

**Item Sufficiency in Educational Assessments When Multiple Cut-Points are Used**

Rebecca L. Norman  
*University of Nebraska-Lincoln*  
rebeccalnorman@yahoo.com

Paper presented at the annual meeting of the  
National Council on Measurement in Education  
San Francisco, California  
April, 2006

## **Abstract**

The purpose of this study was to evaluate the sufficiency of measurement for reporting students' performance at multiple levels using assessments in mathematics that are currently being used in a Midwestern state using Subkoviak's (1988) suggestion of at least 6 items for dichotomous decisions as a guide. The analysis is based on performance level judgments of items and tasks from multiple assessments (CRTs and NRTs) at the Grade 4, Grade 8, and High School level, made by fifty-two teachers. This state applies four performance categories; Beginning, Progressing, Proficient, and Advanced. Results revealed that in the majority of instances, there were enough items to classify students into two or three performance levels, but rarely were there enough to make all four classifications. Implications are provided for those in the field of education.

## **Introduction**

Recent research has focused on validity issues and the growing question of the appropriate number of items necessary to represent a construct in educational assessments. Ryan and DeMark (2002) discuss validity in assessments as the ability to be fair and give an accurate reflection of what students know and can do. Linn (2002) expressed validity in a similar matter, noting that the proper use and interpretation of an assessment score is necessary to support validity. Under the broader heading of validity, Linn (2002) cites "construct under representation" as a problem of not having enough information to properly assess a particular construct.

Information to properly assess a construct can be attained by having a sufficient number of items making up the assessment. Research has sought to answer the question of how many items are required to measure a particular construct. The validity framework for a testing program begins with defining the intended uses therefore the number of items or tasks depends on the type of decision to be made, as well as the characteristics of the domain. In high stakes decisions it is important to have many items or tasks, so people are not misclassified. When the stakes are lower, for example in a classroom quiz, fewer items may be acceptable (Popham, 1990). Millman and Greene (1989) recommend 7 or 8 items as a necessary minimum when making a dichotomous educational decision (e.g. mastery or nonmastery). Subkoviak (1988) suggests that 6 to 8 items with specific characteristics may be needed to make a confident decision. No widely accepted criterion exists for making classifications beyond dichotomous decisions. Because in many instances we have moved beyond dichotomous decisions in educational assessment, additional research is needed to address the sufficiency of items when multiple performance level cut points are used in an assessment.

Another way to characterize the concept of assessment validity is that the assessment needs to have the capacity to provide opportunities for students to demonstrate their abilities that are consistent with the reported performance levels. Brookhart (2005) examined Nebraska mathematics assessments which require students to be classified into one of four performance categories, and determined that the state can only be confident regarding classifications of students below or above the progressing/proficient cut-point. However, at the beginning and advanced levels sufficiency lacked. Bhola et al. (2003) also discuss the issue of multiple performance categories and express the difficulty in making accurate classifications in these situations. Having more than two performance categories requires not only more items, but also

items distributed across the whole range of performance categories in order to make valid decisions. Simply multiplying the recommended number of items by the number of cutpoints or performance categories is not sufficient; the items must be adequately dispersed. In order to comply with NCLB requirements, assessments must have the ability to classify students into each of the performance categories; therefore it is important that each performance category is adequately represented (Bhola et al., 2003).

Two ways in which students are commonly classified into performance categories are through criterion-referenced tests (CRTs) and norm-referenced tests (NRTs). Thorndike (2005) explains that CRTs are concerned with measuring specific skills in classroom situations. These are tests most frequently given by teachers in a classroom in order to determine whether or not a student has mastered a particular skill. The concern of a NRT is to be capable of comparing a student, or group of students, to a general reference group and determine how well he/she, or the group, is showing achievement in comparison to this reference group (Thorndike, 2005).

Many states classify students' performance into multiple categories (e.g., Below Basic, Basic, Proficient, or Advanced) through CRTs and/or NRTs. These assessments may or may not contain items or task that offer breadth and depth of measurement in each of these categories. The purpose of this study was to evaluate the sufficiency of measurement for reporting students' performance at multiple levels using commercially available and teacher produced assessments in mathematics that are currently being used in a Midwestern state using Subkoviak's recommendation of at least 6 items at each performance level.

### **Methods**

The data analyzed in the current study was collected in two phases in a Midwestern state. Phase 1 involved the selection of mathematics standards to be analyzed (which are summarized in Table 1). This Phase also consisted of the researchers, accompanied by personnel from the state's department of education, developing performance level descriptors for these selected mathematics standards in grades 4, 8, and high school. The performance levels for this state are Beginning, Progressing, Proficient, and Advanced. Draft performance level definitions (PLDs) were developed during a one-day meeting with teacher teams at each grade level. These are the performance levels that were used for the operational ratings which took place in Phase 2 of the study. Phase 1 also involved the selection of commercially available (NRTs) and teacher produced assessments (CRTs) for analysis.

#### *Test selection*

Criteria for Commercially available NRTs were 1) that the test was not going to be revised within the next year or two, thus the results would remain current for the near term and 2) that the tests were previously judged to assess the standards that were selected for inclusion in the study. CRT's were chosen based on multiple criteria. First, they were sought from a representative number of districts from each region of the state, West, Central, and East. It was desired to attain 5 to 6 per region. The second criterion required that the assessments are "standardized" within the district. That is, all students in the district are assessed using the same assessment tasks for the standards being selected. Third, districts have assessment tasks for each of the selected standards. Fourth, that each selected district's assessment has been rated at least Very Good in terms

of overall assessment quality. Fifth, that a range of district sizes be included. A district could have provided assessments at all three grade levels.

These criteria were not always met. Districts were hesitant to provide their assessments for review and obtaining cooperation proved to be very difficult. In total only 11 districts provided their Mathematics assessments. Some districts did not provide assessments for all grade levels either because they did not have all grade levels, or did not want to share certain assessments. There were also assessments that were written in ways which rating them would not be possible. One districts' assessment was used for practice, so the operational ratings were done on 10 districts' assessment materials.

