DOCUMENT RESUME

ED 442 818 TM 031 238

AUTHOR	Brown, Richard S.
TITLE	Using Latent Class Analysis To Set Academic Performance Standards.
PUB DATE	2000-04-00
NOTE	44p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150) -- Tests/Questionnaires (160)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Achievement; *Academic Standards; *Junior High School Students; Junior High Schools; Mathematics; Performance Based Assessment; Performance Factors; Responses; *Student Evaluation
IDENTIFIERS	*Latent Class Analysis; *Standard Setting

ABSTRACT

The use of latent class analysis for establishing student performance standards was studied. Latent class analysis (LCA) is an established procedure for investigating the latent structure of a set of data. LCA presumes that groups, classes, or respondents differ qualitatively from one another, and that these differences account for all of the relationships in the data. A student assessment instrument was developed consisting of a 10-item multiple choice component and 2 performance assessment tasks. The instrument was completed by 191 seventh and eighth grade mathematics students. The latent class procedures provided information regarding the qualitative differences among the student responses. A series of latent class models was explored using binary and continuous indicators. For this sample of student responses, the LCA procedure indicated that a two-group structure was the most appropriate model for explaining the differences in student performance. Perhaps the most important finding from the study deals with the extent to which the varying standard-setting procedures rendered comparable conclusions regarding what constitutes proficient performance. Another implication of the findings is that if empirical methods can be shown to render determinations regarding student proficiency that are comparable to the more common, and more costly, judgmental approaches, these methods could be used more often to support the determinations made with the judgmental approaches. An appendix contains the developed student assessment instrument. (Contains 2 figures, 10 tables, and 70 references.) (SLD)

Using Latent Class Analysis to Set Academic Performance Standards

Richard S. Brown
National Center for Research on Evaluation, Standards, and Student Testing
University of California, Los Angeles

Paper presented at the annual meeting of the
American Educational Research Association

New Orleans, LA
April, 2000

RUNNING HEAD: LATENT CLASS ANALYSIS AND PERFORMANCE STANDARDS

Using Latent Class Analysis to Set Academic Performance Standards

Richard S. Brown
UCLA/CRESST

Introduction

Over the years, a variety of techniques have been proposed for determining appropriate performance standards (see Berk, 1986, for a review). Although these procedures vary widely, generally speaking, subject matter experts are gathered together and asked to make judgments about what level of performance on a specified task or examination reflects a given level of competence. These judgments may be reached using any number of methods, and often vary depending on a number of factors, including the methods employed (Andrew & Hecht, 1976; Behuniak, et al., 1982; Berk, 1986; Bowers & Shindoll, 1989; Brennan & Lockwood, 1980; Brown, 1993; Cantor, 1989; Cross, et al., 1984; Garrido, 1985; Impara & Plake, 1997; Koffler, 1980; Melican & Plake, 1985; Mills, 1983; Norcini, et al., 1987; Plake, 1995; Reilly, et al, 1984; Rock, Davis, & Werts, 1980; Skakun & Kling, 1980; Smith & Smith, 1988), the subject matter experts chosen to participate (Garrido, 1985; Hamberlin, 1985; Hurtz, Sanders, & Hurtz, 1996; Longford, 1996; Plake, Impara, & Potenza, 1994; Smith & Smith, 1988), and the characteristics of the testing instrument (Norcini, 1987; Plake & Melican, 1989; Smith, 1987; Taylor, 1987). Most often, the testing instruments used have been comprised of multiple-choice items of varying levels of difficulty.

However, more recent investigations into standard setting procedures using more complex, performance based assessments have appeared (Hambleton & Plake, 1995; Jaeger, 1995; Luecht & DeChamplain, 1998; Putnam, Pence, & Jaeger, 1995). In fact, the journal Applied Measurement in Education recently devoted an entire special issue to the topic of standard setting for complex performance tasks (Volume 8, Number 1, 1995), and followed up with an additional special issue dedicated to much the same idea, titled

Setting Consensus Goals for Academic Achievement (Volume 11, Number 1, 1998). In it, the editors emphasize the emerging importance of standard setting procedures using complex performance tasks by stating, "As the field of measurement moves toward increasingly frequent administration of performance tests and other measures of complex behavior for decision making about individuals, be it in elementary school settings or in licensure or certification settings, the need to undertake sound practices for standard setting is essential" (Impara & Plake, 1995, p. 1). This study explores one approach to setting performance standards in just such a context.

The idea of performance standards has evolved over time. Until and including the 1970's setting performance standards generally meant establishing a cut-score on some continuous measure of achievement (often a scale score) that was based on defined content standards for a given content domain. Since that time, performance standards have included the use of "anchor items" to demonstrate performance at arbitrarily determined points along the achievement continuum. With the 1988 reauthorization of the National Assessment of Educational Progress (NAEP), performance standards were established as categorical achievement levels "presented with descriptions of varying detail and exemplar items in order to give them meaning to lay audience" (Linn, Koretz, & Baker, 1996, p. 26). These achievement levels reflect judgments about how students at a given grade level in a particular subject area should perform. The use of these achievement levels for establishing and reporting performance standards has had both positive and negative effects. Although the lay audience can better grasp the meaning of the performance standards using these levels than they otherwise would using only scale score values, some reports in the popular press oversimplify the achievement levels and misrepresented student achievement as discontinuous (Linn, Koretz, & Baker, 1996). Nevertheless, achievement levels continue to be one basis for evaluating student academic performance on the NAEP assessment, despite arguments regarding the adequacy of their descriptions, their validity, and their credibility (Burstein, et al, 1993, Linn, Koretz, Baker, & Burstein, 1991; Sugrue, et al., 1995)

There is little doubt that performance assessments are increasingly becoming a part of a comprehensive assessment system for student achievement or that standards are becoming more and more important (Baker & Linn, 1997). Recent initiatives in most states and in many of the largest public school districts in the nation have included performance assessments along with standardized, norm-referenced tests, as part of their comprehensive student assessment systems.

This trend in assessment and the need to understand how various components of the standard setting procedure impact the quality of standards motivates this study. Many of the most frequently used methods for setting achievement levels use judgmental approaches, meaning they use human raters, presumably experts, to make judgments or decisions about where the standard should be established. Given that human judgment is notoriously fallible and subject to a variety of influences (Dawes, Faust & Meehl, 1989; Gilovich, 1991; Kahneman, Slovic, & Tversky, 1982; Nisbett & Ross, 1980), it is worthwhile to investigate non-judgmental approaches. This study intends to do just that, by exploring the use of latent class analysis for establishing student performance standards.

Standard Setting Approaches

The area of developing and applying performance standards has been an active field for research for several decades. Early procedures for unidimensional, multiple-choice tests were developed over forty years ago (Nedelsky, 1954), and expanded upon greatly over the past twenty years. Initial standard setting procedures dealt primarily with the assessment instrument at the item level. Later approaches either incorporated additional information, or focused on a different type of information altogether. For example, recent approaches involve judgments of response profiles rather than individual assessment items.

One of the earliest procedures, the Nedelsky (1954) method, asks a panel of experts to identify response alternatives in a multiple-choice that a "minimally competent" respondent would recognize as incorrect. The expected chance score is then computed from the remaining response alternatives for each item and summed across items to yield the cut-score or performance standard.

Other early, but seminal work in this area was conducted by William Angoff (1971), whose initial standard setting approach is the most widely used and modified procedure to date (Sireci & Biskin, 1992). This approach requires expert judges to estimate the probability that a minimally competent respondent would correctly answer a given item. The estimates for all items comprising the assessment are summed, and a cumulative passing score is established. This procedure has been lauded for its ease of administration and facility in describing to expert raters.

