RUNNING HEAD: A Generalizability Investigation of Cognitive Demand and Rigor

A Generalizability Investigation of Cognitive Demand and Rigor Ratings of Items and Standards in an Alignment Study

Allison Lombardi, PhD

Mary Seburn, PhD

David Conley, PhD

Eric Snow, PhD

Educational Policy Improvement Center

720 E. 13th Ave., Suite 202

Eugene, OR 97401

541-346-6153

allison_lombardi@epiconline.org

allisonl@uoregon.edu

Presented at the annual conference of the American Educational Research Association

Denver, CO

April 2010

Abstract

In alignment studies, expert raters evaluate assessment items against standards and ratings are used to compute various alignment indices. Questions about rater reliability, however, are often ignored or inadequately addressed. This paper reports the results of a generalizability theory study of cognitive demand and rigor ratings of assessment items and college-readiness standards in the context of an alignment study of college admissions tests to a set of college readiness standards. Results indicate a higher level of generalizability for Math item and standard ratings than for English item and standard ratings, as well as a higher level of generalizability for cognitive demand ratings than for rigor ratings. Results also suggest that the standard of 5-6 raters in alignment studies may be insufficient for obtaining desired reliability. These findings may be used to carefully plan more robust alignment studies in the future so that higher levels of reliability across raters will be attained.

A Generalizability Investigation of Cognitive Demand and Rigor Ratings of Items and Standards in an Alignment Study

In a measurement context where expert raters are used to evaluate student performance questions, the reliability of ratings is paramount and is investigated using empirical evidence. In alignment studies, however, where expert raters are used to evaluate assessment items against a set of standards, basic questions about rater reliability are addressed either partially or not at all (Herman, Webb & Zuniga, 2007; D'Agostino, et. al., 2008). Most alignment studies use between three and ten expert raters, but the extent to which the ratings are generalizable across raters and the influence of disagreements in ratings on alignment conclusions are often not critically examined (e.g., Achieve, Inc., 2007; Webb, 1997, 1999, 2002; Porter, 2002).

Generalizability theory is particularly useful in the context of alignment studies, as it provides a model for disentangling and identifying the multiple sources of error that may influence the consistency of ratings. That is, generalizability theory describes the amount of error that is attributable to raters and describes the extent to which the ratings generalize beyond the individual raters to the intended construct domain. The error attributable to raters should be as small as possible. When it is not, it implies that ratings were not consistently made across raters. Generalizability Theory is used to conduct a *G Study*, where G coefficients resulting from analyses conducted on objects and facets of measurement are used to evaluate reliability. G Study results are then used in a Decision Study (or D Study) to forecast the G coefficients that would be obtained with varying numbers of raters (Mushquash & O'Connor, 2006).

Aligning College Readiness Standards to College Admission and Placement Tests

The number of college freshmen requiring remedial education ranges from 30% to 60% (NCES, 2004), suggesting a gap between what is taught in high school and what students are

expected to know and do in college. A student may be *college-eligible*—that is, able to meet college admissions requirements—without being *college ready*—able to enroll and succeed in credit-bearing general education courses at the postsecondary level without remediation (Conley, 2005, 2007, 2010). Typically, college admissions and placement assessments are used to measure college readiness and scores from these assessments often determine whether a student takes a remedial course prior to enrollment in a credit-bearing course. At the same time, in an attempt to address the growing problem of high school graduates requiring remediation, some states, such as Texas, have developed and adopted career and college readiness standards meant to guide educators in developing curriculum and assessments with the underlying goal of college and career readiness for all graduating students. In light of these newly developed standards, it is important to determine the degree of alignment to widely-used college admissions and placement tests. However, evidence from content analysis and alignment studies (e.g., Conley, 2003; Conley & Brown, 2007; Brown & Niemi, 2007; Achieve, Inc., 2007) suggest these assessments may not be aligned strongly enough with college-readiness standards to be useful tools for providing feedback to high-school students and teachers concerning college readiness or remediation needs. For this reason, it is imperative that we critically examine the alignment between these assessments and college readiness standards (AERA, APA & NCME, 1999). Such critical examination requires a more robust study of rater reliability in the context of alignment studies.