#### *Meeting locations*

To encourage participation, three locations for the operational classification of assessment tasks into performance categories, were selected. One meeting was held in each region of the state (West, Central, and East). It was the intention that in addition to one NRT being rated, assessments from that region would also be rated. Because of the lower than expected level of cooperation by school districts in sharing their assessments, one district's assessment was rated in two locations (East and West).

#### *Selecting Teachers*

Because the study had two phases and because the meaning of the performance level definitions was a critical element of the rating of assessment tasks, teachers who had participated in phase one, development of performance level definitions provided a cadre of participants for phase two. Thus, two groups of teachers were recruited to participate in this study. Group 1 participated in both the determination of performance level definitions and the rating of assessment tasks, and Group 2 participated only in rating assessment tasks. Thus it was anticipated that up to six teachers (two at each grade level) from group 1 would participate in each of the three regional meetings to rate assessment tasks. The rationale for the Group 1 teachers was that they could provide insights into the development and meaning of the performance level definitions during the rating process. This, however, was not the case. The time frame for the study was such that most of the teachers who participated in the development of the definitions in late July 2003 did not recall the deliberations or outcomes in sufficient detail by the time they met again to make the ratings to be helpful in the rating process.

As was the case in obtaining assessment materials for review, recruiting teachers was also a challenge. Not all of the Group 1 teachers elected to participate in the subsequent regional meetings to rate assessments. In addition, obtaining additional teachers often proved difficult. The number of teachers at each of the meeting sites is shown in table 1.

Table 1. Number of teachers who participated in rating assessment tasks at each meeting.

Region	Grade	Teachers
East	4	9
	8	6
	H.S.	6
Central	4	6
	8	5
	H.S.	6
West	4	4
	8	7
	H.S.	5

### *Conducting the meetings*

Six meetings were held in three locations, one in each region of the Midwestern state. Each meeting started at 1:00 in the afternoon and ended the following afternoon. Prior to each of the meetings the assessments that were to be evaluated were examined and for each assessment task that could be rated a rating form was created. Teachers were given copies of a power point presentation that was used for an orientation, copies of the performance level descriptions, a non-disclosure form, an informed consent form, a demographic information form, and an evaluation form for each meeting. Teachers were also provided a travel reimbursement form.

A total of 52 teachers met at one of three locations across the state. Each meeting began with an orientation that described the purpose of the meetings and the context in which the study was being conducted. They were assured that districts that provided assessments would be anonymous in reports to the state's department of education. The first task for the teachers was to read through the PLDs and discuss them. The objective was for the teachers to understand what these definitions were and how the four performance levels were differentiated in terms of knowledge, skills, and abilities in mathematics. Next teachers evaluated and rated a practice assessment. Those teachers who had been part of Group 1 had already rated these, however their original ratings were not shared with the Group 2 teachers until their ratings were complete.

Teachers were split into grade level groups and began by making independent judgments on rating forms about the performance level that an item or task measured. Items that were rated included; multiple-choice items, short answer items, and performance type items that are scored using a rubric. In order to make their decision, teachers were advised to review the performance level description associated with the standard that the assessment task was intended to assess. Then decide if the majority (about 66%) of Beginning students would answer it correctly. If not, then ask if the majority of the Progressing students would be able to answer correctly. If the assessment task was too difficult for these lower proficiency levels, continue the process asking about Proficient, and if necessary, Advanced. For dichotomous items, it was only necessary to mark the lowest performance level of students who would likely answer correctly. Items polytomously scored with a rubric (performance tasks), teachers

evaluated how many points a student at the Beginning level would likely earn, and then did the same for the three remaining categories based on the number of points possible.

Following the independent judgments, the teachers were asked to come to consensus on each of the items and tasks for each assessment. In the consensus process, each group member was asked to indicate the performance level they selected for each item or task. Generally, if a majority of the group members agreed, no further discussion occurred. If there was disagreement discussion occurred until consensus was reached. At the conclusion of the meetings, teachers were given the opportunity to evaluate the process using a form that was created for that purpose. The entire meeting process was overlooked by a facilitator who organized test materials, looked over the rating forms for completed tasks, and was available to answer questions. The facilitator assigned a different teacher to serve as a Table Leader for each assessment to lead discussion. This eliminated the possibility of the facilitator driving decisions about classifications.

The analysis of the teacher item ratings involved examining the data to determine item sufficiency for classifying students into the four categories (beginning, progressing, proficient, and advanced). This was done by examining each possible dichotomous decision; Beginning or above, Progressing or below (or above), Proficient or below (or above), and Advanced or below. Sobkoviak's (1988) suggestion of at least 6 items (or measurement opportunities was used to make evaluations, and at least 4 of the 6 points must be obtained to be classified at a particular level. Each NRT and CRT was analyzed independently. Some CRTs were analyzed at more than one meeting; specifically one CRT was rated both at the Eastern and Western meeting, and two district assessments were analyzed at all meetings.

Table2. Description of Education Standards.

