guides under this criterion. These were resolved by revisiting each interpretive guide and determining whether or not the statement was a specific misinterpretation. Exactly half of the different ratings were reconciled as meeting the criterion and the other half as not meeting the criterion. No suggestion to improve the criterion was provided; rather it appears that the interpretive guides need to improve on the clarity of the stated misinterpretations.

Many of the interpretive guides met criterion #4 by identifying the specific curriculum that the test was assessing. The interpretive guides listed either the name of the curriculum, a partial listing of the curriculum, or the entire curriculum being assessed. Only three of the interpretive guide reviewed did not contain any reference to the curriculum being assessed.

The title of the interpretive guide typically provided the needed information to meet criterion #5. For example, the Washington interpretive guide is named "2006

Reaching Higher." However, the South Dakota interpretive guide, while published in 2005, did not state anywhere the year(s) for which the interpretive guide is appropriate. There was no consistency among states about how often new interpretive guides were drafted. Some states produced new interpretive guides each year, and others may produce a new guide once every three years. Therefore, it is necessary to define clearly for which year(s) the guide is appropriate. Eighty-six percent of the interpretive guides reviewed provided this information.

Criterion #6 was met by 2 of the 37 interpretive guides reviewed. This criterion was considered met when an interpretive guide stated what qualifications were needed to interpret test scores. The Colorado interpretive guide met this criterion by stating, "These individuals include, but are not limited to, classroom teachers, principals, school psychologists, superintendents, district staff, State Department of Education staff, and educational research and policy professionals" (p. 19). For the majority of the interpretive guides, it is unclear whether anyone can interpret the scores, or if the interpretive guide only states the interpretations that target audience should be making.

Criterion #7 was not met by any of the interpretive guides reviewed. Not one interpretive guide had any reference to supporting documentation or studies that validated the stated use(s) for the tests. One subject matter expert expressed concern during their review of the criterion that adding such documentation might make the guide too technical for some readers.

Four of the 37 interpretive guides reviewed met criterion #8. Several interpretive guides mentioned the different types of administrations available (e.g., braille, large print, read-aloud accommodation). However, many of the interpretive guides did not specify

whether or not scores from students who took the test under different administration methods can be compared. The South Dakota interpretive guide states that, "An appropriate or reasonable accommodation should not interfere with the interpretation of a student's score" (p. 29). This statement, or a similar statement, would allow the interpretive guide to meet criterion #8. This criterion had the lowest score reliability between coders. The coders gave each interpretive guide the same rating only 50% of the time. The interpretive guides were often vague in the area of comparing scores. The coders found that it was difficult to determine in several interpretive guides whether or not the information was about comparing different content areas or about comparing the scores of students who had different accommodations. No recommendations were suggested to clarify this criterion. Rather, the interpretive guides needed to be clearer as to what they were referring.

If an interpretive guide provided any additional information about the test construction, then it met criterion #9. The Arkansas interpretive guide met this criterion by using a question and answer format. One of the questions was, "The test takes too long. Why does this test take so much longer than other tests?" (p. 2). The Montana interpretive guide listed a great deal of general information about the test in an appendix. There was a range in detail provided in the guides, which took part in lowering the inter-rater reliability for this criterion. The coders gave the same rating to an interpretive guide 70% of the time. A suggestion to improve the clarity of the criterion is to provide more examples in the description of what information is required to satisfy the criterion.

Criterion #10 was met by two interpretive guides. The South Dakota interpretive guide met this criterion by listing both student/home factors and school factors that

should be considered when interpreting the test results. Some of the interpretive guides described the test results as a 'snapshot' in time of a student's performance, but very few went a step further to suggest that other factors may have influenced the student's performance while they took the test.

Ninety-two percent of the interpretive guides reviewed met Criterion #11 by providing a general statement about the intended use of the test results. One of the general purposes stated by the Indiana interpretive guide was, "To what extent an individual student has mastered the Indiana Academic Standards" (p. 7). The criterion appears to be clear and direct, and the coders had different ratings for only two of the interpretive guides.

The *Standards* state that, "Score reports should be accompanied by a clear statement of the degree of measurement error" (p. 146). This statement is part of what criterion #12 is based on, and 30% of the interpretive guides reviewed met this criterion. The Oklahoma interpretive guide showed an example of the measurement error in a score report. The Massachusetts interpretive guide provided examples of interpretive guides, but did not show or describe the measurement error. The measurement error was difficult to explain in simple terms, and several interpretive guides used language very appropriate for both parents and teachers.

Criterion #13 was the key criterion, as it asks whether the interpretive guide is, in fact, describing how to interpret test scores. Sixty-two percent of the interpretive guides provided statements that specified one or more ways that test scores should be interpreted. This criterion had the second lowest inter-rater reliability. The coders gave 12 of the 20 interpretive guides the same rating. Often a general statement that would satisfy

criterion #11 would be marked as meeting criterion #13 as well. However, closer analysis found that the general statement was not specific enough and did not give a clear direction to interpret an individual student's score. It is unclear why over one-third of the interpretive guides did not state an interpretation that could be made. Many of the interpretive guides referenced the performance level descriptions that are often a part of standard setting meetings. However, referring the reader to these descriptions was not considered meeting this criterion unless the descriptions went beyond stating information in the curriculum. For example, the New England Common Assessment Program provides within its achievement level descriptions information such as "Errors made by these students are few and minor and do not reflect gaps in prerequisite knowledge and skills" (p. 8). This statement appears on the individual student score report and helps parents and teachers determine to what extent interventions are needed.