The purpose of this study was to examine the generalizability of ratings used to compute various alignment indices in the context of a broader alignment study. Raters were trained to make expert judgments concerning the rigor and cognitive demand of a sample of items from six

college admission and placement tests, as well as a validated set of college readiness standards. This study addressed the following research questions:

1. To what extent are cognitive demand and rigor ratings of assessment items generalizable across raters?
2. To what extent are cognitive demand and rigor ratings of standards generalizable across raters?
3. To what extent does the generalizability of cognitive demand and rigor ratings differ for items and standards?
4. What is the ideal number of raters needed to maximize the generalizability of rigor and cognitive demand ratings for items and standards?

Methods

This study employed Generalizability Theory to conduct a G-Study and D-Study in order to address the above research questions in the context of a broader alignment study between college admission and placement test items and a set of college readiness standards. Raters rated the degree of alignment on test items and standards according to two rating scales: Cognitive Demand and Rigor.

Generalizability of Ratings

We investigated the reliability of the cognitive demand and rigor ratings by conducting a generalizability theory analysis (Shavelson and Webb, 1991). Specifically, we used a design with items crossed by raters (i x r design) and standards crossed by raters (s x r design) in which the sources of variation were treated as random. Because our intent was to measure the rigor and cognitive demand of items and standards, the items and standards were the objects of measurement and the raters were considered the sources or error (or facet in g-theory

terminology). We report the phi coefficient, called the index of dependability (Shavelson and Webb, 1991), which can be considered as a reliability-like coefficient for absolute decisions (Herman, Webb & Zuniga, 2007; Thompson, 2003). We focus on reporting reliability for absolute decisions as opposed to relative decisions because the primary interest is to identify the absolute level of cognitive demand and rigor of an item or standard rather than to rank order the items or standards. We also report the absolute error variance associated with each phi coefficient, as it indicates the overall consistency of item and standard ratings across raters.

In the larger alignment study, we used the ratings of six experts as the basis for our alignment computations and decisions. Six is the standard number of reviewers required in similar alignment studies to obtain sufficient reliability (Herman, Webb and Zuniga, 2005; Webb 1997, 1999, and 2002). Consistent with our treatment of rigor ratings and cognitive demand ratings as quasi-quantitative measures (rather than as strictly categorical measures), we conducted a generalizability analysis with items and standards crossed by raters (Shavelson and Webb, 1991). Other measures of agreement are appropriate for categorical ratings (e.g., kappa coefficients, percent agreement, etc.). We used the results of the generalizability theory analysis in a decision study to determine the ideal number of raters needed to maximize the generalizability of item and standard ratings. Coefficients and error variances were calculated using SPSS version 16.0 (SPSS, Inc., 2007; Mushquash & O'Connor, 2006).

Rating Scales: Cognitive Demand & Rigor

We used the first four levels of Marzano's (2001) taxonomy - retrieval, comprehension, analysis, and knowledge utilization - as the basis for the cognitive demand scale. Under this taxonomy, cognitive demand is defined as the level of information processing and the degree of conscious thought needed to complete a task. In our case, the object of measurement was

represented by an assessment item or standard (rather than a task). One characteristic of the taxonomy is that each level builds off of the prior one and each requires a higher degree of cognitive processing than the previous one; as such, we can use these levels to create a set of ordered categories for scoring purposes. The rating scale, which ranges from 1 (lowest) to 4 (highest), employs the following definitions:

1 = **Retrieval:** Recognizing, recalling, executing
2 = **Comprehension:** Integrating and symbolizing
3 = **Analysis:** Matching, classifying, analyzing errors, generalizing, specifying
4 = **Knowledge utilization:** Decision making, problem solving, experimenting, and investigating