Additional methods were proposed by Ebel (1972) and Jaeger (1982). The Ebel (1972) approach is more cumbersome and has received less use than the previously mentioned methods, requiring judges to sort items into categories according to perceived difficulty and relevance. The Jaeger (1982) method differs slightly from the earlier protocols in that it asks judges to identify which items an examinee *should* be able to answer correctly, rather than the likelihood or probability that an examinee *would* get the answer correct. Under this approach, the cut-score is determined as the sum of those items the respondent should answer correctly, as opposed to the sum of item probabilities in the Angoff method.

Other approaches have been suggested which deal with judges evaluating examinee responses rather than assessment items. The contrasting group approach (Livingston & Zieky, 1982) requires raters to categorize respondents into either clearly masters or clearly non-masters, and the cut-score is identified as a point somewhere between the distributions of the two identified groups. Similarly, the borderline group method (Livingston & Zieky, 1982) requires categorization of respondents, but in this case judges are asked to identify those respondents who are clearly neither masters nor non-masters. The cut-score is then established at or near the mid-point of the distribution of scores for this mid-range group. More recently developed but less widespread approaches include judgmental policy capturing (Jaeger, 1995) and the dominant profile method (Plake, Hambleton, & Jaeger, 1997; Putnam, Pence, & Jaeger, 1995), which seek to determine which characteristics of the examinees' performance on multi-dimensional assessments lead to decisions regarding levels of mastery.

Several researchers have looked into this mass of varying procedures. The most comprehensive review of standard setting methodology was conducted by Berk (1986; see also Jaeger, 1989). In it, he identified 38 different standard setting procedures, though many were derivatives of other approaches, notably the Angoff and Ebel procedures. Berk concludes his review by suggesting that a modified Angoff approach may be the procedure best suited for certification testing applications.

In another review of performance standard setting procedures, Cascio et al., (1988) provide an insight into the legal issues around setting cutoff scores in addition to looking at the psychometric and methodological issues. These researchers provide an interesting review of how the courts have weighed in on the issue of setting standards via cut-scores on tests and assessments. Still other researchers have commented on the area, offering guidelines and advice for standard setting procedures (Cizek, 1996; Maurer, et al., 1991, Mills, 1995; Shepard, 1980).

Many investigations have been undertaken comparing standard setting procedures (Andrew & Hecht, 1976; Behuniak, et al., 1982; Berk, 1986; Bowers & Shindoll, 1989; Brennan & Lockwood, 1980; Brown, 1993; Cantor, 1989; Cross, et al., 1984; Garrido, 1985; Impara & Plake, 1997; Koffler, 1980; Melican & Plake, 1985; Mills, 1983; Norcini, et al., 1987; Plake, 1995; Reilly, et al, 1984; Rock, Davis, & Werts, 1980; Skakun & Kling, 1980; Smith & Smith, 1988; Van der Linden, 1982). Generally, these studies show that different standard setting approaches yield differing standards (see Jaeger, 1989). For example, the Nedelsky procedure has been shown in several investigations to produce more lenient standards than the Angoff approach (Andrew & Hecht, 1976; Brennan & Lockwood, 1980; Rock, Davis, & Werts, 1980; Skakun & Kling, 1980). However, the research is far from conclusive. An interesting recent meta-analysis came to a different conclusion altogether, suggesting that "different standard setting procedures do not systematically yield different cut scores." (Bontempo, Marks, & Karabatsos, 1998, p. 3).

Latent Class Approach

One method to establish performance standards which differs considerably from the earlier judgmental approaches is the purely empirical latent class approach. Latent class analyses (LCA) is an established procedure for investigating the latent structure of a set of data (Bergan, 1983; Bergan & Stone, 1985; Dayton & MacReady, 1976), and has been used to assess latent structures using achievement items (Dayton, 1991; Haertel, 1984, 1989; Luecht & DeChamplain, 1998). This procedure differs in many of the underlying assumptions of the previous methods. The previous approaches assume that a continuous, unmeasured trait underlies student performance, and that somewhere along that continuum there is at least one location, or cut-point, where a meaningful distinction should be made.

In contrast to such approaches, latent class analyses does not presume that a continuous trait underlies performance, but rather that groups, or classes, of respondents differ *qualitatively* from one another, and that these differences account for all of the relationships in the data. Models specifying varying numbers of latent classes can be fit to the data, parameters estimated, and the model tested to see how well these proposed structures capture the relationships among the data. Moreover, whereas previous approaches presume a consistent item response probability for all respondents (i.e., item difficulty parameters are estimated to be the same for all students), latent class analysis allows for differing item response probabilities across classes, while retaining the assumption that within classes, item responses are independent.

In general, latent class analysis seeks to identify the number of latent classes, the proportion of subjects in each latent class, and the conditional item probabilities within each class. LCA can also make predictions regarding class membership for each response pattern. Using the notation expressed in Dayton (1991), let:

y_i = (y_{ij}) be the vector of 1/0 responses by the i^{th} respondent (i = 1,..., n) to the k items (j = 1,...,k).

α_{jc} = the conditional item probability of item j in latent class c, where (c = 1,…,C)

Θ_c = the proportion of respondents in each latent class. The sum of these proportions across all classes must sum to 1.

Then, the conditional probability of a response pattern given a particular latent class is estimated using the following product-multinomial:

$$P(y_i \mid c) = \prod_{j=1}^{k} (\alpha_{jc})^{y_{ij}} \bullet (1-\alpha_{jc})^{1-y_{ij}}$$

The unconditional probability of a given response pattern is estimated by using a weighted sum (weighted by the corresponding latent class proportion) across all latent classes:

$$P(y_i) = \sum_{c=1}^{C} \Theta_c \left[\prod_{j=1}^{k} (\alpha_{jc})^{y_{ij}} \bullet (1-\alpha_{jc})^{1-y_{ij}} \right]$$

The probability of membership in a latent class given a particular response pattern is then estimated using Bayes' Theorem:

$$P(c \mid y_i) = \frac{P(y_i \mid c) \bullet P(c)}{\sum_{c=1}^{C} P(y_i \mid c) \bullet P(c)}$$

Thus, by using latent class analysis, it is possible to use a sample of student responses to accomplish four goals: determine the extent to which a specified latent structure fits student performance data; determine which latent structure best represents the relationships in the data; obtain estimates of item parameters for each latent class; and identify which class within that latent structure each response pattern most likely belongs. Individual students could thereby be assigned to a given class based on their response pattern, and these categorizations could then be compared to assignments made using other performance standard setting procedures.

Procedures

Student Assessment Instrument

A student assessment instrument was developed consisting of a 10-item multiple choice component and 2 performance assessment tasks. The multiple choice component was created from the probability and data representation cluster of items from the Third International Mathematics and Science Study (TIMSS) released item set for seventh and eighth grade mathematics students. These items range in difficulty from fairly easy (p-value = .85) to rather difficult (p-value = .41).

The two performance assessment items were developed specifically to address probability and statistics knowledge for seventh grade students as part of a Center for Research on Evaluation, Standards, and Student Testing (CRESST) assessment development project for the Los Angeles Unified School District (LAUSD). The first performance assessment item is titled "A New Plan" and deals with probability by asking students to develop a plan for rendering a specific decision using three coins all tossed at once. This item draws upon students understanding of how to identify all possible outcomes of tossing three coins simultaneously, determining the probability of each outcome, and developing a decision rule for ensuring fair and equitable judgments for the participants.

The second performance assessment item, titled "The Food Spinner", also deals with probability, but in a slightly different way. In this exercise, students are provided with an image of a spinner device on which various food selections occupy divided up regions of a circular space. The regions of the circle for each food selection are not equal in size. In addition, students are presented with a table of results from 20 spins of the spinner, showing the number and percentage of times the spinner landed on each food selection. Students are asked to explain whether or not the observed frequencies are consistent with what they would expect, and if what they would expect the results to look like following an additional 100 spins. This task deals with students' understanding of theoretical probabilities and variations from those probabilities as a function of sample size.