The inter-rater reliability was low for criterion #14. The coders did not have the same rating for 6 of the 20 interpretive guides, and the coders suggested that the criterion be clarified. The disagreement often happened because it was unclear concerning which groups the criterion was referring. After further review, it was determined that the criterion was referring to subgroups of students based on gender, ethnicity, or accommodation. The criterion was not referring to groups of students by classroom, school, or district. Thirty-eight percent of the interpretive guides met criterion #14. The interpretive guides that did meet criterion #14 clearly stated exactly what accommodations were provided to specific groups of students. The low amount of interpretive guides that met this criterion is due likely to the fact that many of the interpretive guides focused solely on the individual score report without addressing

accommodations. One of the reasons that the interpretive guides did not go beyond the individual score report is because students and parents do not see the higher level group reports.

Criterion #15 was met by 11% of the interpretive guides. As stated in the paragraph above, several interpretive guides did not provide information beyond the individual score report. For the interpretive guides that did, only four showed and explained measurement error present in the group data.

After coding the interpretive guides and analyzing the data, there appear to be some benefits to reorganizing the criteria as one subject matter expert suggested. However, when coding the interpretive guides, it was found that each interpretive guide is unique, and few shared a set pattern or organization. Rearranging the criteria may assist future coders and others who try to create interpretive guides based on the criteria, because several criteria were often satisfied by one section in an interpretive guide. Therefore, if additional research is done using the criteria, the reordering of the criteria would be appropriate.

This research attempted to determine whether interpretive guides published for state-level, grade 5, mathematics, customized, criterion-referenced tests that were administered during the 2005-06 academic year met criteria based on the *Standards for Educational and Psychological Testing*. From the data collected, it appears that such interpretive guides do not meet the criteria. Due to the lack of research published on interpretive guides and because the *Standards* do not specify how well interpretive guides should conform to the *Standards*, the findings must be thoughtfully weighed. Interpretive

guides are written for a variety of audiences, but past and current research shows that different audiences all have a limited understanding of the standardized tests in question.

Validity

Due to the lack of prior research in the area of interpretive guides, the face validity of this research is in question. Subject matter experts review of the criteria prior to coding offers evidence of face validity. The subject matter experts had several years of experience in one or more of these three professions:

- Teaching
- State department of education (assessment division) employee
- Test publishing company employee

Comments from the subject matter experts are integrated in the project as either edits to the criteria or as suggestions considered. Therefore, a triangulation of the criteria was achieved by using the expertise of the author, support of the literature, and approval by subject matter experts. This triangulation ensured that the criteria being used were appropriate, consistent with the *Standards*, and yielded information deemed appropriate to experts in the field.

Content and construct validity of the project were supported by using member checking after the data were collected, analyzed, and summarized. Subject matter experts with the same expertise of those who commented on the criteria were used. Several of the same subject matter experts were used, along with previously unused experts in the same fields. Comments from the subject matter experts appear in this final project either as:

- edits to the conclusions
- support for existing conclusions

- support for debasing existing conclusions
- new conclusions

Triangulation of the data were achieved by using the literature and subject matter expert input related to the conclusions. Member-checking ensured that the conclusions were appropriate, consistent with the *Standards*, and yielded new information that could be applied to the field. These steps helped address weakness in the areas of content validity and construct validity. Because of the nature of the study, criterion validity was not addressed.

Reliability

The research design brings about concerns in the area of reliability. Having two individuals code the interpretive guides independently provided a way to measure inter-rater reliability and internal consistency. This level of agreement indicates that the criteria were clearly written overall, and that the training that the raters received was adequate for the task. Having a low number of errors indicates that the majority of the criteria were applied consistently. Also the low number of errors indicates that one or more criteria were difficult to apply and that one or more interpretive guides did not lend themselves to be evaluated by the criteria.

The results of the McNemar's test is reflected in the data by the fact that coder 1 gave a 0 rating and coder 2 gave a 1 rating for a criterion over twice as often as the inverse occurred. When the coders disagreed, the final reconciliation agreed with coder 1 66% of the time. The reconciliation process that the coders participated in increased the overall internal consistency reliability of the raters.

A final attempt to ensure inter-rater reliability of the data is the detail that has been given to the criteria. Each criterion is supported by the *Standard(s)* upon which it is based. Also, each criterion has a description specifically written to address the use of the criterion with the type of interpretive guides being rated. Providing this level of detail in the criteria helped ensure that the coders applied the criteria consistently. The detail in the criteria also assists in the area of future research by applying the criteria consistently if the research is repeated or if the criteria are applied in another context.

Threats to Internal Validity

The *Standards*, published in 1999, are now eight-years old. This poses a threat to the history aspect of internal validity because the relevance of the *Standards* may be fading. Research has not yet shown the need for revised, or even new, standards, but it is also an anticipated outcome that the application of the *Standards* for this purpose may cause them to be re-critiqued. Also, by the time this research is completed, many interpretive guides written for 2006-07 academic year tests will have been published. This also poses a threat to the history aspect of internal validity because many updates may have been made to interpretive guides and the data collected would not be as applicable. However, interpretive guides tend to change little unless new tests are created or new contractors are hired.

The limited number of subject matter experts presents a threat to the selection aspect of internal validity. The individual subject matter experts may not have equivalent content and/or assessment expertise in the area of interpretive guides. Therefore, their comments may be unclear or may even be in conflict with each other. Basing each criterion on specific standards will help reduce the potential for ambiguous comments

from subject matter experts. Another threat to the selection aspect of validity is the use of a convenience sample. Each subject matter expert provides a unique view from either a corporation, a state department, or both. However, each state department and each corporation is unique. Therefore, a random sample would have better supported this aspect of validity.

Threats to External Validity

A major threat to the ecological aspect of external validity is the lack of research currently published on the topic of evaluating interpretive guides. There are no previous criteria to use or with which to compare. Also, there are no previous results of evaluating interpretive guides to compare the results of this study.

Because of the interaction between the setting and the criteria, the results of the study cannot be generalized for all interpretive guides. Specific criteria can be created to evaluate other interpretive guides for other types of tests, but the criteria created for this study is specific to the interpretive guides evaluated in the study.