Grade	Standard	Description
Fourth	4.1.2	By the end of fourth grade, students will make change and count out in amounts up to \$20.00.
	4.2.1*	By the end of fourth grade, students will estimate and accurately calculate without and with calculators and solve problems involving addition, subtraction, multiplication, and division of whole numbers and understand the relationships among the operations.
	4.3.4	By the end of fourth grade, students will determine the perimeter of a many-sided figure (without a formula) using both standard and nonstandard units of measure.
	4.4.2	By the end of fourth grade, students will identify and draw points, lines, line segments, rays, and angles.
	4.5.1*	By the end of fourth grade, students will collect, organize, represent, and interpret numerical and categorical data and clearly communicate the findings.
	4.6.2	By the end of fourth grade, students will identify, describe, and extend arithmetic patterns, using concrete materials and tables.
Eighth	8.1.4	By the end of eighth grade, students will apply appropriate number of theory such as prime and composite, factors and multiples, divisibility, powers, properties, and identities.
	8.2.2*	By the end of eighth grade, students will identify the appropriate operation and do the correct calculations to solve word problems.
	8.3.2	By the end of eighth grade, students will convert units within measurement systems using proper conversion factors (standard and metric).
	8.4.1	By the end of eighth grade, students will identify, describe, compare, and classify geometric figures such as plane figures like polygons and circles; solid figures like prisms, pyramids, cones, spheres, and cylinders; and lines, line segments, rays, angles, parallel and perpendicular lines.
	8.5.2*	By the end of eighth grade, students will read and interpret tables, charts, and graphs to make comparisons, predictions, and inferences.
	8.6.3	By the end of eighth grade, students will describe and represent relations, using tables, graphs, and rules.
Twelfth	12.1.2	By the end of twelfth grade, students will describe and compare the relationships among all subsets of real numbers.
	12.2.1*	By the end of twelfth grade, students will solve theoretical and applied problems using numbers in equivalent forms, radicals, exponents, scientific notation, absolute values, fractions, decimals, and percents, ratios and proportions, order of operations, and properties of real numbers.
	12.3.1	By the end of twelfth grade, students will select and use appropriate measuring units, tools, and/or technology to achieve a specified degree of accuracy and precision.
	12.4.5	By the end of twelfth grade, students will apply right triangle trigonometry to solve problems.
	12.5.1*	By the end of twelfth grade, students will apply sampling techniques to gather data, organize, display, and interpret data to solve complex problems.
	12.6.3	By the end of twelfth grade, students will apply and solve problems involving systems of equations, and systems of inequalities and matrices.

\* measured by NRT's available from commercial publishers

## Results

### *Criterion Referenced Tests*

Grade 4 CRT's across districts and standards contained anywhere from 1 item to 120 items per standard. The three districts' CRTs that were rated in the Eastern regional meeting included one district that was also rated in the Western regional meeting. As shown in table 3, District 1-E had between six and 23 items associated with the relevant standards. Three of the assessments (for standards 4.1.5, 4.3.4, and 4.6.2) were not broad enough to make any performance level classifications. For the remaining three standards, the assessment for standard 4.2.1 provided sufficient evidence to classify students as Advanced, Proficient, or Below Proficient (by inference); the assessments for standards 4.4.2 and 4.5.1 each permit classifications of Proficient and Below Proficient. The assessment for 4.5.1 also allows for the classification of Beginning.

Table 3. Consensus item ratings from the Eastern regional meetings for CRT's in Mathematics for Grade 4 for Standards 4.1.5, 4.2.1, 4.3.4, 4.4.2, 4.5.1, and 4.6.2 for Eastern districts (non-common)

Standard	District 1-E					District 2-E					District 3-E				
	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A
Standard 4.1.5	6	0	4	2	0	6	1	3	1	1	16	2	9	4	1
Standard 4.2.1	21	0	4	11	6	26	0	8	17	1	25	0	1	22	2
standard 4.3.4	6	0	3	3	0	1	0	0	1	0	30	0	9	11	10
Standard 4.4.2	10	2	3	5	0	19	4	10	5	0	24	0	10	14	0
Standard 4.5.1	23	4	8	10	1	20	3	4	13	1	25	0	5	18	2
Standard 4.6.2	8	0	1	4	3	8	0	4	4	0	25	2	6	10	7

As shown in table 4, in the Central region all standards had assessments that were rated except for District 3-C for standard 4.2.1 for which no standard was evaluated. The number of items per standard varied greatly in this region, from 1 to 120 tasks. In District 1-C, standard 4.1.5 had only a single assessment task, not allowing any performance classifications. That same district had 120 tasks covering standard 4.2.1, distributed across all performance categories, allowing classification into the four performance levels. The remaining assessments permit classification into two performance categories: standard 4.3.4 permits classification of students as either Progressing or Beginning (by inference); standard 4.4.2 permits classification as either Advanced or Below Advanced; and the assessments for standards 4.5.1 and 4.6.2 permit classification of students into Proficient or Below Proficient categories.

The assessments for District 2-C, provide some classification information for all standards (note that any inferences made for standard 4.4.2 should be with extreme caution). The assessment for standard 4.1.5 permits classification of students as either Advanced or Not Advanced. Standard 4.2.1's assessment with 40 items permits classification of students into all categories except Advanced, as does the assessment for standard 4.3.4. In contrast to the latter two assessments, the assessments for standards 4.4.2 and 4.5.1 permit classification of students as either Proficient or Below Proficient by the assessment used for standard 4.6.2.

Table 4. Consensus item ratings from the Central regional meetings for CRT's in Mathematics for Grade 4 for standards 4.1.5, 4.2.1, 4.3.4, 4.4.2, 4.5.1, and 4.6.2 for Central districts (non-common)

Standard	District 1-C					District 2-C					District 3-C				
	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A
Standard 4.1.5	1	0	0	1	0	11	2	2	1	6	14	0	1	1	12
Standard 4.2.1	120	44	30	34	12	40	6	10	21	3	NR				
standard 4.3.4	10	0	8	2	0	16	0	10	6	0	12	0	5	7	0
Standard 4.4.2	12	2	3	3	6	7	0	0	5	2	16	1	4	4	7
Standard 4.5.1	10	0	1	7	2	15	0	4	10	1	20	0	0	10	10
Standard 4.6.2	10	0	0	10	0	19	1	3	8	7	8	0	0	5	3

Table 5 displays item classifications for Grade 4 Western districts. District 1-W's assessments provide some opportunities to classify students into two or more performance levels for all six standards. The assessment for standard 4.1.5 permits classification of students as either Progressing or Beginning (by inference). The other assessment that permits only two levels of performance classification is for standard 4.5.1, which permits assigning students to either the Proficient or Below Proficient categories. The assessments for standards 4.2.1 and 4.3.4 permits classifying students into three performance levels, Advanced, Proficient, or Below Proficient and Proficient, Progressing, or Beginning (by inference), respectively. The assessments for both standards 4.3.4 and 4.6.2 permit classification into all four performance levels, but the Proficient classification for standard 4.4.2 should be made with caution, because it requires a high degree of inference.