These performance items were scored on a scale of 1 to 4, using an adaptation of a rubric developed and validated in previous CRESST research (Baker, Freeman, & Clayton, 1991; Niemi, 1996). A complete copy of the assessment instrument is provided in the Appendix.

Student answers to this assessment instrument were obtained from several Los Angeles area junior high or middle schools. A total of 9 classrooms in 3 schools provided data for 191 students. These students were seventh and eighth grade mathematics students, comprised of 94 males and 97 females. The assessment instrument was administered during a single regular class period without disruption to the general course of instruction. Each class scheduled the assessment administration to occur subsequent to the instructor having addressed the issues of probability and statistics in their classes, so the students would have had some instruction in the subject matter assessed.

Results

In general, students performed better than expected on the multiple-choice items, less so on the performance items. The international mean from the TIMSS administration for the ten items selected was 6.10 items correct. In this group of students, the mean number of total multiple choice items correct was 7.02. The total sample means and standard deviations for each item are presented in Table 1.

Table 1.
Means and Standard Deviations for Student Response Data

Assessment Item	Mean	SD	P-value.
MC1	.53	.50	.60
MC2	.49	.50	.48
MC3	.94	.23	.85
MC4	.85	.36	.73
MC5	.79	.41	.48
MC6	.56	.50	.51
MC7	.35	.48	.41
MC8	.95	.22	.79
MC9	.89	.32	.81
MC10	.58	.50	.44
MCTotal	7.02	1.88	
Food Spinner	1.77	.073	N/A
New Plan	2.02	.077	N/A

Note. P-values are not available for performance items.

Fitting Latent Class Models to the Data

The student performance data were submitted to a series of latent class models using a comprehensive modeling program developed by Muthen & Muthen (1998). Six models were fitted to the data, three models using each of the items in the assessment instrument as binary indicators, and three models using the multiple choice and performance tasks as three continuous indicators. The three models using twelve binary indicators involved 1, 2 and 3 latent classes. For these binary models, the multiple choice items were scored as either right (1) or wrong (0), while the performance task scores were dichotomized at the mid-point of the score range. That is, these variables were recoded such that a score of 3 or 4 indicated sufficient performance (1) while a score of 1 or 2 indicated poor performance (0). Similarly, the three models using the multiple choice and performance tasks as continuous measures involved one each for 1, 2, and 3 latent classes. Each model yielded separate parameter estimates (such as item probabilities in the binary case, or mean scores in the continuous case) for each latent class. In addition, each model produced estimates of latent class proportions, likelihoods of assignment to each class for each respondent, and fit indices for the model. These results for each model are presented in Tables 2-7.

In reviewing the separate parameter estimates for each class in the 2-class binary model, we see that the estimated probabilities for each item are quite different between the classes (see Table 3). Although for the easier items (items 3, 8, & 9; p-values all above .82), both groups had comparable estimated probabilities of getting the item correct, there were large differences in the likelihoods for several of the other items, particularly the performance tasks. For instance, item 7 had just over a 10% probability of being answered correctly by members of Class 1 while for the Class 2 this estimated leaped to over 62%. Similar disparities are found for items 1, 2, 6, and 10. Thus, it appears performance on these items discriminated between the two classes the best. In addition, the performance tasks generated very low probabilities for one group of students (just over 6% and 16% for the New Plan and Food Spinner task, respectively) and much higher likelihoods in the other class (25.4% and 38.4%, respectively).

Table 2.
Parameter Estimates for 1 Class Model with 12 Binary Indicators

Items	Class 1
MC Item 1	.531
MC Item 2	.495
MC Item 3	.943
MC Item 4	.845
MC Item 5	.794
MC Item 6	.562
MC Item 7	.345
MC Item 8	.948
MC Item 9	.887
MC Item 10	.577
Food Spinner	.149
New Plan	.263
Class Proportions	1.00
N	181

Table 3.
Parameter Estimates for 2 Class Model with 12 Binary Indicators

Items	Class 1	Class 2
MC Item 1	.346	.747
MC Item 2	.350	.665
MC Item 3	.936	.951
MC Item 4	.768	.936
MC Item 5	.670	.940
MC Item 6	.291	.879
MC Item 7	.106	.626
MC Item 8	.944	.954
MC Item 9	.827	.957
MC Item 10	.416	.766
Food Spinner	.061	.254
New Plan	.160	.384
Class Proportions	.540	.460

Table 4.
Parameter Estimates for 3 Class Model with 12 Binary Indicators

Items	Class 1	Class 2	Class 3
MC Item 1	.113	.771	.269
MC Item 2	.000	.646	.411
MC Item 3	.871	.952	.946
MC Item 4	.097	.920	.885
MC Item 5	.676	.935	.659
MC Item 6	.118	.848	.323
MC Item 7	.233	.622	.063
MC Item 8	.500	.957	.987
MC Item 9	.720	.956	.838
MC Item 10	.432	.741	.423
Food Spinner	.117	.251	.044
New Plan	.294	.380	.131
Class Proportions	.0719	.4830	.4451

The parameter estimates for the continuous indicator model were equally interesting (see Table 6). Whereas the mean multiple choice score for the lower group was only a little more than six and a half correct out of a possible ten (just above 6.6), the higher performing group achieved nearly two full points higher, at nearly eight and a half correct. Similarly, for the performance tasks, the higher performing groups achieved a mean score of about one full score point above the lower performing group (1.5 vs. 2.6 for the New Plan task; 1.8 vs. 2.6 for the Food Spinner task)

Table 5.
Parameter Estimates for 1 Class Model with 3 Continuous Indicators

Items	Class 1
Multiple Choice	7.033
Food Spinner	1.768
New Plan	2.017
Class Proportions	1.00

Table 6.
Parameter Estimates for 2 Class Model with 3 Continuous Indicators

Items	Class 1	Class 2
Multiple Choice	6.617	8.472
Food Spinner	1.517	2.637
New Plan	1.835	2.645
Class Proportions	.776	.224

Table 7.
Parameter Estimates for 3 Class Model with 3 Continuous Indicators

Items	Class 1	Class 2	Class 3
Multiple Choice	4.433	6.854	8.547
Food Spinner	1.393	1.541	2.665
New Plan	1.679	1.858	2.656
Class Proportions	.075	.712	.213

Table 8.
Latent Class Model Comparisons

Model	1-Class	2-Class	3-Class
12 Binary Indicators			
Loglikelihood	-1184.81	-1125.60	-1117.63
Free parameters	12	25	36
AIC	2393.62	2301.19	2307.25
BIC	2432.84	2382.89	2424.90
3 Continuous Indicators			
Loglikelihood	-779.14	-754.71	-753.64
Free parameters	6	10	14
AIC	1570.27	1529.42	1535.28
BIC	1589.46	1561.40	1580.06

In determining model fit for latent class models, several indices can be investigated (Sclove, 1987), including the loglikelihood value relative to the number of parameters, the Akaike information criterion (AIC; Akaike, 1987), and the Bayesian information criterion (BIC; Schwartz, 1978). Table 8 provides a summary of the fit indices for the six models. It appears that in both conditions, the 2-class models provide a better fit to the data than the 1- or 3- class models, as indicated by higher values of the loglikehood measure and lower values on the AIC and BIC measures.

Review of Class Proportions

The latent class proportions for all of the latent class models are presented in Table 9. It is interesting to note that for the two 2-class models using different input measures (binary items vs. continuous scores), the resultant latent class proportions are quite different. Though in both cases, the 2-class models were clearly the best fitting models of the three alternatives, they yielded quite different latent classes. In the case of the binary measures, the 2-class models identified one class constituting 54% of the respondents and another comprising 46%. For the model with the continuous indicators, the classes were much less equivalently distributed. The lower performing class contained 77.6 of the respondents compared to only 22.4% designated into the higher performing class. This suggests that the standards set with the latent class procedure using these different measures are not consistent – that using binary measures like items resulted in a lower standard while using the three continuous measures generated higher standards.