Additionally, the subject matter expert comments are a threat to external validity because they are composed of a limited sample of responses, and they may not be representative of experts in general in the field of interpretive guides. This interaction between the setting and the participants decreases that chance that peer-reviewers will agree with the conclusions.

Limitations

Weaknesses in the design of this study that can be addressed by future researchers include:

1. Application of the criteria to interpretive guides could be improved by specifying more clearly how well an interpretive guide must address each criterion before it is considered to have met the criterion.
2. Having a small number of subject matter experts to both review the criteria and the resulting findings.
3. Having a small number of coders rate the interpretive guides.
4. Not being able to locate an interpretive guide for each test currently being administered.

Conclusions

Based on the research conducted, the criteria listed in Appendix A are appropriate to use to evaluate interpretive guides published for state-level, grade 5, mathematics, customized, criterion-referenced tests. Both the literature and subject matter experts have confirmed this statement through multiple articles over recent years and through a specific analysis of each criterion. Subject matter experts have also confirmed that the use of the *Standards* as the supporting document for the criteria was appropriate even thought the standards are 8 years old. Not only were the *Standards* found to be rigorous, they were also found to be current.

It is the conclusion of this research that interpretive guides published for state-level, grade 5, mathematics, customized, criterion-referenced tests that were administered during the 2005-06 academic year do not meet criteria based on the *Standards*. Subject

matter experts agreed that there are certain criteria that must be met by all interpretive guides, and that there is a minimum number of criteria that interpretive guides should meet. However, there was disagreement among the subject matter experts as to whether the interpretive guides are rigorous enough. The conclusion that interpretive guides are not rigorous enough is based upon:

- the variability in the number of criteria met by interpretive guides,
- the number of interpretive guides meeting less than 50% of the criteria, and
- the number of criteria met by less than 50% of the interpretive guides

Because on this information and the comments from subject matter experts, it was found that the vast majority of the interpretive guides provided insufficient information upon which to make valid interpretations.

Criterion #13 states, "The interpretive guide states how the test scores of individuals should be interpreted." (p. 65). Criterion #13 is often the main purpose of interpretive guides. The coders revealed that only 62% of the interpretive guides stated how the test scores of individuals should be interpreted. This is further confirmation that interpretive guides are not providing sufficient information upon which to make valid interpretations.

The lack of research about interpretive guides for NCLB tests is concerning, as is the lack of overall research pertaining to other ancillary documents that support such tests. The methodology of this research provides one of several ways to generate criteria to evaluate ancillary materials developed for standardized tests.

As results from NCLB tests are used increasingly for accountability purposes in the future, scrutiny of the ability of the tests to assist teacher interventions will increase.

Interpretive guides are one method of sharing information with teachers and parents about how to interpret test data; possibly leading to interventions for students. While interpretive guides are not the only method for disseminating the information, many states use this method. Interpretive guides, at a minimum, must contain the information needed by parents and teachers to make informed and valid interpretations about student test results. Currently, the majority of interpretive guides lack this information.

Anticipated Outcomes

This study will provide information about the overall content of interpretive guides so that they may be evaluated and compared to each other. In addition, this study will provide the developers of interpretive guides criteria from which to determine whether their interpretive guides meet the breadth of what is expected in the *Standards*. By comparing the contents of interpretive guides across states, further clarity may be brought to the criteria; thus suggesting a sort of minimum standard on which to evaluate future interpretive guides. This would then lead to improvements to interpretive guides across states. Also, there is the possibility that this project may be referenced when the *Standards* are revised, to determine whether the *Standards* are clear enough concerning what should be contained in interpretive guides.

Suggestions for Further Research

This research generated criteria that can be used to determine if interpretive guides meet certain standards. Additional research should be conducted to determine if any additional criteria are needed. For example, many of the interpretive guides provided contact information for state department employees. Also, many interpretive guides stated that group scores would be reported only if the sample count was above a certain

number. It is unclear whether such criterion should be listed. Lastly, criterion #1 was not addressed because there was not enough time or information to determine if an interpretive guide met the criterion. Further research should investigate if criterion #1 is appropriate and/or is being met by interpretive guides.

It has been shown clearly in this research that the criteria have been applied in a non-rigorous manner. Further research should focus on individual criterion to determine how rigorous interpretive guides must be to be considered as having 'met' the criterion.. For example, criterion #3 states that, "The interpretive guide cautions against anticipated misuses of test results." (p. 55). Should there be a certain number of misuses listed? Are there specific misuses that should be listed depending on how the test was built? There is much to explore in each criterion.

Lastly, it is unknown what percent of parents and teachers actually obtain, read, and apply the information contained in interpretive guides. Research should be conducted to determine this information, as well as to determine if the interpretive guides are effective in relaying their messages.

References

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

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *The Standards for Competence in the Educational Assessment of Students.* Retrieved January 18, 2007, from http://www.unl.edu/buros/bimm/html/article3.html

Amrein, A.L., & Berliner, D.C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, *10*(18). Retrieved February 7, 2007, from http://epaa.asu.edu/epaa/v10n18/.

Barton, P.E., & Coley, R.J. (1994). *Testing in America's schools: Policy information report.* Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED366616)

Bauer, S.C. (2000). Should achievement tests be used to judge school quality? *Education Policy Analysis Archives*, *8*(46). Retrieved October 29, 2006, from http://epaa.asu.edu/epaa/v8n46/.

Bettesworth, L.R. (2007, April). *Data analysis and interpretation used to inform decisions in school*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois.

Brookhart, S. M. (2001, March). *The "standards" and classroom assessment research*. Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, Dallas, TX.

Brown, D. F. (1993, April). *The political influence of state testing reform through the eyes of principals and teachers*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Camara, W.J., & Lane, S. (2006). A historical perspective and current views on the Standards for Educational and Psychological Testing. *Educational Measurement: Issues and Practice, 25*(3), 35–41.