Five of the assessments rated for District 2-W permit comfortable classification of students into two or more performance levels. These are the assessments for standards 4.2.1 (classifies students as Proficient or Below Proficient), 4.3.4 (classifies students as Progressing or Beginning-by inference), 4.4.2 (classifies students as Proficient or Below Proficient), 4.5.1 (classifies students as Progressing or Beginning - by inference\_, and 4.6.2 (classifies students as Proficient or Below).

Table 5. Consensus item ratings from the Western regional meetings for CRT's in Mathematics for Grade 4 for standards 4.1.5, 4.2.1, 4.3.4, 4.4.2, 4.5.1, and 4.6.2 for Western districts (non-common)

Standard	District 1-W					District 2-W					District 3-W				
	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A
Standard 4.1.5	16	1	10	4	1	5	0	3	2	0	5	0	4	1	0
Standard 4.2.1	25	0	2	13	10	10	0	3	7	0	18	0	3	11	4
standard 4.3.4	30	0	17	11	2	10	0	7	3	0	3	0	3	0	0
Standard 4.4.2	24	1	8	4	11	15	0	4	9	2	5	1	1	1	2
Standard 4.5.1	25	0	3	21	1	16	0	13	3	0	6	0	0	6	0
Standard 4.6.2	25	1	12	6	6	11	3	3	5	3	8	0	4	4	0

The two districts that had their assessment rated at all three regional meetings each had one standard for which no assessments were included (4.1.5 in District 1-All, and 4.6.2 in District 2-All). In District 1-all only two standards provided enough measurement information to make differential student classifications. Assessments for standards 4.2.1 and 6.6.2 permit classifying students as Advanced, Proficient, or Below

Proficient. The assessments from the remaining standards do not permit any classifications.

For District 2-All there are three standards for which classifications can be made. The assessment for standard 4.2.1 provides sufficient information for classifying students as Advanced, Proficient, or Below Proficient. For standard 4.4.2 a classification of either Proficient or Below Proficient is possible, whereas for standard 4.5.1 a classification of either Progressing or Beginning can be made. See table 6 for item ratings.

Table 6. Consensus item ratings for the district assessments rated in all three regions for Standards 4.1.5, 4.2.1, 4.3.4, 4.4.2, 4.5.1, and 4.6.2 (common districts).

Standard	District 1-All					District 2-All				
	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A
Standard 4.1.5	NR					10	1.7	1.7	4	2.7
Standard 4.2.1	47	0	4.3	27.7	15	28	0	2	7.3	18.7
Standard 4.3.4	4	0	2	2	0	4	0	1	2.3	0.7
Standard 4.4.2	10	2	2.7	1.7	3.7	17	1.3	4	9	2.7
Standard 4.5.1	7	0.3	3.3	2.3	1	28	2	6.3	15.3	4.3
Standard 4.6.2	17	0.7	0.7	10.3	5.3	NR				

The number of tasks per standard on each assessment on Grade 8 CRT's ranged from 4 to 92. As was the case in Grade 4, one district that was rated in the Eastern district was also rated in the Western district. Only one standard in one district was not rated. For most of the standards, the number of items was greater than 20.

In District 1-E the 12 assessment tasks that focus on standard 8.1.4 are divided evenly across the Proficient and Progressing categories, thus permitting classifying students into these two categories and into the Beginning category (by inference). Measurement opportunities are more limited for standards 8.2.2 and 8.3.2 such that only classifications of Progressing or Beginning (by inference) can be made. The large number of assessment tasks associated with standards 8.4.1 and 8.5.2 are distributed across all performance classifications, permitting students to be classified into all four performance levels. The five items for standard 8.6.3 are not sufficient to make any classification decision.

The one standard for which there were no items rated was standard 8.2.2 for District 2-E. Although all other standards for this district contained between 10 and 40 items, the assessments provided no information at the Advanced level. The assessments for standards 8.1.4, 8.3.2, 8.4.1, and 8.5.2 provided sufficient information to place students into Proficient, Progressing, and Beginning (by inference) levels. The assessment for standard 8.6.3 permits assigning students to Proficient or Below Proficient.

Each of the assessments rated for District 3-E provided enough information for multiple classifications. For standards 8.1.4, 8.2.2, 8.3.2, 8.4.1, and 8.5.2 the assessments permit classification into Proficient, Progressing, and Beginning levels. The Beginning classifications are based on inference. Standard 8.6.3 provides enough information to classify students into all four classifications.

Table 7. Consensus item ratings from the Eastern regional meetings for CRT’s in Mathematics for Grade 8 for Standards 8.1.4, 8.2.2, 8.3.2, 8.4.1, 8.5.2, and 8.6.3 for Eastern districts (non-common)

Standard	District 1-E					District 2-E					District 3-E				
	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A
Standard 8.1.4	12	0	6	6	0	16	1	6	7	2	22	0	14	8	0
Standard 8.2.2	8	0	6	2	0	NR					22	2	6	14	0
standard 8.3.2	13	0	10	3	0	30	0	22	8	0	23	0	11	8	4
Standard 8.4.1	60	8	27	15	10	27	6	16	5	0	27	1	20	6	0
Standard 8.5.2	40	8	17	7	8	40	6	20	13	1	26	2	10	11	3
Standard 8.6.3	5	0	2	3	0	10	0	4	6	0	26	1	5	12	8

Assessments from two districts were rated at all three meetings. Table 7 displays the average items per standard across the three meetings. In District 1-All, the assessments for standards 8.5.2 and 8.6.3 did not provide enough tasks to make any performance level classifications. The assessments for standards 8.1.4, and 8.3.2 permit classifying students as Proficient, Progressing, or Beginning. The assessments of standards 8.2.2 and 8.4.1 permit classifying students as Progressing or Beginning. In District 2-All, the assessment for standards 8.2.2, 8.5.2, and 8.6.3 do not provide enough information to classify students into any performance categories. The assessment for standard 8.1.4 contains 92 items; even so the items do not represent every category adequately making it possible only to classify students as Proficient, Progressing, and Beginning. The assessment for standard 8.3.2 allows classification into all four categories, while the one for 8.4.1 permits classification into the three lowest categories.