Table 9.
Latent Class Model Comparisons - Proportions in Each Latent Class

Model	Class 1	Class 2	Class 3
12 Binary Indicators			
1- Class Model	1.00		
2- Class Model	.540	.460	
3 - Class Model	.072	.483	.445
3 Continuous Indicators			
1- Class Model	1.00		
2- Class Model	.776	.224	
3 - Class Model	.075	.712	.213

Clustering of Designation Based on Response Profiles

It is interesting to look at how the two different latent class models of interest designated student responses by viewing a scatterplot of these designations as a function of their profiles. To present this in a two-dimensional space, the performance item scores were combined and represent the x-axis, while the multiple choice score represents the y-axis. Plotted on this plane are the class designations for the lower and higher performing groups. From this perspective, we can see, albeit in a crude way, the decision rules used by the latent class models for designating certain student performances into a specified group. It appears that for the binary model, a score of 7 or better on the collection of multiple choice items plus a combined score of 5 or better on the performance task assures a higher group designation, but such a designation can be obtained without necessarily doing well on the performance tasks. In many cases, a higher designation was achieved by very high scores on the multiple choice items and very low scores on the performance tasks. This finding is not surprising considering that each of the performance tasks were treated as just another item on the assessment in this model.

In contrast, the scatterplot for the continuous model clearly shows a much less compensatory approach. In this model, a higher level designation was almost never achieved without a minimum combined performance task score of 5. There was one exception for a profile that had 9 multiple choice items correct, a score of 1 on the New Plan task, and a score of 3 on the more difficult Food Spinner task. Clearly, and not surprisingly, the continuous model considers the respondents' performance on the performance tasks much more heavily in determining latent class designations.

Figure 1
Scatterplot of Class Assignment Using Binary Indicators

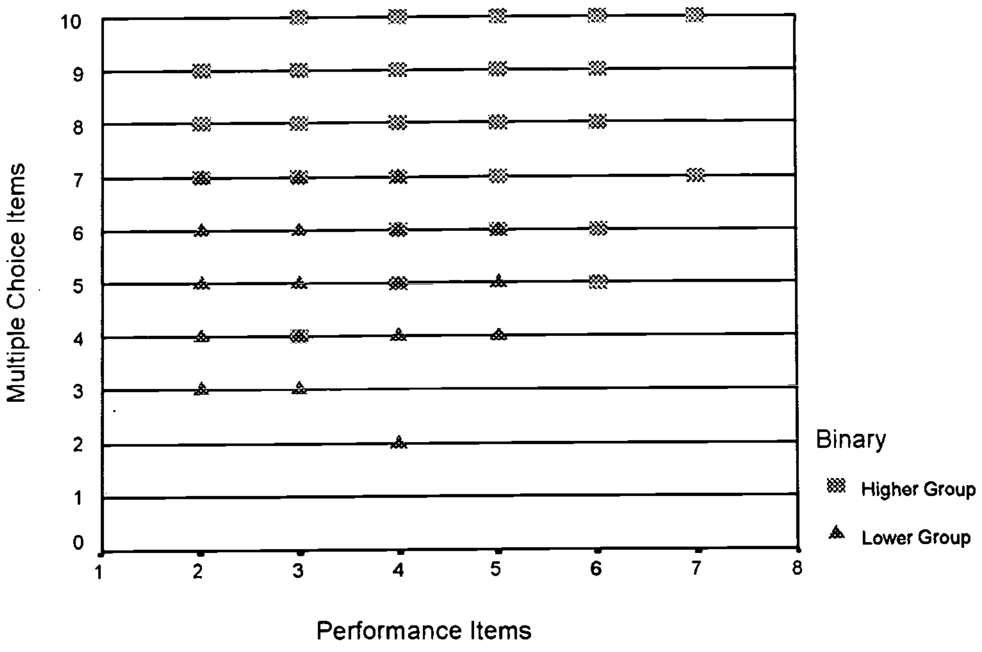
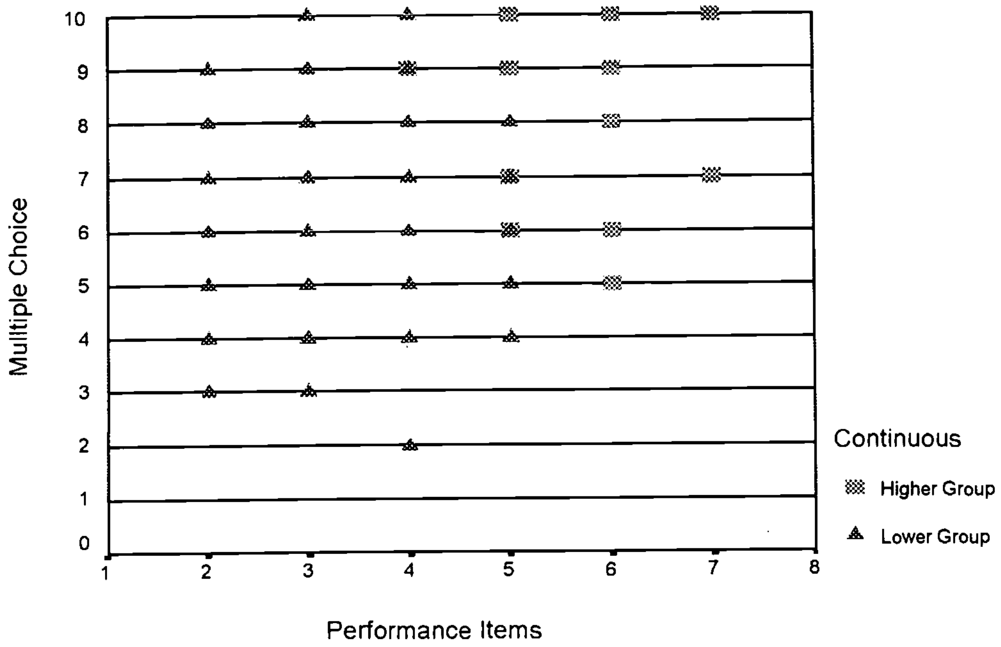


Figure 2
Scatterplot of Class Assignment Using Continuous Indicators



Summary of Latent Class Analysis

The latent class procedures provided information regarding the qualitative differences among the student respondent on the assessment instrument. A series of latent class models were explored using both binary and continuous indicators. For both types of indicators, a 2-class model fit the data better than a single class model, indicating that the respondents did differ qualitatively in their response patterns. In addition, it was seen that a 3-class model does not fit the data better than the 2-class model, suggesting that attempts at defining three distinct categories of student performance are not justified by these data. Rather, based on this analysis, student respondents could be classified into two distinct groups, each having different probabilities of getting each item correct (in the binary case) or different mean scores on the continuous measures. However, at what level of academic performance this demarcation is established is determined by how the data are presented. Using item level data resulted in a categorization of students wherein a high percentage of the students were deemed higher performers. In contrast, the model using the continuous measures yielded a breakdown wherein less than a quarter of the students were designated into the higher performing category. Clearly, using the scale scores placed more emphasis on the quality of the responses to the performance tasks and generated a more discriminating higher level group. How these designations coincide with determinations made using judgmental approaches will be explored next.

Comparisons Among Methods

In a related study (Brown, 1999), two judgmental approaches were applied to the student assessment instrument discussed here; a modified Angoff approach and a profile rating approach based on the judgmental policy capturing method of Jaeger (1995). The standards derived from these two methods, as applied to the sample of student responses from obtained in this study, can be compared with determinations made using the latent class procedure.

A series of 2 x 2 (doesn't meet/meets standard by standard setting method) cross tabulations were analyzed to investigate how well the various methods agreed with one other. These analyses generated some very interesting results.