Chester, M.D. (2005). Making valid and consistent inferences about school effectiveness from multiple measures. *Educational Measurement: Issues and Practice*, *18*(4), 40–51.

Cimbricz, S. (2002). State-mandated testing and teachers' beliefs and practice. *Education Policy and Analysis*, 10(2). Retrieved October 29, 2006, from http://epaa.asu.edu/epaa/v10n2.html

Darling-Hammond, L. (1997). Using standards and assessments to support student learning. *Phi Delta Kappan*, *79*(3), 190–199.

Darling-Hammond, L., & Wise, A. E. (1985). Beyond standardization: State standards and school improvement. *The Elementary School Journal*, *85*(3), 315-336.

Grant, S. G. (2000). Teachers and tests: Exploring teachers' perceptions of changes in the New York state testing program. *Education Policy and Analysis*, 8(14). Retrieved October 29, 2006, from http://epaa.asu.edu/epaa/v8n14.html

Gullickson, A. R. (1986). Teacher education and teacher-perceived needs in educational measurement and evaluation. *Journal of Educational Measurement*, *23*(4), 347–354.

Harris, W.G. (2006). The challenges of meeting the *Standards*: a perspective from the test publishing community. *Educational Measurement: Issues and Practice*, *25*(3), 42–45.

Ingram, D., Louis, K. S., & Schroeder, R. G.(2004). Accountability policies and teacher decision making: Barriers to the use of data to improve practice. *Teachers College Record*, *106*(6), 1258–1287.

Jones, B.D., & Egley, R.J. (2004). Voices from the frontlines: Teachers' perceptions of high-stakes testing. *Education Policy Analysis Archives*, *12*(39). Retrieved February 7, 2007, from http://epaa.asu.edu/epaa/v12n39/.

Kearney, P.C. (1983). Uses and abuses of assessment and evaluation data by policymakers. *Educational Measurement: Issues and Practice*, *2*(3), 9–12.

Klein, S.P., Hamilton, L.S., McCaffrey, D.F., & Stecher, B.M. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives*, *8*(49). Retrieved February 7, 2007, from http://epaa.asu.edu/epaa/v8n49/.

Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources*, *37*(4), 752–777.

Koretz, D. (2006). Steps toward more effective implementation of the Standards for Educational and Psychological Testing. *Educational Measurement: Issues and Practice, 25*(3), 46–50.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for

categorical data. *Biometrics, 33*, 159-174.

Mabry, L., & Margolis, J. (2006). NCLB: Local implementation and impact in southwest Washington state. *Educational Policy Analysis Archives*, *14*(23). Retrieved February 7, 2007, from http://epaa.asu.edu/epaa/v14n23/.

McCoy, J.D. (2007, April). *School improvement test members' preparedness for data-driven decision making*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Mertler, C.A. (2002). *Using standardized test data to guide instruction and intervention. ERIC digest.* Washington DC: ERIC Clearinghouse on Assessment and Evaluation. (ERIC Document Reproduction Service No. ED470589)

Mertler, C. A., & Campbell, C. (2005, April). *Measuring teachers' knowledge & application of classroom assessment concepts: Development of the "Assessment Literacy Inventory"*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

National Archives and Records Administration. (2002). *No Child Left Behind Act of 2001* (Public Law 107-110). Washington, DC: U.S. Government Printing Office.

Natriello, G., & Pallas, A. M. (1999). *The development and impact of high stakes testing*. (Report No. TM031516). (ERIC Document Reproduction Service No. ED443871)

Nolen, S. B., Haladyna, T. M., & Haas, N. S. (1992). Uses and abuses of achievement test scores. *Educational Measurement: Issues and Practice*, *11*(2), 9–15.

O'Sullivan, R.G., & Johnson, R.L. (1993, April). *Using performance assessments to measure teachers'competence in classroom assessment*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Plake, B.S., Impara, J.C., & Fager, J.J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice, 12*(4), 10–12.

Popham, J.W. (1999). Why standardized tests don't measure educational quality. *Educational Leadership, 56*(6), 8–15.

Popham, J.W. (2004) Why assessment illiteracy is professional suicide. *Educational Leadership*, *62*(1), 82–83.

Quilter, S. M., & Gallini, J. K. (2000). Teachers' assessment literacy and attitudes. *The Teacher Educator*, *36*(2), 115–131.

Rudman, H. C. (1989). Integrating testing with teaching. *Practical Assessment, Research & Evaluation*, *1*(6). Retrieved February 7, 2007, from http://PAREonline.net/getvn.asp?v=1&n=6.

Sanders, W. L., & Horn, S. P. (1995). Educational assessment reassessed: The usefulness of standardized and alternative measures of student achievement as indicators for the assessment of educational outcomes. *Education Policy and Analysis*, *3*(6). Retrieved October 29, 2006, from http://epaa.asu.edu/epaa/v3n6.html

Stiggins, R. (2004). New assessment beliefs for a new school mission. *Phi Delta Kappan*, *86*(1), 22–27.

Tinsley, H.E.A., & Bradley, R.W. (1986). Testing the test: Test interpretation. *Journal of Counseling and Development*, *64*, 462–466.

Yeh, S. S. (2005). Limiting the unintended consequences of high-stakes testing. *Education Policy and Analysis*, *13*(43). Retrieved October 29, 2006, from http://epaa.asu.edu/epaa/v13n43.html

Zancanella, D. (1992). The influence of state-mandated testing on teachers of literature. *Education Policy and Analysis*, *14*(3). Retrieved January 4, 2007, from http://epaa.asu.edu/epaa/v14n3.html.