Rigor differs from cognitive demand in that it focuses not only on the mental activity required to answer an item successfully or to perform the expectation stated in the standards, but on the relative challenge and difficulty of doing so. Entry-level college expectations for students serve as the point of reference. The scale, which ranges from 1 (lowest) to 3 (highest), employs the following definitions:

1 = ***Below*** *the level at which an entry-level college student should perform*
2 = ***At*** *the level at which an entry-level college student should perform*
3 = ***Above*** *the level at which an entry-level college student should perform*

Background of Alignment Study

The current Generalizability Study was conducted in the context of a larger alignment study. Researchers have developed several methods for evaluating alignment between assessment items and educational standards (Rothman, 2003; Webb, 1997, 1999; Porter, 2002, ACHIEVE, 2007; Wixson, Fisk, Dutro, McDaniel, 2002). In the alignment study, we

implemented a modified version of Webb's methodology (Webb, 1997, 1999) that focused on the use of item and standard ratings to compute three commonly used alignment metrics: categorical concurrence, depth-of-knowledge consistency, and range of knowledge.

Assessments. This study included operational items from a combination of six mathematics and English/Language Arts college admissions and placement assessments. Test developers provided item sets for use in the study that were representative of the test specifications. Some item sets consisted of intact forms, and others were sampled from item banks, depending on whether the test was paper and pencil (fixed form) or computer adaptive where there are no identifiable test forms *per se*. There are other means for determining the alignment of item pools, such as drawing a sample of items where the sample size is determined by the length of the test administered or sampling test content across administrations using multiple tests administered to students. However, most alignment studies comparing the alignment of test forms and test item pools do not specifically address the lack of comparability between the two (Achieve, 2007; Brown & Niemi, 2007).

College readiness standards. The college readiness standards were developed as part of a larger state-wide initiative to improve the alignment between the K-12 and postsecondary systems. These standards describe the content knowledge, thinking skills, and cognitive strategies students need to know to succeed in entry-level postsecondary courses without remediation.

Rater recruitment and training. Six English and six math content area experts were recruited to participate in the alignment study. All raters were active college faculty members at postsecondary institutions from around the U.S. Most had previous experience in the process used for rating assessment items and educational standards. We recruited six raters for each

content area, as previous research in similar alignment studies indicates that six raters are required to obtain sufficient reliability (Herman, Webb and Zuniga, 2005; Webb 1997, 1999, 2002).

The six English and six math raters were trained in content-specific groups during an iterative process using sets of non-operational sample assessment items. The raters convened to review and discuss the standards, items, and rating scale definitions. They first reviewed and rated the standards and a set of sample items individually and then met as a group via teleconference to discuss their ratings and judgments. Through iterative discussion and practice applying the rating scales to multiple sets of sample items and standards, they identified and refined decision rules to apply to consistently make ratings and alignment determinations. As discrepancies arose, the group discussed and reached consensus on a resolution, and then added decision rules that would help them resolve similar discrepancies in future ratings. A team leader was recruited from each group to facilitate consensus and address content and alignment questions from reviewers throughout the study. This process was repeated until the raters agreed on their ratings of the sample assessment item sets and standards.

Rating process. Following the completion of training, the math and English raters accessed the assessment items via a secure online tool that collected their ratings of rigor and cognitive demand. They first rated the standards, then assessment items, providing rigor and cognitive demand ratings for all.

Results and Discussion

Table 1 shows the generalizability results for the cognitive demand and rigor ratings of the Math and English items across assessments. The phi coefficients for the cognitive demand

ratings are close to or above the conventional .80 criterion for reliability (Mushquash & O'Connor, 2006). The phi coefficients for the rigor ratings were lower than those for the cognitive demand ratings, with none of the coefficients reaching the conventional criterion for reliability. These results indicate the six raters reached an acceptable level of dependability for estimating Math and English items' level of cognitive demand, but not for level of rigor. Results also indicate higher reliability for math items in both cognitive demand and rigor ratings.