Generally, the judgments from the different methods agreed with each other quite well, with the exception of the aforementioned differences between the latent class binary and continuous models. These two empirical approaches had the lowest level of exact agreement at 61.9%.

More interesting, however, was the concurrence between the Angoff and profile rating procedures. These two judgmental methods rendered identical determinations for 85.7% of the student responses. Such high levels of agreement between these two approaches is supportive evidence for the judgments of either method. That two distinct judgmental approaches provided such concurrence argues against the premise that the method makes a substantial difference. This replication of decisions across judgmental methods is even more compelling when one considers the internal consistency measures of each of the procedures. The agreement between the methods (85.7%) is on par with the level of agreement these raters had within the same method. Internal consistency measures for judgments for the Angoff approach was .91, while for the profile rating task it was .86.

The comparisons between empirical and judgmental approaches also generated support for the convergence of methods. Since judgments made using the Angoff rating procedure used item level information rather than scale values, it would be expected that the determinations made using this approach would concur more with designations made using the binary latent class procedure than with designations made using the continuous latent class procedure. Likewise, since the profile rating task relied on scale scores rather than item level data, the designations from this procedure should agree more with determinations made from the continuous latent class approach than with determinations made from the binary latent class method. Both of the expectations were supported by the data (see Table 10). The classifications of students based on the standards set by the Angoff and binary latent class procedure agreed a remarkable 92.2%, while agreement between the Angoff and continuous latent class approach lagged at 66.3%. Similarly, categorizations from the profile rating approach and the continuous latent class method agreed 87.2% of the time, while the profile rating method and binary latent class approach agreed only 77.1%. In total, these results clearly indicate the concurrence of agreement among the various standard setting methods using the same data elements regarding what constitutes proficient student achievement.

Table 10.
Percent of Agreement for Decisions Made Using Different Standard Setting Methods

Method	Angoff	Profile Rating	LCA-Binary
Angoff	--		
Profile Rating	**85.7**	--	
LCA-Binary	**92.2**	77.1	--
LCA-Continuous	66.3	**87.2**	61.9

Discussion

This study addresses an important issue in the area of setting performance standards around academic achievement, particularly when the assessment upon which the judgment is determined is comprised of both traditional multiple choice components and more recent performance assessment tasks. But before too much is made of the results from the current study, it is important to recognize the limitations and shortcomings which beset it.

First, as a single study, these results should be taken as one data point in the full data set of standard setting research. Though the results of this study may be compelling, they must be replicated in future samples of raters and student respondents before we can be sure the answered we think we've found are sound and stable. In addition, this study deals with only a single subject area (mathematics) for a single grade range (8th grade) with a single assessment instrument. Future research into other grade areas and subject matters should be undertaken and, it is hoped, support the findings demonstrated here. It would also be desirable to obtain similar results using other assessment instruments, especially those in wide scale use such as the standardized, norm referenced tests so heavily relied upon across the country for assessing academic achievement.

The latent class procedure indicates that for this sample of student responses, a two group structure was the most appropriate model for explaining the differences in student performance. This was consistent using both binary and continuous indicators. The implication of this finding is that there may not be more than two distinctly different groups of respondents, and thus attempts at identifying more than two groups would not be supported by the data. Further, this study showed that the results from the empirical procedure depend on the type of indicators presented to a greater extent than the judgmental approaches do. Agreement between the judgmental approaches using binary and continuous measures (85.7%) was much higher than the agreement between empirical procedures using different indicator types (61.9%).

Perhaps the most important finding from this study deals with the extent to which the varying standard setting procedures rendered comparable conclusions regarding what constitutes proficient performance. Comparisons across methods level revealed the approaches agree with each other at a very high level – a level equal to the internal consistency measures for the judgmental methods themselves. The implications of this finding are several. First, if the determinations made from two distinct judgmental approaches are comparable to decisions made using the same method on more than one occasion, it could be argued that the method doesn't really matter much. This finding, though consistent with recent meta-analytic work (Bontempo, Marks, & Karabatsos, 1998), runs counter to the prevailing belief in standard setting research. Berk (1996, p. 216) quoted a National Academy of Education report that stated, "The most consistent finding from the research literature on standard setting is that different methods lead to different results. Not only do judgmental and empirical methods lead to different results,. . . . but different judgmental methods lead to different results." Perhaps the results from this study and the recent quantitative research synthesis by Bontempo and colleagues will cause researchers in this area to reconsider this conclusion.

Another implication of this finding is that if empirical methods can be shown to consistently render determinations regarding student proficiency comparable to the more common, and more costly, judgmental approaches, these methods could be used more frequently to support the determinations made with those judgmental approaches. Having sound quantitative support to concur with human judgmental efforts would surely strengthen the basis for the established standards. The empirical approaches might even be used more in situations where convening human raters is too costly or otherwise prohibitive. In any event, these findings provide practitioners charged with the task of setting performance standards with another means of supporting or determining their decisions regarding where appropriate standards should be set for a given level of achievement. In addition, the judgmental approaches could be used interchangeably or to validate one another in a given context and for a given purpose.

This may assist practitioners and certification boards struggling with the question of which method to employ, or whether an empirical approach could be defensible. These results may also help educational policymakers in state education departments and local school districts in making their decisions about how to combine multiple student measures to make important, high stakes retention and graduation decisions. Many states and districts are currently facing tough decisions about just these issues.

References

Akaike, H. (1987). Factor analysis and AIC. Psychometrika, 52(3), 317-332.

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 36, 35-50.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike, (Ed.), Educational Measurement. Washington, DC: American Council in Education.

Baker, E. L., Freeman, F., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), Testing and cognition (pp. 131-153). Englewood Cliffs, NJ: Prentice Hall.

Baker, E. L., & Linn, R. L. (1997). Emerging educational standards of performance in the United States. CSE Technical Report 437.

Behuniak, P., Archambault, F. X., & Gable, R. K. (1982). Angoff and Nedelsky standard setting procedures: Implications for the validity of proficiency test score interpretation. Educational & Psychological Measurement, 42(1), 247-255.

Bergan, J. R. (1983). Latent-class models in educational research. In E. W. Gordon (Ed.), Review of research in education (Vol. 10, pp. 305-360). Washington, DC: American Educational Research Association.

Bergan, J. R., & Stone, C. A. (1985). Latent class models for knowledge domains. Psychological Bulletin, 98(1), 166-184.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56(1), 137-172.

Berk, R. A. (1996). Standard setting: The next generation (Where few psychometricians have gone before!). Applied Measurement in Education, 9(3), 215-235.

Bontempo, B. D., Marks, C. M., & Karabatsos, G. (1998). A meta-analytic assessment of empirical differences in standard setting procedures. Paper presented at the annual meeting of the American Educational Research Association. San Diego, CA.

Bowers, J. J., & Shindoll, R. R. (1989). Angoff, Beuk, and Hofstee: A comparison of multiple methods for setting a passing score. Paper presented at the annual meeting of the National Council for Measurement in Education. San Francisco, CA.

Brennan, R. D., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4, 219-240.

Brown, W. L. (1993). A study of the nature and extent of the discrepancies among three methods for setting passing scores. Paper presented at the annual meeting of the National Council for Measurement in Education. Atlanta, GA.

Brown, R. S. (1999). A comparison and validation of setting performance standards using judgmental and empirical approaches. Unpublished doctoral dissertation. University of California, Los Angeles.

Burstein, L., Koretz, D., Linn, R., Sugrue, B., Novak, J., Baker, E. L., & Harris, E. L. (1993). The validity of interpretations of the 1992 NAEP achievement levels in mathematics (August). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Cantor, J. A. (1989). A validation of Ebel's method for performance standard setting through its application with comparison approaches to a selected criterion-referenced test. Educational & Psychological Measurement, 49(3), 709-721.