Appendix A: Criteria for Evaluating Interpretive Guides

The *Standards for Educational and Psychological Testing* (*Standards)* state, "Evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every standard in this document, and acceptability cannot be determined by a checklist" (p. 4). The criteria listed in this appendix are not intended to act as a checklist, nor is it suggested that each criterion be satisfied by each interpretive guide.

Each criterion is accompanied by the *Standard(s)* that support them. Also, an additional description is provided to clarify how each criterion is being applied to interpretive guides. By reading each criterion, the supporting *Standard(s)* and the description, coders will obtain better score reliability. Also, by providing detail for each criterion, future research may support or disintegrate the appropriateness and improve the framework by which we gauge interpretive guides.

These criteria should be employed for their intended use, which is to determine, in a broad sense, how well interpretive guides are complying with the *Standards*. Concluding that a guide complies with the standards if it meets all criterion listed here would be a misinformed statement. This guide does not attempt to provide a rigorous analysis of the content of the interpretive guide, nor should conclusions about the results yield such information.

Criterion 1:
The interpretive guide should be available at the time when the test is published or at the time when the test is released for use.

Based on *Standard(s)*:
6.1 Test documents (e.g, test manuals, technical manuals, user's guides, and supplemental material) should be made available to prospective test users and other qualified persons at the time a test is published or released for use.

Description:
This criterion is satisfied when the interpretive guide is made available at or before the test results and/or data are in the hands of the test users.

Criterion 2:
The interpretive guide should be written clearly for the intended reader.

Based on *Standard(s)*:
5.10 When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common mis-interpretations of test scores, and how scores will be used.
6.2 Test documents should be complete, accurate, and clearly written so that the intended reader can readily understand the content.
6.8 If a test is designed to be scored or interpreted by test takers, the publisher and test developer should provide evidence that the test can be accurately scored or interpreted by the test takers. Test that are designed to be scored and interpreted by the test taker should be accompanied by interpretive materials that assist the individual in understanding the test scores and that are written in language that the test taker can understand.

Description:
This criterion is satisfied when the guide is written for a reader who does not have specialized knowledge or background with assessments. The guide may also be written on from both a conceptual and a technical standpoint to allow understanding for the largest amount of readers.

Criterion 3:
The interpretive guide cautions against anticipated misuses of test results.

Based on *Standard(s)*:
5.10 When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common mis-interpretations of test scores, and how scores will be used.
6.3 The rational for the test, recommended uses of the test, support for such uses, and information that assists in score interpretation should be documented. Where particular misuses of a test can be reasonably anticipated, cautions again such misuses should be specified.
13.15 In educational settings, reports of group differences in test scores should be accompanied by relevant contextual information, where possible, to enable meaningful interpretation of these differences. Where appropriate contextual information is not available, users should be cautioned against misinterpretation.

Description:
This criterion is satisfied when general or specific limitations of the data are stated, such that the reader knows that care must be taken using the data for certain purposes. An exhaustive list is not necessary, but at least one specific interpretation that should be avoided should be clearly stated.

Criterion 4:
The interpretive guide identifies and describes what the test is assessing.

Based on *Standard(s)*:
5.10 When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common mis-interpretations of test scores, and how scores will be used.
6.6 When a test relates to a course of training or study, a curriculum, a textbook, or packaged instruction, the documentation should include an identification and description of the course or instructional materials and should indicate the year in which these materials were prepared.

Description:
This criterion is satisfied when a list of standards which are being assessed is given, or when the website location of the standards is given.

Criterion 5:
The interpretive guide states what test year(s) that the interpretive guide is appropriate to use for.

Based on *Standard(s)*:
6.6 When a test relates to a course of training or study, a curriculum, a textbook, or packaged instruction, the documentation should include an identification and description of the course or instructional materials and should indicate the year in which these materials were prepared.
6.14 Every test form and supporting document should carry a copyright date or publication date.
13.16 In educational settings, whenever a test score is reported, the date of test administration should be reported. This information and the age of any norms used for interpretation should be considered by test users in making inferences.

Description:
This criterion is satisfied when the interpretive guide clearly states what test year(s) the guide is appropriate for. This information should not have to be inferred from or assumed by the copyright year of the document.

Criterion 6:
The interpretive guide states what qualifications of individuals are required to interpret the test scores accurately.

Based on *Standard(s)*:
6.7 Test documents should specify qualifications that are required to administer a test and to interpret the scores accurately.
13.12 In educational settings, those who supervise others in test selection, administration, and interpretation should have received education and training in testing necessary to ensure familiarity with the evidence for validity and reliability for tests used in the educational setting and to be prepared to articulate or to ensure that others articulate a logical explanation of the relationship among the tests used, the purposes they serve, and the interpretations of the test scores.
13.13 Those responsible for educational testing programs should ensure that the individuals who interpret the test results to make decisions within the school context are qualified to do so or are assisted by and consult with qualified persons.

Description:
This criterion is satisfied when the interpretive guide states what knowledge, certification, training, or degree is required to interpret the test data. Stating who the guide is intended for (teachers, parents, etc…) does not, on its own, meet this criterion.

Criterion 7:
The interpretive guide references studies and/or support relevant to general and specific uses of the test.

Based on *Standard(s)*:
6.9 Test documents should cite a representative set of the available studies pertaining to general and specific uses of the test.
13.2 In educational settings, when a test is designed or used to serve multiple purposes, evidence of the test's technical quality should be provided for each purpose.
13.9 When test scores are intended to be used as part of the process for making decisions for educational placement, promotion, or implementation of prescribed educational plans, empirical evidence documenting the relationship among particular test scores, the instructional programs, and desired student outcomes should be provided. When adequate empirical evidence in not available, users should be cautioned to weigh the test results accordingly in light of other relevant information about the student.

Description:
This criterion is satisfied when the interpretive guide cites research or other appropriate documentation that is supported by research to support the stated use(s) of the test.