Table 8. Consensus item ratings for CRT’s in Mathematics for Grade 8 for Standards 8.1.4, 8.2.2, 8.3.2, 8.4.1, 8.5.2, and 8.6.3 for Common districts.

Standard	District 1-All					District 2-All				
	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A
Standard 8.1.4	29	0	23.7	5.7	0	92	1	56.7	29.7	4.7
Standard 8.2.2	16	3.7	9	3.3	0	4	0	0	4	0
Standard 8.3.2	20	0	12.3	7	0.7	42	1.7	15.7	15	9.3
Standard 8.4.1	40	7.7	28.7	3.7	0	26	0	15	9.7	1.3
Standard 8.5.2	7	2	2	2.3	0.7	6	0.7	2	3.3	0
Standard 8.6.3	4	0	1	2	1	5	0	2	3	0

District 1-W’s assessments permitted classifications at the three lowest performance categories for five of the six standards in this study, standards 8.1.4, 8.2.2, 8.3.2, and 8.5.2. The assessment for standard 8.4.1 had items that limited classification to only Progressing and Beginning performance levels. Standard 8.6.3’s assessment provided sufficient information to assign students to all four performance levels. District 3-W had only between 4 and 10 items per standard, making no student classifications possible. District 2-W was not much better, with anywhere from 5 to 15 items per standard. Table 8 displays Western district ratings.

Table 9. Consensus item ratings from the Western regional meetings for CRT's in Mathematics for Grade 8 for Standards 8.1.4, 8.2.2, 8.3.2, 8.4.1, 8.5.2, and 8.6.3 for Western districts (non-common)

Standard	District 1-W					District 2-W					District 3-W				
	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A
Standard 8.1.4	22	0	15	7	0	10	0	6	4	0	6	0	2	4	0
Standard 8.2.2	22	3	6	13	0	15	6	8	1	0	4	0	0	3	1
standard 8.3.2	23	1	11	11	0	14	0	4	6	0	8	0	2	6	0
Standard 8.4.1	27	4	21	2	0	14	5	9	0	0	8	0	5	3	0
Standard 8.5.2	26	4	13	5	4	5	1	2	2	0	7	0	3	3	1
Standard 8.6.3	26	1	9	2	14	12	0	7	2	3	10	0	1	7	2

Table 9 displays data for two districts' assessments that were rated at the Central meeting. The assessment for standard 8.1.4 in District 2-C can be used to classify students as either Proficient or Below Proficient. No classification decisions can be made with the items associated with 8.2.2. For standard 8.3.2, 8.4.1, and 8.6.3 students can be classified as Progressing or Beginning using the appropriate assessments. For standard 8.5.2, a classification of Above Beginning could be made for students who answer more than 6 items correctly.

In District 3-C, three standards had no assessments rated. The items associated with standard 8.3.2 permit classification of students into Proficient, Progressing, and Beginning (by inference). The assessment for standard 8.4.1 permits classification into the categories Progressing and Beginning. Standard 8.6.3's assessment provides the opportunity to say that students are Above Progressing, Progressing, or Beginning (by inference).

Table 10. Consensus item ratings from the Central regional meetings for CRT's in Mathematics for Grade 8 for Standards 8.1.4, 8.2.2, 8.3.2, 8.4.1, 8.5.2, and 8.6.3 for Central districts (non-common)

Standard	District 2-C					District 3-C				
	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A
Standard 8.1.4	15	0	4	11	0	NR				
Standard 8.2.2	5	0	2	3	0	NR				
Standard 8.3.2	12	0	8	4	0	15	1	6	8	0
Standard 8.4.1	32	7	25	0	0	24	3	19	2	0
Standard 8.5.2	8	0	3	5	0	NR				
Standard 8.6.3	15	2	8	4	1	22	3	13	5	1

High school CRT's contained anywhere from 2 to 26 items covered by an assessment per standard. As was the case in Grade 4 and 8, two districts were rated in all three meeting locations. Table 11 provides information on item distributions from the Eastern regional meeting. In District 1-E, two of the standards contain only seven tasks, permitting no classifications to be made. Standard 12.4.5 contains 11 tasks too diverse in difficulty to make any decisions. The assessments for standards 12.1.2 and 12.2.1 can classify students as either Progressing or Beginning, the assessment for standard 12.3.1 permits only an inference about Beginning students.

District 2-E had two assessments that were not rated (for standards 12.3.1 and 12.5.1) and two other standards 12.1.2 and 12.4.5) that had too few items to permit any

student performance level classification. For standard 12.2.1 the assessment permits three classification of student performance, Advanced, Proficient, or Below Proficient (by inference). The assessment for standard 12.6.3 permits classifying students as either Proficient or Below Proficient.

All except one assessment from District 3-E permitted at least one level of performance classification. The one exception is the assessment for standard 12.5.1. The assessment for standard 12.3.1 permits assigning students to only the Beginning level of performance. Students can be assigned to either the Proficient, Progressing, or Beginning performance levels based on the assessments for standards 12.1.2 and 12.2.1, whereas for standards 12.4.5 and 12.6.3 students can be classified only as Proficient or Below Proficient.

Table 11. Consensus item ratings from the Eastern regional meetings for CRT’s in Mathematics for High School for Standards 12.1.2, 12.2.1, 12.3.1, 12.4.5, and 12.6.3 for Eastern districts (non-common).