Cascio, W. F., Alexander, R. A., & Barret, G. V. (1988). Setting cutoff scores: Legal, psychometric, and professional issues and guidelines. Personnel Psychology, 41, 1-24.

Cizek, G. J. (1996). Standard-setting guidelines. Educational Measurement: Issues & Practice, 15(1), 13-21.

Cross, L. H., Impara, J., C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. Journal of Educational Measurement, 21(2), 113-129.

Dawes, R. M., Faust, D., & Meehl, P. (1989). Clinical versus actuarial judgment. Science, 243, 1668-1674.

Dayton, C. M. (1991). Educational applications of latent class analysis. Measurement and Evaluation in Counseling and Development, 24(3), 131-141.

Dayton, C.M., & MacReady, G. B. (1976). A probabilistic model for validation of behavior hierarchies. Psychometrika, 41, 189-204.

Ebel, R. L. (1972). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice Hall.

Garrido, M. (1985). An experimental study of the effect of judges' knowledge of item performance data on two forms of the Angoff standard setting method. Dissertation Abstracts International. 1985 Nov. 46 5-A : p.1228-1229.

Gilovich, T. (1991). How we know what isn't so. New York: Free Press.

Haertel, E. (1984). An application of latent class models to assessment data. Applied Psychological Measurement, 8(3), 333-346.

Haertel, E. (1989). Using restricted latent class models to map the skill structure of achievement items. Journal of Educational Measurement, 26(4), 301-321.

Hamberlin, M. K. (1985). Influence of item response theory and type of judge on a standard set using the iterative Angoff standard setting method. Dissertation Abstracts International. 1993 Feb. 53 8-A : p.2781

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. Applied Measurement in Education, 8(1), 41-55.

Hurtz, G. M., Sanders, L. M., & Hertz, N. R. (1996). An investigation of the optimal number of raters for setting pass points with the Angoff method. Paper presented at the annual meeting of the Society of Industrial and Organizational Psychologists. San Diego, CA.

Impara, J. C., & Plake, B. S. (1995). Editor's note. Applied Measurement in Education, 8(1), 1-2.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. Journal of Educational Measurement, 34(4), 353-366.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 4, 461-476.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp., 485-514). New York: Macmillan.

Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. Applied Measurement in Education, 8(1), 15-40.

Kahneman, D., Slovic, P., & Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press.

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 17(3), 167-178.

Linn, R. L., Koretz, D., & Baker, E. L. (1996). Assessing the validity of the National Assessment of Educational Progress: NAEP Technical Review Panel white paper (CSE Tech. Rep. No. 416). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Linn, R. L., Koretz, D. M., Baker, E. L., & Burstein, L. (1991). The validity and credibililty of the achievement levels for the 1990 National Assessment of Educational Progress in mathematics (CSE Tech. Rep. No. 330). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Livingston, S. A., & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Longford, N. T. (1996). Reconciling experts' differences in setting cut scores for pass-fail decisions. Journal of Educational & Behavioral Statistics, 21(3), 203-213.

Luecht, R. M., DeChamplain, A. (1998). Applications of latent class analysis to mastery decisions using complex performance assessments. Paper presented at the annual meeting of the American Educational Research Association. San Diego, CA. April, 1998.

Maurer, T. J., Alexander, R. A., Callahan, C. M., Bailey, J. J., & Dambrot, F. H. (1991). Methodological and psychometric issues in setting cutoff scores using the Angoff method. Personnel Psychology, 44, 235-262.

Melican, G. J. & Plake, B. S. (1985). Correction for guessing and Nedelsky's standard-setting method: Are they compatible? Journal of Psychoeducational Assessment, 3(1), 31-36.

Mills, C. N. (1983). A comparison of three methods of establishing cut-off scores on criterion-referenced tests. Journal of Educational Measurement, 20(3), 283-292.

Mills, C. N. (1995). Comments on methods of setting standards for complex performance tests. Applied Measurement in Education, 8(1), 93-97.

Muthen, L. K., & Muthen, B.O. (1998). Mplus: A comprehensive modeling program for applied researchers. Los Angeles: Muthen & Muthen.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Niemi, D. (1996). Assessing conceptual understanding in mathematics: Representations, problem solutions, justifications, and explanations. Journal of Educational Research, 89, 351-363.

Nisbett, R. E. & Ross, L. (1980). Human inference: Strategies and shortcomings of social judgment. Englewood Cliffs, NJ: Prentice Hall.

Norcini, J. (1987). The answer key as a source of error in examinations of professionals. Journal of Educational Measurement, 24(4), 321-331.

Norcini, J. J., Lipner, R. S., Langdon, L. O., & Strecker, C. A. (1987). A comparison of three variations on a standard-setting method. Journal of Educational Measurement, 24(1), 56-64.

Plake, B. M. (1995). An integration and reprise: What we think we have learned. Applied Measurement in Education, 8(1), 85-92.

Plake, B. S. & Melican, G. J. (1989). Effects of item context on intrajudge consistency of expert judgments via the Nedelsky Standard Setting Method. Educational & Psychological Measurement, 49(1), 45-51.

Plake, B. S., Impara, J. C., & Potenza, M. T. (1994). Content specificity of expert judgments in a standard-setting study. Journal of Educational Measurement, 31(4), 339-347.

Plake, B. S.; Hambleton, R. K.; Jaeger, R. M. (1997). A new standard-setting method for performance assessments: The dominant profile judgment method and some field-test results. Educational & Psychological Measurement, 57(3), 400-411.

Putnam, S. E., Pence, P. & Jaeger, R. M. (1995). A multi-stage dominant profile method for setting standards on complex performance assessments. Applied Measurement in Education, 8(1), 57-83.

Reilly, R. R., Zink, D. L., & Israelski, E. W. (1984). Comparison of direct and indirect methods for setting minimum passing scores. Applied Psychological Measurement, 8(4), 421-429.

Rock, D. A., Davis, E. L., & Wertz, C. (1980). An empirical comparison of judgmental approaches to standard setting procedures. Research Report, Educational Testing Service.

Schwartz, G. (1978). Estimating the dimensions of a model. The Annals of Statistics, 6(2), 461-464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. Psychometrika, 52(3), 333-343.

Shepard, L. (1980). Standard setting issues and methods. Applied Psychological Measurement, 4(4), 447-467.

Sireci, S. G., & Biskin, B. H. (1992). Measurement practices in national licensing examination programs: A survey. Clear Exam Review, 21-25.

Skakun, E. N., & Kling, S. (1980). Comparability of methods for setting standards. Journal of Educational Measurement, 17, 229-235.

Smith, R. L. (1987). The item characteristics judges use when setting standards. Dissertation Abstracts International, 47, 8-A, 2966.

Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. Journal of Educational Measurement, 25(4), 259-274.

Sugrue, B, Novak, J., Burstein, L., Lewis, E., Koretz, D., & Linn, R. (1995). Mapping test items to the 1992 NAEP mathematics achievement level descriptions: Mathematics educators' interpretations and their relationship to student performance (CSE Tech. Rep. No. 393). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Taylor, C. S. (1987). The relationship between test length, standard setting method and decision consistency reliability. Dissertation Abstracts International,48 4-B : p.1183.

Van der Linden, Wim J. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. Journal of Educational Measurement, 19(4), 295-308.

Appendix

Student Assessment Instrument

This math test has three parts. Part A is a multiple-choice test. The multiple-choice test consists of ten questions.

In Part B and Part C you will read tasks that will require you to solve problems and write explanations.

You may underline numbers or words in the booklet, write notes, or draw pictures. You are to record all of your answers in this test booklet. It is important that you write clearly so that the person who reads your answers will understand what you meant to say.

Multiple-Choice Test

Directions to the Student

This is a test of how well you understand certain topics in math.