Table 1

G-Study Coefficients for Math and English Items Across Assessment

		Cognitive Demand		Rigor	
		Coefficients	Error Variance	Coefficients	Error Variance
Subject	Item *N*	Phi	Absolute	Phi	Absolute
Math	1460	0.859	0.035	0.505	0.007
English	1239	0.703	0.053	0.446	0.020

Table 2 shows the generalizability results for the cognitive demand and rigor ratings of the English and Math standards. The phi coefficients for the cognitive demand ratings are close to or above the conventional .80 criterion for reliability. The phi coefficients for the rigor ratings were lower than those for the cognitive demand ratings, with none of the coefficients reaching the conventional criterion for reliability. These results indicate that the six raters did reach an acceptable level of dependability for estimating Math and English standards' level of cognitive demand, but not for level of rigor. Therefore according to these findings, for both items and standards six raters appear to be sufficient for cognitive demand, but insufficient for rigor.

Table 2

G-Study Coefficients for English and Math Standard Ratings

		Cognitive Demand		Rigor	
Subject	Standards *N*	Coefficients Phi	Error Variance Absolute	Coefficients Phi	Error Variance Absolute
Math	115	0.855	0.100	0.566	0.038
English	119	0.724	0.095	0.556	0.060

Overall, these results indicate stronger generalizability across raters for Math item and standard ratings than for English item and standard ratings. Additionally, the results indicate stronger generalizability across raters for cognitive demand ratings than for rigor ratings. Interestingly, Herman, Webb, Zuniga (2005) also reported that ratings of cognitive demand were more reliable than were ratings of centrality (similar to rigor in that centrality evaluated the extent that a standard was essential to a topic).

Table 3 shows the residual effects components for cognitive demand and rigor ratings across assessments and standards. With regard to the estimated variance components for standards and raters, we noticed that the components for the residual effects (standard x rater) on cognitive demand were particularly large relative to the individual (absolute) components for standards (as presented in Table 2). As well, the residual effects components (standard x rater) on rigor were larger than the absolute components, although the discrepancy was not as great. With regard to the estimated variance components for items and raters, the residual effects components on cognitive demand (item x rater) were larger than the individual (absolute) components (as presented in Table 1), and the residual effects components (item x rater) on rigor were slightly larger than the individual (absolute) components.

These findings show there are greater differences in residual effects in cognitive demand than rigor, and there are greater differences in standards than items. Further, these findings suggest larger interaction effects for cognitive demand and standards (i.e., raters rank-ordered items and standards differently on rigor and cognitive demand), and/or other sources of error variability not captured in our study design. Other sources of error variability could be a result of raters rating standards before items, each rater receiving the items in a random order, raters conducting their ratings at different times in different settings, and variation in rater experience writing items and conducting alignment studies.

Table 3

Residual Effects Components for Cognitive Demand and Rigor Ratings Across Items and Standards

	Residual Effects	
Subject	Cognitive Demand	Rigor
Item*Rater		
Math	0.193	0.043
English	0.241	0.075
Standard*Rater		
Math	0.550	0.189
English	0.466	0.257

Figures 1 and 2 present the results of the decision study analysis. Figure 1 shows results for the items. These results indicate that, for cognitive demand ratings of Math items, increasing the number of raters from 6 to between 10 and 15 would result in small gains in reliability. For English items, the same increase in raters would result in moderate gains in reliability, and most importantly, these moderate gains would bring the reliability to the acceptable .80 criterion level.

For rigor ratings of Math and English items, increasing the number of raters from 6 to between 15 and 20 raters would results in phi coefficients approaching the conventional criterion for reliability.