Standard	District 1-E					District 2-E					District 3-E				
	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A
Standard 12.1.2	15	6	9	0	0	5	1	2	2	0	19	1	10	6	2
Standard 12.2.1	17	6	11	0	0	20	0	2	11	7	16	1	8	7	0
Standard 12.3.1	9	5	4	0	0	NR					14	9	3	2	0
Standard 12.4.5	11	4	4	2	1	3	0	3	0	0	16	0	2	11	3
Standard 12.5.1	7	0	2	3	2	NR					5	1	3	1	0
Standard 12.6.3	7	3	1	2	1	10	0	0	8	2	12	0	1	9	2

The reviews of the assessments from the three non-common districts reviewed in the Western regional meeting are shown in Table 12. Only the assessments for district 1-W provide sufficient information for making any performance level decisions. In District 1-W, for standard 12.1.2 students may be classified as Proficient, Progressing, or Beginning (by inference) For standard 12.2.1 students may be classified as Proficient or Below Proficient, whereas for standard 12.3.1 students may be assigned to either Progressing or Beginning levels. The assessments for standards 12.4.5 and 12.6.3 permit classifications of students into Proficient or Below Proficient categories, while the assessment for standard 12.5.1 does not allow any classification. Although there are not enough items at any single level in District 2-W or 3-W to make classification decisions, for standard 12.3.1 in District 2-W students may be classified as at least Progressing if they answer most of the questions correctly. This is also the case for the same standard in District 3-W.

Table 12. Consensus item ratings from the Western regional meetings for CRT’s in Mathematics for High School for Standards 12.1.2, 12.2.1, 12.3.1, 12.4.5, and 12.6.3 for Western districts (non-common).

Standard	District 1-W					District 2-W					District 3-W				
	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A
Standard 12.1.2	19	0	10	5	4	8	3	2	3	0	5	0	4	1	0
Standard 12.2.1	16	0	5	8	3	6	1	4	1	0	5	0	0	5	0
Standard 12.3.1	14	7	4	3	0	10	3	4	3	0	7	1	3	3	0
Standard 12.4.5	6	0	1	12	3	5	0	0	5	0	5	0	4	0	1
Standard 12.5.1	5	4	0	1	0	5	5	0	0	0	5	0	2	3	0
Standard 12.6.3	12	0	1	9	2	5	0	0	5	0	5	0	0	5	0

In District 1-All, at least two levels of classification can be made for all standards except 12.5.1, for which no classifications are possible. The assessments for standards 12.1.2 and 12.3.1 permit assigning students as either Progressing or Beginning performance levels. For standards 12.2.1, 12.4.5, and 12.6.3, the assessments permit classifications of Proficient or Below Proficient. None of the assessments permit a classification of Advanced.

Three standards (12.2.1, 12.3.1, and 12.5.1) were not rated for District 2-All. Of the remaining standards, the assessment for standard 12.1.2 permits classifications of Advanced or Below Advanced. The assessment for standard 12.6.3 permits classifications of Proficient or Not Proficient. The assessment for standard 12.4.5 permits three levels of classification, Proficient, Progressing, and Beginning (by inference).

Table 13. Consensus item ratings from the Eastern regional meetings for CRTs in Mathematics for High School for Standards 12.1.2, 12.2.1, 12.3.1, 12.4.5, 12.5.1, and 12.6.3.

Standard	District 1-All					District 2-All				
	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A
Standard 12.1.2	17	3.7	12	1.3	0	14	0	0	3.3	10.7
Standard 12.2.1	15	0	3.3	8.3	3.3	NR				
Standard 12.3.1	12	2.7	8	1.3	0	NR				
Standard 12.4.5	11	0	3	6	2	34	1.3	17.3	13	2.3
Standard 12.5.1	7	1	1.7	2.7	2	NR				
Standard 12.6.3	16	0	0.3	14	1.7	11	0	0.3	8.3	2.3

Table 14 contains the data for the two non-common districts that provided mathematics assessments for review. Two assessments permit performance level classifications for District 1-C. The assessments for standard 12.1.2 and 12.3.1 permit classifications of Progressing, or Beginning. The other two assessments that were rated for standard 12.4.5 and 12.6.3 do not permit classification.

In District 2-C, students can be classified on standard 12.1.2 as Progressing or Below Progressing. The assessment for standard 12.2.1 permits classification as Proficient, Progressing, or Beginning (by inference). The assessment for standard 12.4.5 permits classification as Proficient or Below Proficient. Assessments for the remaining standards (12.3.1, 12.5.1, and 12.6.3) do not provide enough measurement information to make any performance level judgments.

Table 14. Consensus item ratings from the Central regional meetings for CRTs in Mathematics for High School for Standards 12.1.2, 12.2.1, 12.3.1, 12.4.5, 12.5.1, and 12.6.3

Standard	District 1-C					District 2-C				
	Items	B	Prog	Prof	A	Items	B	Prog	Prof	A
Standard 12.1.2	15	0	15	0	0	11	2	6	3	0
Standard 12.2.1	NR					26	0	15	11	0
Standard 12.3.1	12	4	7	1	0	6	3	3	0	0
Standard 12.4.5	5	0	0	5	0	11	0	5	4	2
Standard 12.5.1	NR					7	0	3	4	0
Standard 12.6.3	2	0	1	1	0	4	0	0	3	1

Table 15 summarizes tables three through 14. For Grade 4 assessments, it was most common to have sufficient information to make only two performance level classifications. Assessments for standards at the High School level displayed a lack of sufficiency of information, and it was most common to not be able to make any performance classifications. No standard contained sufficient breadth and depth to classify students into all four levels for Grade 12. Grade 8 assessment results were more promising, with three classifications most commonly permitted, yet only 5 instances was there sufficient information to make all four performance decisions.

Table 15. Summary of number of Performance Classifications Permitted at each meeting by grade level on CRTs for standards covered by assessments.