Criterion 8:
The interpretive guide states, for tests where different administration methods are used, the interchangeability of test results and guidance for interpreting test results.

Based on *Standard(s)*:
6.11 If a test is designed so that more than one method can be used for administration or for recording responses – such as marking responses in a test booklet, on a separate answer sheet, or on a computer keyboard – then the manual should clearly document the extent to which scores arising from these methods are interchangeable. If the results are not interchangeable, this fact should be reported, and guidance should be given for the interpretation of scores obtained under the various conditions or methods of administration.
7.8 When scores are disaggregated and publicly reported for groups identified such as gender, ethnicity, age, language proficiency, or disability, cautionary statements should be included whenever credible research reports that test scores may not have comparable meaning across these different groups.

Description:
This criterion is satisfied when the interpretive guide clearly states whether or not scores across subgroups can be compared directly, or cautions against such comparisons.

Criterion 9:
The interpretive guide states general information that can be used to determine appropriate uses for the test in specific contexts.

Based on *Standard(s)*:
5.10 When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common mis-interpretations of test scores, and how scores will be used.
6.15 Test developers, publishers, and distributors should provide general information for test users and researchers who may be required to determine the appropriateness of an intended test use in a specific context. When a particular test use cannot be justified, the response to an inquiry from a prospective test user should indicate this fact clearly. General information also should be provided to test takers and legal guardians who must provide consent prior to a test's administration.

Description:
This criterion is satisfied when general information about the test is given such as: administration setting, testing time, accommodations offered, test blueprints, etc. Some or all of this information may be listed in the guide, or a website location may be given where the information can be found.

Criterion 10:
The interpretive guide states whether the time, date, and/or context in which the test was given should be considered when interpreting the test results.

Based on *Standard(s)*:
7.5 In testing applications involving individualized interpretations of test scores other than selection, a test taker's score should not be accepted as a reflection of standing on the characteristic being assessed without consideration of alternate explanations for the test taker's performance on that test at that time.
13.11 In educational settings, test users should ensure that any test preparation activities and materials provided to students will not adversely affect the validity of test score inferences.

Description:
This criterion is satisfied when the interpretive guide clearly states whether or not information such as time, date, and/or context that the test was given will impact the interpretation of test scores.

Criterion 11:
The interpretive guide states the intended uses of the test results.

Based on *Standard(s)*:
5.10 When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common mis-interpretations of test scores, and how scores will be used.
13.1 When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test user.

Description:
This criterion is satisfied when the interpretive guide clearly states how the test scores are meant to be used in a general sense. This information is provided at a high level, and specific uses do not have to be given to satisfy this criterion.

Criterion 12:
The interpretive guide states the degree of measurement error in an individual's score.

Based on *Standard(s)*:
5.10 When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common mis-interpretations of test scores, and how scores will be used.
13.14 In educational settings, score reports should be accompanied by a clear statement of the degree of measurement error associated with each score or classification level and information on how to interpret the scores.

Description:
This criterion is satisfied when the degree of measurement error is stated in the text of the interpretive guide, or there is information describing where this information can be found on the score reports. Information may be given as to where else the degree of measurement error can be found, such as a technical manual or a website.

Criterion 13:
The interpretive guide states how the test scores of individuals should be interpreted.

Based on *Standard(s)*:
5.10 When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common mis-interpretations of test scores, and how scores will be used.
13.14 In educational settings, score reports should be accompanied by a clear statement of the degree of measurement error associated with each score or classification level and information on how to interpret the scores.

Description:
This criterion is satisfied when specific decision(s) that can be made based upon the data are given. This criterion is not met by explaining where the student's score can be found on the score report.

Criterion 14:
The interpretive guide provides relevant contextual information for interpreting differences between groups.

Based on *Standard(s)*:
13.15 In educational settings, reports of group differences in test scores should be accompanied by relevant contextual information, where possible, to enable meaningful interpretation of these differences. Where appropriate contextual information is not available, users should be cautioned against misinterpretation.

Description:
This criterion is satisfied when a description of subgroups, accommodations given to different subgroups, test settings unique to certain subgroups, or other information that assists in interpreting the data for subgroups is provided. Information may be given as to where else the subgroup information can be found, such as a technical manual or a website.

Criterion 15:
The interpretive guide states the degree of measurement error in score reports for groups.

Based on *Standard(s)*:
5.10 When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common mis-interpretations of test scores, and how scores will be used.
13.14 In educational settings, score reports should be accompanied by a clear statement of the degree of measurement error associated with each score or classification level and information on how to interpret the scores.

Description:
This criterion is satisfied when the degree of measurement error for groups is stated in the text of the interpretive guide, or there is information describing where this information can be found on the score reports. Information may be given as to where else the degree of measurement error for groups can be found, such as a technical manual or a website.

Appendix B: Interpretive Guides Reviewed

Alabama

Alabama Department of Education. (n.d.). *2006 interpretive guide Alabama accountability system*. Retrieved December 24, 2006, from ftp://ftp.alsde.edu/documents/100/6-C%20ACCOUNTABILITY%202006%20Interpretive%20Guide%20Chapter%20Format.pdf

Alaska

Alaska comprehensive system of student assessment (CSSA) guide to test interpretation for the grade 5 standards based assessments for parents and students spring 2006. (n.d.). Retrieved February 3, 2007, from http://www.eed.state.ak.us/tls/assessment/sba/Spring06/GTIs/8x11Spring2006PandSGTIgrade5.pdf

Arizona

Arizona Department of Education, CTB/McGraw-Hill. (2006). *Grades 3 through 8 guide to test interpretation Arizona's instrument to measure standards AIMS DPA spring 2006*. Retrieved December 4, 2006, from http://www.ade.az.gov/standards/aims/Administering/