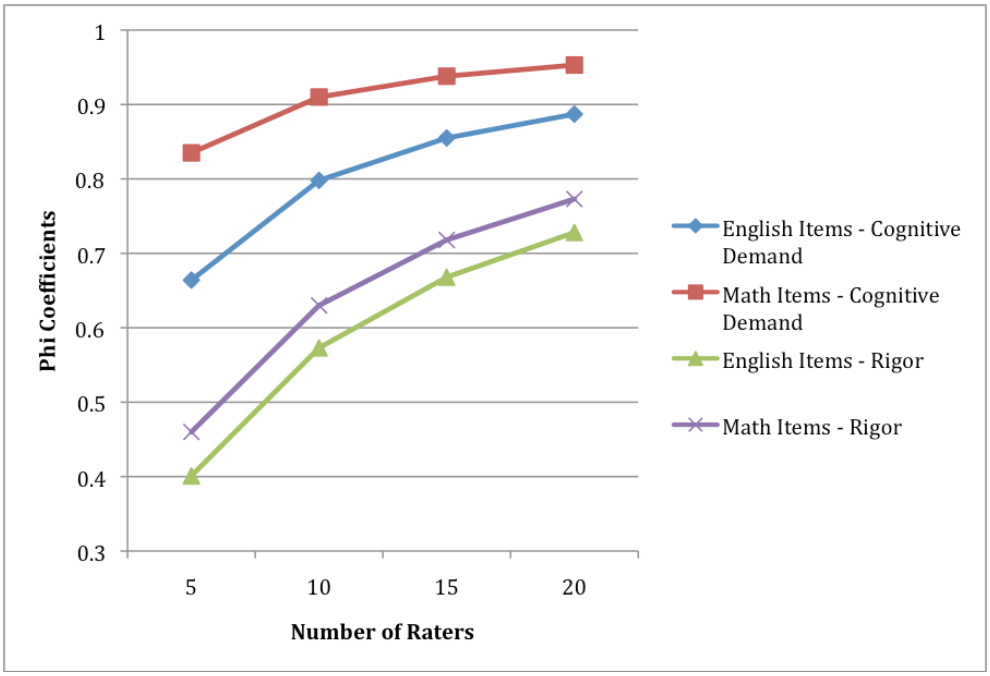

Figure 1. *D-Study results for English and Math items*

Figure 2 shows results for the standards. For Math standards, these results indicate that increasing the number of raters from 6 to between 10 and 15 would result in small gains in reliability in cognitive demand. For English standards, the same increase in raters would result in moderate to large gains in reliability, and the reliability would exceed the acceptable .80 criterion level. For rigor ratings of Math and English standards, increasing the number of raters from 6 to between 15 and 20 raters would results in phi coefficients approaching the conventional criterion for reliability.