Performance Classifications Permitted					
Grade	Meeting	0-1	2	3	4
4	Eastern	6	6	4	2
	Central	0	11	4	1
	Western	6	8	2	2
	All	3	2	3	0
	Total	15	27	13	5
8	Eastern	1	3	10	3
	Central	1	6	2	0
	Western	2	9	6	1
	All	5	2	5	1
	Total	9	20	23	5
12	Eastern	7	7	3	0
	Central	5	4	1	0
	Western	13	4	1	0
	All	1	7	1	0
	Total	26	22	6	0

### *Norm Referenced Tests*

As shown in table 16, there are two subtests in which items were aligned with standards 4.2.1 and 4.5.1. For standard 4.2.1 there are 33 items and for standard 4.5.1 there are 12 items. The 33 items associated with standard 4.2.1 are concentrated in the Progressing and Proficient categories, permitting classification of students as being

Proficient, Progressing, or Beginning (by inference). The items aligned with standard 4.5.1 also permit three levels of classification, but these levels are Advanced, Proficient, and Below Proficient (by inference).

Table 16. Consensus item ratings from the Central regional meeting for NRT-2 in Mathematics for Grade 4 for Standards 4.2.1 and 4.5.1.

Subtest	Items	Beginning	Progressing	Proficient	Advanced
<b>Math Concepts and Problem Solving</b>					
Standard 4.2.1	11	0	1	8	2
Standard 4.5.1	12	0	0	6	6
<b>Math Computation</b>					
Standard 4.2.1	22	0	10	12	0
<b>Overall</b>					
Standard 4.2.1	33	0	11	20	2
Standard 4.5.1	12	0	0	6	6

NRT-3 of Grade 4 had a total of 35 items that had been aligned with standard 4.2.1 and 12 items with standard 4.5.1. The 35 items that measure 4.2.1 are distributed across the three highest performance levels and there are enough items at each level to make performance level classifications into all four categories (placement into the Beginning category by inference). For standard 4.5.1 there are only 8 items and permit students to be classified as Proficient or Below Proficient.

Table 17. Consensus item ratings from the Western regional meeting for NRT-3 in Mathematics for Grade 4 for Standards 4.2.1 and 4.5.1.

Subtest	Items	Beginning	Progressing	Proficient	Advanced
<b>Math Concepts and Problem Solving</b>					
Standard 4.2.1	21	0	7	9	5
Standard 4.5.1	12	0	0	5	3
<b>Math Computation</b>					
Standard 4.2.1	14	0	4	6	4
<b>Overall</b>					
Standard 4.2.1	35	0	11	15	9
Standard 4.5.1	8	0	0	5	3

For Grade 8, NRT-1 (shown in table 18) has one subtest that includes only 14 items related to each of the two standards (8.2.2 and 8.5.2) appropriate to this study. The preponderance of items associated with standard 8.2.2 are at the Proficient level of performance, suggesting that students can be classified as either Proficient or Below Proficient. For standard 8.5.2, however, the items provide the opportunity to classify students as Proficient, Progressing, or Beginning (by inference).

Table 18. Consensus item ratings from the Eastern regional meeting for NRT-1 in Mathematics for Grade 8 for Standards 8.2.2 and 8.5.2.

Subtest	Items	Beginning	Progressing	Proficient	Advanced
<b>Math Concepts and Data Interpretation</b>					
Standard 8.2.2	14	0	3	10	1
Standard 8.5.2	14	1	6	6	1
<b>Overall</b>					
Standard 8.2.2	14	0	3	10	1
Standard 8.5.2	14	1	6	6	1

NRT-2 for Grade 8 was rated in the Central regional meeting. This NRT has two subsets that contain items aligned to standard 8.2.2 and one subtest having items related to standard 8.5.2. The six items aligned with standard 8.5.2 are distributed too widely to provide enough measurement information to make any performance level classifications. The 25 items (across both subsets) aligned with standard 8.2.2 permit classification into Proficient, Progressing, and Beginning (by inference) categories.

Table 19. Consensus item ratings from the Central regional meeting for NRT-2 in Mathematics for Grade 8 for Standards 8.2.2 and 8.5.2.

Subtest	Items	Beginning	Progressing	Proficient	Advanced
<b>Math Concepts and Problem Solving</b>					
Standard 8.2.2	10	0	3	6	1
Standard 8.5.2	6	0	3	2	1
<b>Math Computation</b>					
Standard 8.2.2	15	3	3	9	0
<b>Overall</b>					
Standard 8.2.2	25	3	6	15	1
Standard 8.5.2	6	0	3	2	1

Table 20 displays item ratings from the one subtest that includes Mathematics items associated with standards 8.2.2 and 8.5.2. There are 10 items aligned with each of these standards. The items for standard 8.2.2 are distributed in such a way that permits classification of Above Progressing (for those who answer almost all of the items correctly) or Below Proficient. The 10 items associated with standard 8.5.2 permit classifying students as either Proficient or Below Proficient.

Table 20. Consensus item ratings from the Western regional meeting for NRT-3 in Mathematics for Grade 8 for Standards 8.2.2 and 8.5.2.

Subtest	Items	Beginning	Progressing	Proficient	Advanced
<b>Mathematics Subset</b>					
Standard 8.2.2	10	0	2	5	3
Standard 8.5.2	10	0	3	6	1
<b>Overall</b>					
Standard 8.2.2	10	0	2	5	3
Standard 8.5.2	10	0	3	6	1

Table 21 displays the teachers' ratings of NRT-1 for high school standards 12.2.1 and 12.5.1. Items in two subsets are aligned with standard 12.2.1 resulting in a total of 28 aligned items. These 28 items were judged to be evenly divided between the Proficient and Progressing categories, thus permitting classifying students as either Proficient, Progressing, or Beginning (by inference). There were only 7 items aligned with standard 12.5.1 and these items ranged across three performance levels without sufficient focus to make any classification decisions.

Table 21. Consensus item ratings from the Eastern regional meeting for NRT-1 in Mathematics for Grade 12 for Standards 12.2.1 and 12.5.1.

