

An Alternative Strategy for Estimating Decision Consistency Reliability

Chad W. Buckendahl

Yongwei Yang

Abdullah Ferdous

University of Nebraska – Lincoln

April, 2003

## Abstract

Assessment and accountability systems that incorporate locally developed information may face challenges demonstrating reliability using traditional measures. For many educators, these skills are beyond the assessment literacy training they received. However, without evidence of reliability, validity of the inferences made about students' performance is questionable. What practical methods are available that would provide classroom educators reliability evidence to defend the use of their students' scores on these assessments? This paper examines the feasibility of a strategy for estimating reliability by comparing results from a traditional internal consistency measure and a proposed decision consistency approach. Results are presented from analyses conducted using multiple mathematics assessments as an illustration.

### An Alternative Strategy for Estimating Decision Consistency Reliability

When local assessment information is used to inform state accountability needs, districts may be required demonstrate the technical quality of their assessments. Unfortunately, few states have assessment literacy requirements for teachers or administrators (Stiggins, 1999). For educators without training in measurement or assessment literacy skills, these technical quality requirements may be daunting. Although their local assessments may be aligned to the district's curriculum and instruction, the lack of technical skills may force them to select commercial or other external assessments for their students that may not be as instructionally relevant. To add credibility to the information districts provide, new technical quality strategies need to be explored that allow districts to demonstrate the utility of their locally developed assessments for purposes beyond the classroom. To the extent possible, an additional goal is that any new strategies be psychometrically sound, yet educator-friendly.

A technical area that may be especially challenging for educators is the concept and common methodologies for estimating score or decision consistency reliability in the context of assessment. Traditional strategies for estimating reliability may not be efficient as districts may lack the time, money, or resources to meet the assumptions of the methods. For example, very few school districts would develop multiple forms of a test that would not be used for high stakes purposes. Thus, an alternate forms reliability calculation would not be feasible. Another methodology that districts might consider is a test-retest estimate of stability. However, in an educational setting, it may not be possible to re-administer an assessment after a reasonable interval without additional instruction or learning in the measured area of the content domain. For single administration

assessments, estimates of internal consistency like KR20, KR21 or coefficient alpha tend to be popular choices in the psychometric community (Traub, 1994). The utility of these formulas to educators at the district level to the instructional process may not be readily evident.

### Rationale for the Study

Brennan (2001) discusses the concept of reliability in terms of replication of scores. Although this is generally accepted within the measurement community, replication alone is rarely the end goal of these analyses. Researchers generally want to go beyond simply observing replicated scores to making inferences about the scores. However, because the assumptions for traditional reliability measures may not be met for criterion-referenced assessments, some researchers have sought to extend this replication concept beyond scores.

Popham and Husek (1969) were early critics of using traditional reliability indices for criterion-referenced tests because of the perceived lack of variability among examinees' scores. Although in reality there is likely sufficient examinee variance to estimate reliability using these measures (Popham, 1990), the concept appears contradictory to many educators who have a mastery-learning orientation. If assessments are constructed to measure a narrow section of the ability continuum, there may not be sufficient variability in students' scores to produce an acceptable internal consistency value. If there is a wide ability range in the tested student population, this low value may accurately reflect the confidence in the stability of the scores. However, if there is a narrow band of abilities among the tested population, an internal consistency measure

may be inappropriate. In these situations, other methods are needed to demonstrate the reliability of inferences about student performance.

An alternative strategy is to consider reliability as decision consistency where the inferences made about the decisions about performance on assessments are evaluated. Given its close ties to the idea of criterion validity, this concept is described in the context of validation theory as being a “decision-based interpretation” Kane (2002). This is in slight contrast to traditional concepts of reliability and the replication of scores, however, decision consistency can be thought of as the replication of decisions. Because decisions about the scores are many times a goal, analyses of this concept may be more meaningful to educators.

Various strategies have been proposed for estimating decision consistency for criterion referenced tests. Some of these methodologies are based on the assumption that there are multiple forms of the assessment. Cohen’s (1960) coefficient kappa offers a method that calculates decision consistency as agreement between decisions from two measures and includes a correction for chance agreement. Another strategy proposed by Swaminathan, Hambleton, and Algina (1974) estimates the agreement between pass-fail decisions from alternate forms of assessments designed to measure the same construct.

Additional strategies have been proposed for estimating the decision consistency reliability of criterion-referenced tests that only require a single form of the assessment. Livingston (1972) proposed a formula that estimated reliability as a function of the expected squared deviation from the cut score. However, this value may not be especially useful beyond illustrating how much better reliability is when compared to a cut point that is not close to the mean (Livingston, 2002). Subkoviak (1976) offered a decision

consistency method that is appropriate for tests with dichotomously scored items and estimates the probability of correct classification of pass/fail decisions given a cut scores.

Breyer and Lewis (1994) suggested a simplified strategy for estimating decision consistency by calculating the probability of correctly classifying examinees as masters or non-masters as if the examinees had taken two alternate forms of the test. Livingston and Lewis (1995) also proposed a single administration decision consistency method that estimates the consistency of the decisions on alternate forms of a test and the accuracy of those decisions relative to the examinees' true score which is calculated using an artificial distribution estimated from the observed scores on the single administered form of the test. Huynh (1976) and Kane and Brennan (1980) have also offered alternative approaches for estimating decision consistency for criterion referenced tests. Many of these methods, though, may be too challenging for individuals not trained in measurement and who do not see the utility in the exercise. What practical methods are available that would provide classroom educators reliability evidence to defend the use of their students' scores for making inferences about performance?

One consideration may be an extension of a simplistic form of decision consistency that is typically calculated using the percent of agreement between two assessments designed to measure the same construct (Traub, 1994). Although it may be assumed that these assessments are written items or tasks, it may be possible to expand the definition of an assessment. If one accepts that a teacher's professional judgment about the abilities of his or her students within a content domain is an assessment, the concept of decision consistency and its utility may be extended. This paper compares the results of traditional internal consistency and a proposed simplified decision consistency

reliability estimates that are calculated using teacher's professional judgments about their students' proficiency and actual students' performance on mathematics assessments in a consortium of Midwestern school districts.

### Methods

Data for this study were collected from 57 mathematics assessments in a consortium of approximately 20 Midwestern school districts. Written assessments were designed to measure the state's mathematics content standards and represent elementary (4<sup>th</sup>), middle (8<sup>th</sup>), and high school (11<sup>th</sup>) levels. For the study, the consortium selected a representative sample of 75 or more students at each grade level from a representative number of school districts (8-10). Reliability analyses were conducted using a commonly used traditional internal consistency estimate, coefficient alpha (Cronbach, 1951), and a proposed decision consistency strategy (percent agreement) that uses teacher judgments of student proficiency level and a written assessment that empirically classifies performance as the two assessments.

Analyses of decision consistency were calculated at two levels, proficient and above and progressing and below classifications and at four levels, beginning, progressing, proficient, and advanced proficiency classifications. This calculation was conducted at two levels because proficient and above represents students who have met the standard the assessment was designed to measure and progressing and below represents students who have not yet met the standard. The additional level of analysis was included to examine whether decision consistency was equivalent across the desired proficiency classification categories. Results presented in the next section show the comparison between the two methods as different alternatives for demonstrating

reliability information. A Pearson correlation was also conducted between the calculated coefficient alpha estimate and each decision consistency estimate to provide a broad estimation of the relationships among these methodologies.

### Results

The results of analyses that compared the results of a traditional internal consistency measure and two levels of a proposed decision consistency estimate are presented by grade level. The school districts in this consortium designed an assessment to measure each of the mathematics content standards in numeration and number sense, computation and estimation, measurement, geometry and spatial concepts, data analysis, statistics, and probability, and algebraic concepts. One of the 8<sup>th</sup> grade assessments was designed to measure two of the data analysis, statistics, and probability standards.

Analyses were conducted by assessment and grade level.

[Insert Table 1 Here]

Table 1 above shows the results of the 4<sup>th</sup> grade reliability analyses. Coefficient alpha values range from .30 to .91. A reasonable target value for making group decisions is an internal consistency value of .70. Half of the assessments at this grade level meet this criterion decision rule. The two level decision consistency analysis asked teachers classify their students into proficiency categories of proficient and above or progressing and below. These judgments were matched against the empirical classification of student proficiency based on their test performance relative to a predetermined cut score. The calculated levels of agreement ranged from .59 to .81, meaning that on average, teachers were able to classify their students in terms of meeting versus not meeting the content standard about 70% of the time.

A four level analysis was also conducted to determine the level of agreement across the four desired levels of inference for the school districts (beginning, progressing, proficient, and advanced). The values for these agreement analyses ranged from .22 to .42. Note that for the four level analyses it was necessary to set three cut scores prior to calculating the decision consistency values. For some assessments, there was insufficient information to allow for three cut scores to be set. These assessments are identified in the 4x4 decision consistency column as “Not Available” (N/A) and also apply to 8<sup>th</sup> and 11<sup>th</sup> grade data.

The correlation between the results of coefficient alpha and the two level decision consistency was .25, whereas the correlation between the results of coefficient alpha and the four level decision consistency was .35. The correlation between the two level and four level decision consistency was .88.

[Insert Table 2 Here]

Table 2 shows the results of the 8<sup>th</sup> grade reliability analyses. Coefficient alpha values range from .10 to .96. All but three assessments at this grade level meet this criterion decision rule of .70 for minimum acceptability. The two level decision consistency analyses resulted in calculated levels of agreement ranging from .49 to .81, meaning that on average, teachers were able to classify their students in terms of meeting versus not meeting the content standard about 70% of the time. The four level analyses resulted in values ranging from .16 to .55.

The correlation between the results of coefficient alpha and the two level decision consistency was -0.15, whereas the correlation between the results of coefficient alpha

and the four level decision consistency was .45. The correlation between the two level and four level decision consistency was .65.

[Insert Table 3 Here]

Table 3 shows the results of the 11<sup>th</sup> grade reliability analyses. Coefficient alpha values range from .38 to .93. Most assessments at this grade level meet this criterion decision rule of .70 for minimum acceptability. The two level decision consistency analyses resulted in calculated levels of agreement ranging from .65 to .81, meaning that on average, teachers were able to classify their students in terms of meeting versus not meeting the content standard about 70% of the time. The four level analyses resulted in values ranging from .27 to .46.

The correlation between the results of coefficient alpha and the two level decision consistency was .19, whereas the correlation between the results of coefficient alpha and the four level decision consistency was .27. The correlation between the two level and four level decision consistency was .73.

### Discussion

This study evaluated the level of agreement between a commonly used internal consistency reliability method and a proposed decision consistency strategy. A goal of the study was to determine if the relationship between methods was sufficient to consider using the results of the proposed decision consistency approach as a reasonable proxy for more traditional methods. Given its conceptual ease and simplicity of calculation, the method may offer educators a practical strategy for defending the use of their students' scores for decisions about proficiency levels within the context of mathematics assessments.

Although many of the internal consistency estimates were acceptable in terms of generally accepted psychometric standards, very few of the decision consistency estimates at either two levels or four levels were at minimally acceptable levels. Correlations between coefficient alpha values and decision consistency values were low. It is noted that because internal consistency focuses specifically on the relationships among items that comprise the scores and decision consistency focuses specifically on the decisions that result from the scores, that there is not an expectation of a high correlation between the methods. However, these results suggest that for this application, the proposed decision consistency strategy did not produce values that would support a district's assertion that decisions about student performance on these assessments were reliable. For this consortium, reporting their internal consistency values would provide better evidence of reliability of the scores, rather than relying on a decision consistency approach to estimate the level of agreement between two assessments designed to measure the same area of the content domain.

It was somewhat surprising to see that the level of agreement between dichotomous classification of student performance (proficient and above or progressing and below) was so low. Because these teachers had considerable experience with both the content and the students they were predicting performance, we expected to see higher levels of agreement (85%+) for the two level decision. Although lower levels of agreement were expected for the four level decision, many of the calculated values at this level ranged from 25%-33% which does not provide much support for the reliability or validity of decisions. At least factors may have contributed to these low values for both the two level and four level decision consistency analyses.

First, the specificity of the proficiency level categories may have been too vague. Similar to rubric development, review, and training activities that are part of inter-rater agreement studies, the specification of the proficiency levels with regard to the content is critical to the proposed method. If the proficiency level categories only reflect value terms such as the “Advanced student will be able to accomplish all mathematics tasks without assistance,” or the “Proficient student will be able to accomplish most mathematics tasks without assistance,” teachers are still left to define “all,” “most,” and “mathematics tasks” based on their own varied experiences with curriculum and students. This does not provide much guidance to the meaning of the proficiency category.

Better descriptions would specify the knowledge, skills, and abilities of the advanced or proficient student relative to the subarea of the content domain of interest and what distinguishes students in these proficiency categories from one another. For example, in the 8<sup>th</sup> grade, an advanced student for a given data analysis, statistics, and probability standard may be able to conceptually understand and independently calculate the mean, median, and standard deviation. However, the proficient student may only conceptually understand and calculate the mean and median. They may not conceptually understand standard deviation and may not be able to calculate it without normal assistance from the teacher.

Second, there may not have been sufficient items across the full ability continuum that would allow for potential classification of students into multiple proficiency categories. Because it was important for these districts to be able to classify students into four proficiency categories (beginning, progressing, proficient, and advanced), sufficient items are needed on the assessments that reflect these four broad ability levels. This

means that there needs to be items/tasks on the assessments that measure beginning skills, progressing skills, proficient skills, and advanced skills if the desire is to report student performance in these three proficiency categories. In examining the raw data, this appears to be one of the pervasive problems across the assessments reported in the illustration above, specifically as it applies to the four level decision consistency analyses.

The majority of misclassifications of students into proficiency categories resulted from students being classified into a higher proficiency category than the teacher predicted, specifically, students being identified as “Advanced” by their performance on the test as opposed to the “Proficient” level the teacher predicted. This suggests that either the teachers consistently underestimated the performance of their students or that the test was not sufficiently representative of the full ability continuum such that the teachers’ judgments were empirically supported by the students’ performance on the test.

It is likely that the tests were not designed to provide sufficient information about the full range of students’ abilities. This is not surprising because many classroom teachers are accustomed to district adopted grading scales that many times use 60% or 70% as an arbitrary cut score for passing. There is a limited range in remaining scale as it runs into a ceiling at 100%. Thus, many educators believe if a student achieves a high score on an assessment (95%-100%) then by definition, the student is advanced, even if the student has not demonstrated any advanced level skills on the assessment. The same rationale applies to each proficiency level classification.

Third, it is also possible that judgments were made about students’ proficiency level that included other factors beyond the criterion-referenced definition of performance. Because there were four levels of proficiency, it may be only natural for an

educator to think about these levels as equivalent to their classroom grades of A, B, C, and D. Including this additional information in the judgments may introduce more variability into the method that will lead to misclassifications. Within classroom grades, there are factors such as students' effort, homework performance, class participation, and classroom behavior that may influence grades, but may or may not represent criterion-referenced performance in an area of the content domain. A student that does not turn in their homework or is disruptive in class may have advanced skills in an area that is not reflected in their course grade. The alternative situation is also possible where a student puts forth a tremendous effort and actively participates in class discussions, but has not yet achieved proficient level skills in a given area.

Further research in the area of reliability methods for criterion-referenced assessments would benefit from replication of similar studies that include better descriptions of proficiency categories and have evidence to support that assessments were designed to measure the scope of the desired inferences. It is possible that the decision consistency estimates reported in this study were underestimated because of the limitations of both the assessments and the educators' judgments. Another useful study would be to examine the proposed decision consistency method in this study in comparison to some of the common decision consistency methods that are also employed by the measurement community.

### Conclusions and Implications

This study can be viewed as one attempt to develop strategies that may help to bridge measurement theory and practice. The study is important for educators as it examines the relationship of the results of a commonly accepted internal consistency

method and a proposed decision consistency strategy for estimating reliability. The proposed method may be useful to school districts that are required to demonstrate reliability evidence for their local assessment systems, but may not have the sophistication to use traditional methods. With greater attention placed on testing at various levels, it will continue to be important to provide evidence of the reliability and validity of inferences made about student, school, district, and possibly state performance.

This study examined the feasibility of using teacher judgments as part of a proposed variation on decision consistency reliability. By exploring strategies that incorporate the valuable information that educators possess, they are participants in the assessment process rather than outsiders. The model of decision consistency proposed in this paper, then, may be seen as an effort to develop a strategy that provides empirical support for educators' professional judgments and extends theory on estimating reliability for criterion-referenced assessments.

## References

- Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. Journal of Educational Measurement, *38*(4), 295-317.
- Breyer, F. J. & Lewis, C. (1994). Pass-fail reliability for tests with cut scores: A simplified method. ETS Research Report RR-94-39. Princeton, NJ: Educational Testing Service.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, *16*, 297-334.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, *20*, 37-46.
- Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, *38*, 725-736.
- Kane, M. T. & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. Applied Psychological Measurement, *4*, 105-126.
- Kane, M. T. (2002). Validating high stakes testing programs. Educational Measurement: Issues and Practices, *21*(1), 31-41.
- Livingston, S. A. (1972). Criterion-referenced applications of classical test theory. Journal of Educational Measurement, *9*, 13-26.
- Livingston, S. A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. Journal of Educational Measurement, *32*(2), 179-197.
- Livingston, S. A. (2002). Personal communication on January 18, 2002.

Popham, W. J. & Husek, T. R. (1969). Implications of criterion-referenced measurement. Journal of Educational Measurement, 6, 1-9.

Popham, W. J. (1990). Modern Educational Measurement: A Practitioner's Perspective (2<sup>nd</sup> ed.) pp. 121-145. Englewood Cliffs, NJ: Prentice Hall.

Stiggins, R. J. (2001). The unfulfilled promise of classroom assessment. Educational Measurement: Issues and Practice, 20(3), 5-15.

Subkoviak, M. J. (1976). Estimating the reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 13, 265-276.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability for criterion-referenced tests: A decision-theoretic framework. Journal of Educational Measurement, 11, 263-267.

Traub, R. E. (1994). Reliability for the Social Sciences: Theory and Applications. Thousand Oaks, CA: Sage Publications, Inc.

Table 1.

Comparison of coefficient alpha and two levels of decision consistency for 4<sup>th</sup> grade assessments.

<u>Assessment</u>	<u>Coefficient Alpha</u>	<u>2x2 DC</u>	<u>4x4 DC</u>	<u>Number of Items</u>
4.1.1	.74	.67	.33	31
4.1.2	.91	.66	.26	45
4.1.3	.58	.73	.34	33
4.1.4	.73	.63	N/A	11
4.1.5	.30	.79	N/A	6
4.2.1	.83	.76	.39	37
4.2.2	.72	.77	.42	20
4.2.3	.62	.59	N/A	12
4.3.1	.69	.65	.26	24
4.3.2	.48	.69	.27	23
4.3.3	.74	.81	.38	19
4.3.4	.64	.67	.28	4
4.4.1	.52	.76	.32	16
4.4.2	.76	.77	N/A	10
4.4.3	.83	.76	N/A	14
4.5.1	.49	.64	.25	16
4.6.1	.71	.68	.26	20
4.6.2	.55	.59	.22	7

Table 1.

Comparison of coefficient alpha and two levels of decision consistency for 8<sup>th</sup> grade assessments.

<u>Assessment</u>	<u>Coefficient Alpha</u>	<u>2x2 DC</u>	<u>4x4 DC</u>	<u>Number of Items</u>
8.1.1	.88	.71	.34	10
8.1.2	.84	.77	.34	16
8.1.3	.84	.73	.37	15
8.1.4	.89	.71	.38	40
8.2.1	.96	.80	.55	262
8.2.2	.73	.80	.48	16
8.2.3	.91	.77	.43	50
8.2.4	.82	.77	.41	15
8.2.5	.78	.77	.40	35
8.3.1	.85	.71	N/A	40
8.3.2	.85	.75	.37	20
8.4.1	.85	.49	.16	45
8.4.2	.87	.72	.33	46
8.4.3	.87	.64	.27	16
8.4.4	.90	.71	N/A	12
8.4.5	.43	.78	.27	12
8.5.1/8.5.2	.74	.81	.42	11
8.5.3	.87	.76	.49	15
8.5.4	.27	.84	.29	11
8.6.1	.86	.81	.49	35
8.6.2	.93	.75	.51	40
8.6.3	.10	.69	.27	6

Table 1.

Comparison of coefficient alpha and two levels of decision consistency for 4<sup>th</sup> grade assessments.

<u>Assessment</u>	<u>Coefficient Alpha</u>	<u>2x2 DC</u>	<u>4x4 DC</u>	<u>Number of Items</u>
12.1.1	.83	.75	.41	13
12.1.2	.84	.73	.41	17
12.2.1	.67	.75	.44	15
12.2.2	.48	.81	N/A	8
12.2.3	.77	.73	.35	20
12.3.1	.38	.69	N/A	14
12.3.2	.68	.77	.44	7
12.4.1	.93	.79	.40	18
12.4.2	.70	.69	.36	5
12.4.4	.85	.74	.31	7
12.4.5	.90	.77	.43	11
12.4.6	.68	.79	N/A	5
12.4.7	.64	.65	.27	12
12.6.1	.71	.76	.46	6
12.6.2	.88	.76	.45	32
12.6.3	.92	.75	.45	16
12.6.4	.78	.67	.36	11