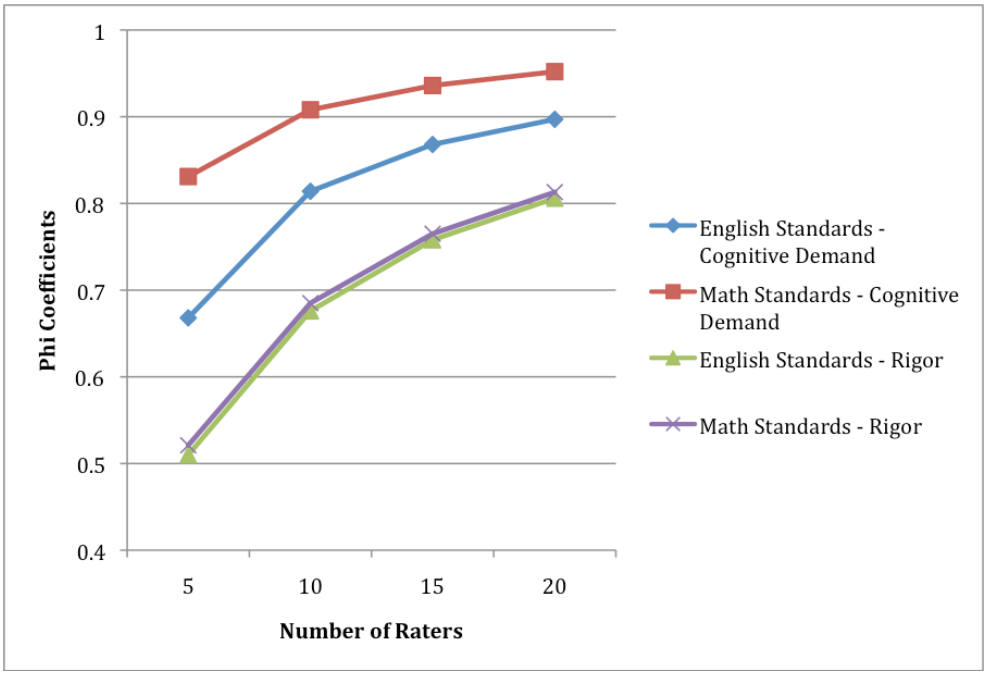

Figure 2. *D-Study results for English and Math standards*

As shown in Figures 1 and 2, rigor ratings for both items and standards would greatly increase if the number of raters were 15 to 20, a notable difference from the suggested 6 (Herman, Webb and Zuniga, 2005; Webb 1997, 1999, 2002). On the other hand, reliability of the cognitive demand ratings would increase in small to moderate amounts if the number of raters were increased from 6 to between 10 and 15. These results indicate more raters are necessary to obtain sufficient reliability in rigor than cognitive demand.

Conclusion and Implications

The purpose of this study was to examine the generalizability of ratings used to compute various alignment indices in the context of a broader alignment study. We examined the extent to which cognitive demand and rigor ratings were generalizable across raters for a sample of college admission and placement test items and a set of validated college readiness standards. In

this examination, we determined if cognitive demand and rigor ratings differed for test items and standards, as well as the ideal number of raters needed to maximize the generalizability of ratings for items and standards.

Several conclusions about rater reliability in alignment studies can be drawn from our findings. First, the primary subject of items and standards rated in alignment studies impact the reliability of ratings. In our findings, Math items and standards showed greater levels of reliability than English items and standards, which is important to consider in designing future alignment studies. Given the objective nature of Math and the more subjective nature of English, it may be easier for raters to agree in alignment of Math items and standards than in English.

A second conclusion drawn from our findings concerns the cognitive demand and rigor scales. These scales were used as the basis for rating items, and clearly impacted the reliability of ratings. We found greater reliability in cognitive demand scales in both content areas, and across items and standards. More research is needed to determine a more precise conclusion as to why cognitive demand ratings tend to be more reliable, but we suspect the difference may be related to the descriptors used to define the rating scales. Recall the cognitive demand rating scale was based on Marzano's (2001) taxonomy, where the descriptors retrieval, comprehension, analysis, and knowledge of utilization comprise the rating scale. Potentially, raters searched for these descriptors within the text of items and standards in order to make their judgments. The rigor rating scale, on the other hand, contained more obscure descriptors that were open to interpretation and subjectivity, where the scale contained three descriptors: below, at, and above the level at which an entry-level college student should perform. Overall, these results indicate the cognitive demand scale may contain better descriptors that elicit more immediate responses with ease and objectivity. The rigor scale, on the other hand, may require more precise

descriptors for raters to easily elicit responses, or perhaps was subject to more sources of error and disagreement in determining adequate skill levels for entry-level college students.

Third, sources of error variability from independently conducted ratings (e.g., variation in rating time/location) not captured in alignment study designs can have a sizable impact on the reliability of item and/or standard ratings. In terms of error variance, there were greater differences between individual (absolute) and residual effects in cognitive demand than rigor ratings, which suggests larger interaction effects between raters and the object of measurement (items or standards) for the cognitive demand scale. It is not known if these differences could be attributed to the interaction effect between raters and the objects of measurement, or facets not included in the design.

Finally, the D-study results suggest that the standard five or six raters typically assumed to be sufficient for obtaining reliable ratings in alignment studies might be insufficient. Particularly, five or six raters may be insufficient when using rating scales similar to the rigor scale. These findings show the number of sufficient raters may differ according to the type of scale used in the alignment study, which is particularly useful knowledge in the design of future alignment studies. Also important to consider, the number of raters may differ according to content area, as our findings show differing levels of reliability for English and Math across both items and standards. For both standards and items, six appears to be a sufficient number of raters for Math and Cognitive Demand, but insufficient elsewhere.

This study was conducted in the context of a larger study addressing alignment of college admissions and placement test items and a set of validated college readiness standards, a policy arena where potential high-stakes decisions could be made based on the alignment study results. As mentioned, the alignment of college admissions and placement tests to college readiness

standards is crucial, as previous studies showed these assessments may not align strongly enough with college readiness standards and are therefore insufficient in providing feedback to high school students and teachers concerning college readiness or remediation needs. Ultimately, these findings reiterate the importance of critically attending to the design of alignment studies, particularly elements regarding the content expertise of raters, the use of multiple rating scales to determine alignment, and rater training and process. These findings clearly demonstrate the value of conducting a generalizability theory analysis to evaluate rater reliability in alignment studies.

References

Achieve, Inc. (2007). *Aligned expectations? A closer look at college admissions and placement tests*. Washington, DC: Author.

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

Brown, R.S. and Conley, D.T. (2007). Comparing state high school assessments to standards for success in entry-level university courses. *Educational Assessment, 12*(2), 137-160.

Brown, R.S., & Niemi, D.N. (2007). *Investigating the alignment of high school and community college assessments in California*. National Center Report #07-3. The National Center for Public Policy and Higher Education.

Conley, D.T. (2003). *Mixed messages: What state high school tests communicate about student readiness for college*. Eugene: University of Oregon, Center for Educational Policy Research.

Conley, D.T. (2007). *Redefining college readiness*. Educational Policy Improvement Center, Eugene, Oregon.

Conley, D.T. (2010). *College and career ready: Helping all students succeed beyond high school.* San Francisco, CA: Jossey-Bass.

D'Agostino, J.V., Welsh, M., Cimetta, A., Falco, L., Smith, S., VanWinkle, W., & Powers, S. (2008). The rating and matching item-objective alignment methods. *Applied Measurement in Education, 21*, 1-21.

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments: A Case study. *Applied Measurement in Education, 20*,101-126.

Herman, J. L., & Webb, N. M. (Eds.) (2007). *Special Issue of Applied Measurement in Education: Alignment Issues, 20*, 1-135.

Marzano, R. J. (2001). *Designing a new taxonomy of educational objectives.* Thousand Oaks, CA: Corwin Press.

Musquash, C., & O'Connor, B. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38* (3), 542-547.

National Center for Education Statistics [NCES] (2004). The condition of education 2004, indicator 18: Remediation and degree completion. Washington, DC: U.S. Department of Education.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3-14.

Rothman, R. (2003). Imperfect matches: The alignment of standards and tests. National Research Council.

Shavelson, R.J., & Webb, N.M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage Publications.

SPSS, Inc. (2007). *SPSS 16.0 for Windows.* Lead Technologies.

Thompson, B. (2003). A brief introduction to generalizability theory. In B. Thompson, (Ed.) *Score reliability* (pp. 43-58). Thousand Oaks, CA: SAGE Publications.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (National Institute for Science Education NISE Res. Monograph No. 6). Madison: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (National Institute for Science Education NISE Res. Monograph No. 18). Madison: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (2002, April). *An analysis of the alignment between mathematics standards and assessments for three states*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Wixson, K.K., Fisk, M.C., Dutro, E., & McDaniel, J. (2002). *The alignment of state standards and assessments in elementary reading*. CIERA Report #3-024. University of Michigan School of Education, Center for the Improvement of Early Reading Achievement: Ann Arbor, MI.