Subtest	Items	Beginning	Progressing	Proficient	Advanced
<b>Mathematics: Concepts and Problem Solving</b>					
<b>Standard 12.2.1</b>	17	0	7	9	1
<b>Standard 12.5.1</b>	7	3	1	3	0
<b>Math Computation</b>					
<b>Standard 12.2.1</b>	11	0	7	4	0
<b>Overall</b>					
<b>Standard 12.2.1</b>	28	0	14	13	1
<b>Standard 12.5.1</b>	7	3	1	3	0

NRT-2 for high school has only one subtest that includes items that were aligned with standards 12.2.1 and 12.5.1 and there are 10 and six items, respectively, associated with these two standards. The majority of items associated with standard 12.2.1 are at the Progressing level, thus permitting classification of students at that level and at the Beginning level (by inference). The six items aligned with standard 12.5.1 are distributed across the Beginning, Progressing, and Proficient performance levels prohibiting any determination of student performance levels.

Table 22. Consensus item ratings from the Central regional meeting for NRT-2 in Mathematics for Grade 12 for Standards 12.2.1 and 12.5.1.

Subtest	Items	Beginning	Progressing	Proficient	Advanced
<b>Mathematics/ Overall</b>					
<b>Standard 12.2.1</b>	10	1	7	1	1
<b>Standard 12.5.1</b>	6	2	2	2	0

NRT-3 provides 29 items across two subtests that are aligned with standard 12.2.1 (presented in table 23). Most of these items are at the Progressing and Proficient performance levels, permitting classifications at those levels and at the Beginning level (by inference). The 16 items aligned with standard 12.5.1 are mostly at the Progressing level, permitting classification at the Progressing and Beginning levels.

Table 23. Consensus item ratings from the Western regional meeting for NRT-3 in Mathematics for Grade 12 for Standards 12.2.1 and 12.5.1.

Subtest	Items	Beginning	Progressing	Proficient	Advanced
<b>Mathematics: Concepts and Problem Solving</b>					
Standard 12.2.1	13	2	6	3	2
Standard 12.5.1	8	6	2	0	0
<b>Math Computation</b>					
Standard 12.2.1	16	0	12	4	0
<b>Overall</b>					
Standard 12.2.1	29	2	18	7	2
Standard 12.5.1	8	6	2	0	0

Item ratings for NRTs revealed similar information as the CRT ratings. The number of possible performance classifications varied greatly by standard, district, and grade. Only in one situation was there sufficient information to permit classification into all four levels. It was more frequently the case that two, three, or zero classifications could be made.

### Discussion

This study revealed that for these assessments there was often insufficient information to support interpretations of performance among four performance levels. The results often support the use of two or three performance categories. For some tests, there was not enough measurement information to classify students into two categories, based on Subkoviak's (1988) recommendation of at least 6 items. For most standards across grade levels, results consistently displayed more items or tasks classified as either *proficient* or *progressing*, and fewer items as *advanced* or *beginning*. It was not unusual to have zero items present in one or more performance level. In some instances there were only 1 or 2 items, making it impossible to make any inferences about a student's performance level. There was great variability in the total number of items present for each standard.

Insufficiency is especially a concern in any high stakes educational situation, such as NCLB, where students are required to be classified into multiple performance categories. This is the case in the state examined in this study, as well as many other states. When assessments do not contain a sufficient number of items distributed adequately across all performance categories, there are likely many students being misclassified, and resulting decisions regarding individual students and schools may not be appropriate. For instance, if a student is misclassified into a higher performance category than they truly belong, and extra help that may be beneficial to this student not provided. This is also a concern for school districts since federal assistance is partially reliant on student classifications.

Performance classification is also a concern at the classroom level; an example of multiple classroom proficiency levels is the use of a letter grading system (e.g. A, B, C, D, and F). Teachers need to be aware that when making these classifications, it is important to have a sufficient number of items, and that each performance category is represented by these items. For instance, there should be sufficient items to distinguish between an "A" student and a "B" student as well as distinguishing a "D" and an "F" student. This may not be a concern in low stakes situations such as frequent classroom quizzes (Popham, 1990). In these situations multiple quizzes are combined to make up a

decision about a student's classification. For situations when one test is important in making decisions, there should be greater concern about the validity of the assessment in terms of item sufficiency.

The issue of validity in the classroom is also related to GPA; having sufficient information to distinguish between ability levels based on this cumulative number. Smith (2003) notes that high school grades play a vital role in college admission, and therefore these grades should provide accurate information concerning the abilities of students. Colleges are often interested in student's high school GPA, along with class rank. One problem related to admissions decisions based on these criteria is grade inflation. When a large proportion of students obtain very high marks, the differences between the students are so small that highly selective colleges have trouble making selections. Thus for these types of decisions, it is important that high school grades have the ability to make accurate classifications of students (Smith, 2003).

The analyses in the current study were based on suggestions of item sufficiency for dichotomous decisions. In many academic situations decisions have become more complex and require multiple classifications. Future research should attempt to determine appropriate rule of thumb guidelines for item sufficiency when applying multiple cut-points in an assessment. This would be beneficial information for teachers and others involved in assessment development to create tests which properly classify students.

### References

- Bohla, D. S., Impara, J. C., and Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, 22(3), 21-29.
- Brookhart, S. M. (2005). The quality of local district assessments used in Nebraska's school-based teacher-led assessment and reporting system. *Educational Measurement: Issues and Practice*, 24(2), 14-21.
- Linn, R. L. (2002). Validation of the uses and interpretations of results of state assessments and accountability systems. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation*. (pp. 27-48). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Millman J., & Greene, J. (1993). The specification and development of tests of achievement and ability. In R. L. Linn. *Educational Measurement (3<sup>rd</sup> ed)*. (pp. 335-366). Phoenix, Arizona: Oryx Press.
- Ryan, J. M., & DeMark, S. (2002). Variation in achievement scores related to gender, item format, and content area tested. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation*. (pp. 67-88). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Popham, J. W. (1990). *Modern educational measurement: A practitioner's perspective (2<sup>nd</sup> ed.)*. New Jersey: Prentice Hall
- Smith, J. K. (2003) Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice* 22(4), 26-33.
- Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, 25(1), 47-56.
- Thorndike, R. M. (2005). *Measurement and Evaluation in Psychology and Education (7<sup>th</sup> Ed.)*. Upper Saddle River, New Jersey: Pearson.