Arkansas

Arkansas Department of Education. (n.d.). *Report Interpretation Guide Benchmark Examinations Grades 3 - 8 February and March 2006 Administration*. Retrieved December 4, 2006, from http://arkedu.state.ar.us/actaap/pdf/rig_benchmark_spr06.pdf

California

California Department of Education. (n.d.). *Standardized Testing and Reporting (STAR) Program Explaining 2006 Reports to Parents and Guardians*. Retrieved December 4, 2006, from http://www.cde.ca.gov/ta/tg/sr/documents/explainstarprnts.pdf

California Department of Education. (n.d.). *Standardized Testing and Reporting (STAR) Program Explaining 2006 Internet Reports to the Public Information for Counties, School Districts, and Schools August 2006*. Retrieved December 4, 2006, from http://www.cde.ca.gov/ta/tg/sr/documents/explintrrpts06.pdf

Colorado

Colorado Department of Education, CTB/McGraw-Hill. (2006). *Colorado student assessment program updated through 2006 guide to test interpretation grades 3 – 8*. Retrieved December 4, 2006, from http://www.cde.state.co.us/cdeassess/documents/csap/2006/GR3-8GTI_CSAP2006.pdf

Connecticut

Connecticut State Board of Education. (n.d.). Connecticut mastery test fourth generation 2006 interpretive guide. Retrieved March 5, 2007, from http://www.csde.state.ct.us/public/cedar/assessment/cmt/resources/misc_cmt/2006_cmt_interpretive_guide.pdf

Delaware

Delaware student testing program grade 5 Your student's test results. (n.d.). Retrieved December 4, 2006, from http://www.doe.state.de.us/AAB/DE06_spg_IF_AllGrades_FinalDH.pdf

Florida

State of Florida, Department of State. (2006). *Understanding FCAT reports 2006*. Retrieved December 4, 2006, from http://www.firn.edu/doe/sas/fcat/pdf/ufr_06.pdf

Georgia

Georgia Department of Education. (n.d.). *2006 CRCT score interpretive guide*. Retrieved September 17, 2006, from http://www.doe.k12.ga.us/ci_testing.aspx?PageReq=CI_TESTING_CRCT

Illinois

Illinois State Board of Education, the Grow Network/McGraw-Hill. (2006). *Illinois student report*. Retrieved December 8. 2006, from http://www.isbe.state.il.us/assessment/pdfs/ISR_Grades_3_5_6_8.pdf

Indiana

State of Indiana Department of Education, CTB/McGraw-Hill. (2005). *Guide to test interpretation*. Retrieved December 8, 2006, from http://www.doe.state.in.us/istep/pdf/GTI/47024W_1stEdGTI_F05IN.pdf

Kansas

Center for Educational Testing and Evaluation, the University of Kansas. (2006, August). *Score report description guide to assist the review of results from the spring 2006 Kansas assessments in reading and mathematics*. Lawrence, KS.

Kentucky

Kentucky Department of Education. (n.d.). *2006 CATS interpretive guide*. Retrieved February 3, 2007, from http://education.ky.gov/NR/rdonlyres/385CAFE2-CB5D-4C59-9312-84645B7F1CDA/0/2006CATSInterpretiveGuide.pdf

Louisiana

Louisiana Department of Education. (2006). *Interpretive guide*. Retrieved September 17, 2006, from http://www.doe.state.la.us/lde/uploads/9725.pdf

Maryland

Analyzing your MSA data. (n.d.). Retrieved February 2, 2007, from http://mdk12.org/data/msa_analyzing/index.asp

Massachusetts

Massachusetts Department of Education. (2006). *Guide to interpreting the spring 2006 MCAS reports*. Retrieved February 3, 2007, from http://www.doe.mass.edu/mcas/2006/interpretive_guides/full.pdf

Michigan

Michigan Department of Education. (n.d.). *Michigan educational assessment program guide to reports grades 3 - 9 Fall 2006*. Retrieved February 3, 2007, from http://www.michigan.gov/documents/mde/F06_MEAP_GTR_3-9_FINAL_181509_7.pdf

Minnesota

Minnesota Department of Education. (n.d.). *Minnesota interpretive guide*. Retrieved December 12, 2006, from http://education.state.mn.us/MDE/Accountability_Programs/Assessment_and_Testing/Assessments/MCA_II/Reports/index.html

Mississippi

Mississippi Department of Education, CTB/McGraw-Hill. (2006). *Mississippi grade level testing program Mississippi curriculum test interpretive guide for teachers and administrators 2006 grades 2-8*. Retrieved December 4, 2006, from http://www.mde.k12.ms.us/acad/osa/Score_Interpretation_Guide_Small.pdf

Missouri

Missouri Department of Elementary and Secondary Education, CTB/McGraw-Hill. (2006). *Missouri assessment program guide to interpreting results communication arts and mathematics revised 2006*. Retrieved December 12, 2006, from http://dese.mo.gov/divimprove/assess/2006_gir.pdf

Montana

Montana Office of Public Instruction. (n.d.). *MontCAS, phase 2 guide to interpreting the 2006 criterion-referenced test and CRT-alternate assessment reports*. Retrieved December 12, 2006, from http://www.opi.mt.gov/PDF/Assessment/CRT/06InterpGuide.pdf

NECAP (New Hampshire, Rhode Island, Vermont)

The New England Common Assessment Program Guide to Using the 2005 NECAP Reports. (n.d.). Retrieved December 23, 2006, from http://education.vermont.gov/new/pdfdoc/pgm_assessment/necap/interpretive_guide_05_supplement.pdf

New Jersey

New Jersey Department of Education. (2006). *March 2006 grade eight proficiency assessment (GEPA) score interpretation manual June 2006*. Retrieved December 12, 2006, from http://www.state.nj.us/njded/assessment/ms/gepa_score_interp_manual.pdf