- At least four possible answers are given for each question. You are to choose the answer you think is better than the others.
- You may use the area to the right of each question as work space to figure out the answer to the problem.
- When you have decided which of the four answers is the correct one, circle that letter on the test paper. You may circle only one letter for each question.

Example:

Sally ate 1/5 of a pie. **What is the portion of pie she ate expressed as a decimal?**

A. 0.50

B. 0.25

[✓ C.] 0.20

D. 1.50

Since 1/5= 0.20, "C" is the correct answer. You should circle the letter "C".

You will have 10 minutes to complete the multiple-choice test. If you finish early you may check your work. Do not go on to the other parts of this test until you are told to do so.

1. In a bag of cards 1/6 are green, 1/12 are yellow, ½ are white and ¼ are blue. **If someone takes a card from the bag without looking, which color is it most likely to be?**

   A. White

   B. Blue

   C. Green

   D. Yellow

2. A drawer contains 28 pens; some white, some blue, some red, and some gray. **If the probability of selecting a blue pen is 2/7 how many blue pens are in the drawer?**

   A. 4

   B. 6

   C. 8

   D. 10

3. This chart shows temperature readings made at different times on four days.

TEMPERATURES					
	6 AM	9 AM	Noon	3 PM	8 PM
Monday	15°	17°	20°	21°	19°
Tuesday	15°	15°	15°	10°	9°
Wednesday	8°	10°	14°	13°	15°
Thursday	8°	11°	14°	17°	20°

When was the highest temperature recorded?

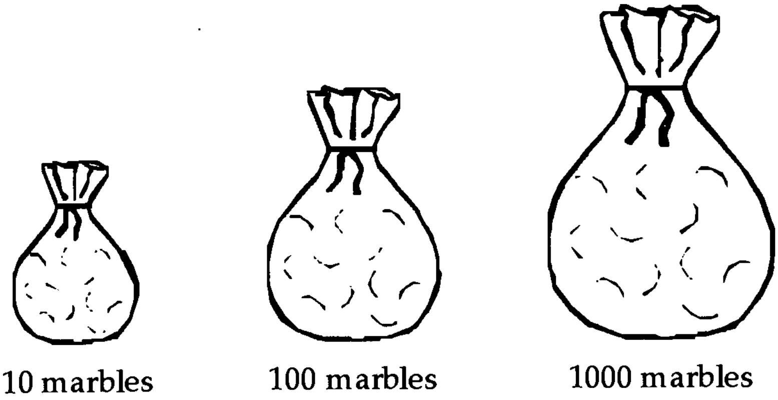
A. Noon on Monday

B. 3 PM on Monday

C. Noon on Tuesday

D. 3 PM on Wednesday

4. There is only one red marble in each of three bags. One bag has 10 marbles, one bag has 100 marbles, and one bag has 1000 marbles. Without looking in the bags, you are to pick a marble out of one of the bags. **Which bag would give you the greatest chance of picking the red marble?**



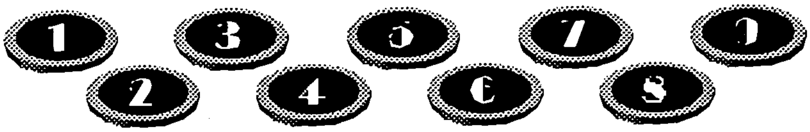
A. The bag with 10 marbles

B. The bag with 100 marbles

C. The bag with 1000 marbles

D. All bags would give the same chance

5. The nine chips shown are placed in a jar and mixed.



Madeline draws one chip from the jar. **What is the probability that Madeline draws a chip with an even number?**

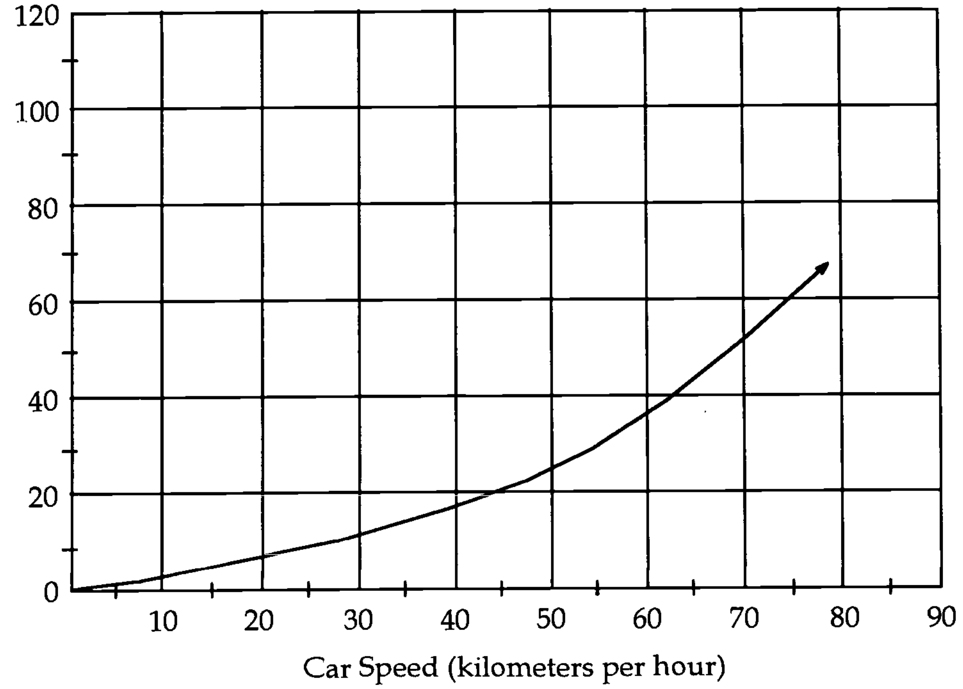
A. 1/9

B. 2/9

C. 4/9

D. 1/2

6. The graph shows the distance traveled before coming to a stop after the brakes are applied for a typical car traveling at different speeds.



A car traveling on a highway stopped 30m after the brakes were applied. About how fast was the car traveling?

A. 48 km per hour

B. 55 km per hour

C. 70 km per hour

D. 160 km per hour

7. Each of the six faces of a certain cube is painted either red or blue. When the cube is tossed, the probability of the cube landing with a red face up is 2/3. **How many faces are red?**

A. One

B. Two

C. Three

D. Four

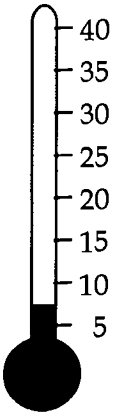
E. Five

8. This table shows temperatures at various times during the week.

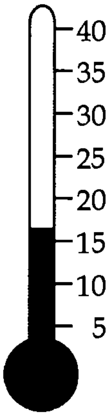
TEMPERATURES					
	6 AM	9 AM	Noon	3 PM	8 PM
Monday	15°	17°	20°	21°	19°
Tuesday	15°	15°	15°	10°	9°
Wednesday	8°	10°	14°	13°	15°
Thursday	8°	11°	14°	17°	20°

Which thermometer shows the temperature at 8 PM on Monday?

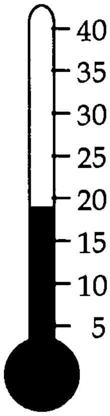
A.



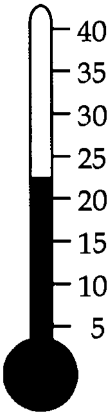
B.



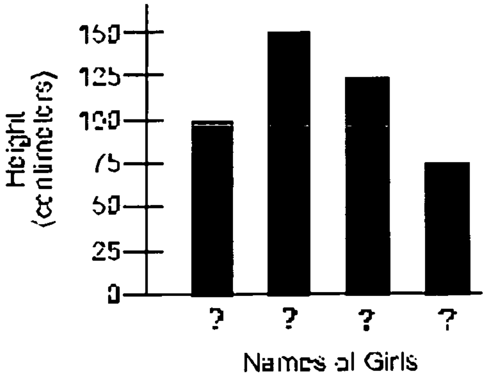
C.



D.



9. The graph shows the heights of four girls.



The names are missing from the graph. Debbie is the tallest. Amy is the shortest. Dawn is taller that Sarah. **How tall is Sarah?**

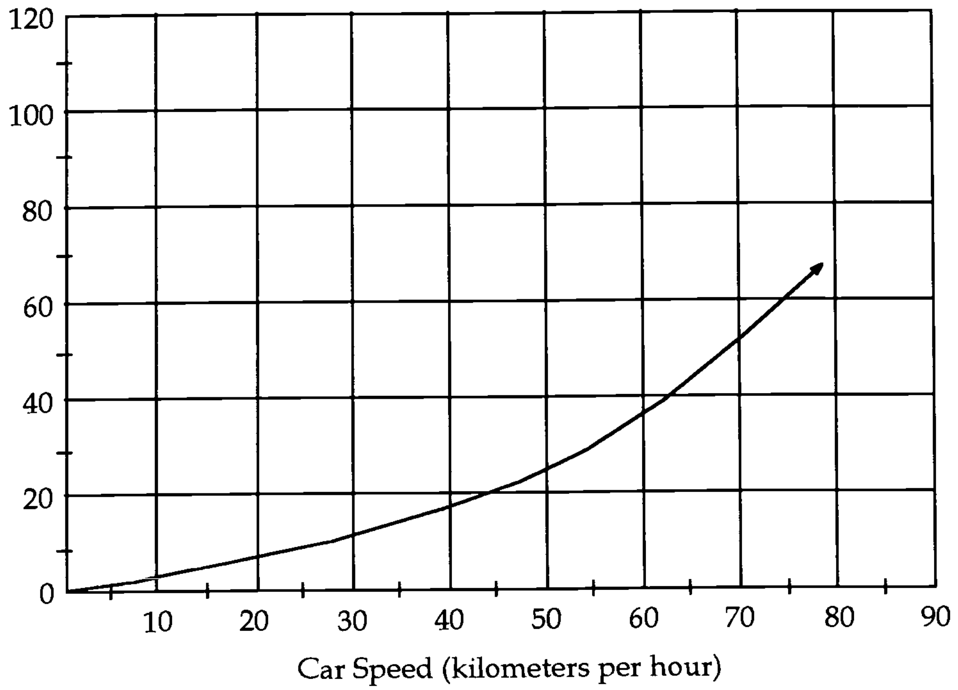
A. 75 cm

B. 100 cm

C. 125 cm

D. 150 cm

10. The graph shows the distance traveled before coming to a stop after the brakes are applied for a typical car traveling at different speeds.



A car is traveling 80 km per hour. **About how far will the car travel after the brakes are applied?**

A. 60 m

B. 70 m

C. 85 m

D. 100m

Time

Parts B & C: Performance Tests

Directions to the Student

In this section of the test you will solve problems and write explanations.

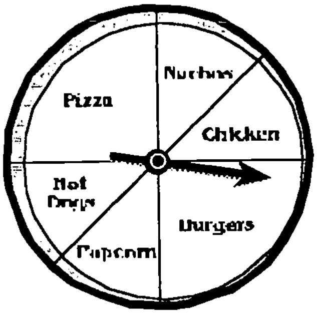
- Read each performance test completely and carefully before you start writing.
- Don't erase. Show all of your work in the space provided. If you make a mistake, draw one line through it.
- Write as complete a response as possible—make your response clear.
- If you are unsure of a response, write down what you do know to show your thinking.

You will have 20 minutes to complete each part. When you have finished Part B, you should go on to Part C of the test. If you finish early you may check your work.

Do **not** turn the page until you are told to do so.

Grade 7
Explanation Task

Food Spinner



The students in Ms. Castillo's classroom made this spinner to do a probability experiment. There are six possible food selections on this spinner.

The table below shows the results after 20 spins.

Frequency Table

Selection	Frequency of Outcome	Percentage
Pizza	3	15%
Hot Dogs	3	15%
Popcorn	4	20%
Burgers	7	35%
Chicken	1	5%
Nachos	2	10%
Total	20	100%

Use your knowledge of probability and statistics to write an explanation for the following questions. Explain your answers as clearly as possible. You may include as many tables and/or examples as you need. **Be sure to answer all of the questions.**

1. What is the probability of the pointer landing on each of the food selections shown in the spinner above? Explain your answer .
2. Are the numbers shown in the table different from what you expect? Explain why or why not.
3. What food selection would you expect on the 21st spin? Explain your answer .
4. What results would you expect if the students in Ms. Castillo's classroom had an

Grade 7
Problem-Solving Task

A New Plan

Pete and Marina ar e twins who always ar gue about what show to watch on T.V. in the evening after their homework is done. T o avoid any more ar guments, their mother came up with a plan. Each evening she will toss a coin. If the coin lands on heads Pete gets to choose the T .V. show. If the coin lands on tails Marina gets to choose.

Although their mother's plan is fair, Pete thinks it is too simple. He wants to design a plan that involves tossing thr ee coins at the same time and using the r esults from all thr ee coins as an outcome. Pete still wants the plan to be fair. Pete wants Marina and himself to have the same chance of winning.

You need to help Pete design the plan. Remember that you have to design a plan that involves tossing thr ee coins at the same time. The plan must use the r esults from all three coins as an outcome. **Be sure to show all of your work.**

In addition, in your solution be sur e to:

1. Identify your design for a new plan.
2. Explain how you know for sur e that your new plan is fair.
3. Describe the steps you took to design a new plan.

AERA

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM031238

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Using Latent Class Analysis to Set Academic Performance Standards

Author(s): Richard S. Brown

Corporate Source: National Center for Research on Evaluation, Standards and Student Testing/ UCLA

Publication Date: May 1, 2000

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature:

Printed Name/Position/Title: Senior Research Associate

Organization/Address:
300 Charles E. Young Dr. North
Los Angeles, CA 90095

Telephone: 310-794- 161

FAX: 310-825-1522

E-Mail Address: rbrown@ucla.edu

Date: May 1, 2000

(over)

ERIC® Clearinghouse on Assessment and Evaluation

University of Maryland
1129 Shriver Laboratory
College Park, MD 20742-5701

Tel: (800) 464-3742
(301) 405-7449
FAX: (301) 405-8134
ericae@ericae.net
http://ericae.net

March 2000

Dear AERA Presenter,

Congratulations on being a presenter at AERA. The ERIC Clearinghouse on Assessment and Evaluation would like you to contribute to ERIC by providing us with a written copy of your presentation. Submitting your paper to ERIC ensures a wider audience by making it available to members of the education community who could not attend your session or this year's conference.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed, electronic, and internet versions of *RIE*. The paper will be available **full-text, on demand through the ERIC Document Reproduction Service** and through the microfiche collections housed at libraries around the world.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse and you will be notified if your paper meets ERIC's criteria. Documents are reviewed for contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our processing of your paper at **http://ericae.net.**

To disseminate your work through ERIC, you need to sign the reproduction release form on the back of this letter and include it with **two** copies of your paper. You can drop of the copies of your paper and reproduction release form at the ERIC booth (223) or mail to our attention at the address below. **If you have not submitted your 1999 Conference paper please send today or drop it off at the booth with a Reproduction Release Form.** Please feel free to copy the form for future or additional submissions.

Mail to:

AERA 2000/ERIC Acquisitions
The University of Maryland
1129 Shriver Lab
College Park, MD 20742

Sincerely,

Lawrence M. Rudner

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

ERIC/AE is a project of the Department of Measurement, Statistics and Evaluation at the College of Education, University of Maryland.