New Mexico

NMHSSA spring 2006 score report interpretive guide. (n.d.). Retrieved December 12, 2006, from http://sde.state.nm.us/div/acc.assess/assess/dl/NMHSSA/NMHSSA_Interpretive_Guide_2006.pdf

New York

The Grow Network/McGraw-Hill. (2006). *Welcome to the grow parent website*. Retrieved February 2, 2007, from http://www.nysparents.com/nys/

North Carolina

NC School report card resources. (n.d.). Retrieved February 2, 2007, from http://www.ncschoolreportcard.org/src/resources.jsp?pYear=2005-2006

North Dakota

North Dakota Department of Public Instruction. (n.d.). *Understanding student achievement within the North Dakota state assessment*. Retrieved December 12, 2006, from http://www.dpi.state.nd.us/testing/assess/understand0406.pdf

Ohio

Ohio Department of Education. (2006). *Ohio achievement tests grades 3 - 8 family report interpretive guide spring 2006*. Retrieved December 12, 2006, from http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=222&ContentID=17597&Content=17597

Oklahoma

Oklahoma State Department of Education. (2006), *Oklahoma School Testing Program 2006 Test Interpretation Manual Grades 3 - 8 Oklahoma Core Curriculum Tests*. Retrieved December 12, 2006, from http://title3.sde.state.ok.us/studentassessment/05-06/3-8/OK-TIM-3-8-Final.pdf

South Carolina

Office of Assessment, South Carolina Department of Education. (n.d.). *PACT score report user's guide*. Retrieved January 30, 2007, from http://ed.sc.gov/agency/offices/assessment/PACT/documents/PACTUserGuide06BW.pdf

South Dakota

Harcourt Assessment, Inc. (2006). *South Dakota state test of educational progress interpretive guide for educators and parents*. Retrieved December 23, 2006, from http://doe.sd.gov/octa/assessment/docs/DakotaSTEPInterpretiveGuide.pdf

Tennessee

State of Tennessee Department of Education, CTB/McGraw-Hill. (2004). Parents' guide to understanding TCAP achievement test results. Retrieved December 23, 2006, from http://tennessee.gov/education/assessment/doc/GTI_for_Parents.pdf

State of Tennessee Department of Education, CTB/McGraw-Hill. (2004).Educators' guide to understanding TCAP achievement test results. Retrieved December 23, 2006, from http://tennessee.gov/education/assessment/doc/GTI_for_Administrators.pdf

Texas

Explanation of TAKS™ results - understanding the confidential student report - a guide for parents. (n.d.). Retrieved February 2, 2007, from http://www.tea.state.tx.us/student.assessment/resources/guides/parent_csr/index.html

TAKS results: what they tell us about our students and programs. (n.d.). Retrieved February 2, 2007, from http://www.tea.state.tx.us/student.assessment/admin/tetn/math.ppt

Student Assessment Division, Texas Education Agency. (n.d.). *Interpreting assessment results: Texas student assessment program*. Retrieved February 2, 2007, from http://www.tea.state.tx.us/student.assessment/resources/guides/interpretive/general_06.pdf

Student Assessment Division, Texas Education Agency. (n.d.).*Texas assessment of knowledge and skills (TAKS) Texas assessment of knowledge and skills–inclusive (TAKS–I)*. Retrieved February 2, 2007, from http://www.tea.state.tx.us/student.assessment/resources/guides/interpretive/TAKS_06.pdf

Virginia

Every child can succeed…A parent's guide to Virginia's standards of learning program. (n.d.). Retrieved January 25, 2007, from http://www.pen.k12.va.us/VDOE/Parents/parentshandbook.pdf

Washington

2006 reaching higher. (n.d.). Retrieved December 23, 2006, from http://www.k12.wa.us/Assessment/pubdocs/2006ReachingHigher3-8.pdf

West Virginia

The Office of Student Assessment Services, West Virginia Department of Education. (n.d.). *Using Assessment Results Systemic Continuous Improvement Process (SCIP)*. Retrieved December 24, 2006, from http://osa.k12.wv.us/pdf/usingassessmentresultsscip.pdf

Wisconsin

Wisconsin Department of Education, CTB/McGraw-Hill. (2006). *Administrator's interpretive guide*. Retrieved December 24, 2006, from http://dpi.wi.gov/oea/pdf/adminguide06.pdf

Wyoming

Wyoming Department of Education. (2005). *Proficiency assessments for Wyoming students*. Retrieved February 2, 2007, from http://www.k12.wy.us/SAA/Paws/math.htm

Appendix C: Interpretive Guides Survey

Directions: Please review the document titled "Interpretive Guides," which contains preliminary findings, along with the criteria used to evaluate the interpretive guides. Then, complete the 5 questions below. Answer each question by placing an "X" under the response option that *BEST* describes your level of agreement or disagreement with the statement. Also, please add any clarifying comments or concerns under each statement.

1. There are certain criteria that all interpretive guides must meet.

Strongly Disagree	Disagree	Agree	Strongly Agree
☐	☐	☐	☐

Please put additional comments or concerns here:

2. There are a minimum number of criteria that interpretive guides must meet.

Strongly Disagree	Disagree	Agree	Strongly Agree
☐	☐	☐	☐

Please put additional comments or concerns here:

3. The criteria being used to evaluate the interpretive guides is based on standards that are outdated.

Strongly Disagree	Disagree	Agree	Strongly Agree
☐	☐	☐	☐

Please put additional comments or concerns here:

4. Typically, interpretive guides are the only method by which state departments disseminate the information stated in the criteria.

Strongly Disagree	Disagree	Agree	Strongly Agree
☐	☐	☐	☐

Please put additional comments or concerns here:

5. Based on data collected for this research, interpretive guides currently are being written adequately for the tests and audiences for whom they were designed.

Strongly Disagree	Disagree	Agree	Strongly Agree
☐	☐	☐	☐

Please put additional comments or concerns here: