DOCUMENT RESUME

ED 422 351 TM 028 903

AUTHOR	Hanson, Bradley A.; Bay, Luz; Loomis, Susan Cooper
TITLE	Booklet Classification Study.
INSTITUTION	ACT, Inc., Iowa City, IA.
SPONS AGENCY	National Assessment Governing Board, Washington, DC.
PUB DATE	1998-04-00
NOTE	39p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Diego, CA, April 14-16, 1998).
CONTRACT	ZA97001001
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Academic Achievement; *Classification; *Cutting Scores; Discriminant Analysis; Elementary Education; Geography; Geography Instruction; *Nonparametric Statistics; Science Instruction; Sciences; Tables (Data); Test Format; *Test Results; United States History
IDENTIFIERS	ACT Assessment; *National Assessment of Educational Progress

ABSTRACT

Research studies using booklet classification were implemented by the American College Testing Program to investigate the linkage between the National Assessment of Educational Progress (NAEP) Achievement Levels Descriptions and the cutpoints set to represent student performance with respect to the achievement levels. This paper describes the process and reports the results of the booklet classification study (BCS) implemented for the science achievement levels. It explores the possibility of using booklet classification as a way to set achievement levels by investigating methodologies for computing achievement level cutpoints using booklet classification data. These methodologies were applied to BCS data for science in this study and had been applied to geography and U.S. history. The BCS for science achievement levels involved grades 4 and 8, with 13 panelists for each grade level. Eighteen booklets were selected from NAEP forms, and 22 from other sources. The BCS for science, geography, and U.S. history have all resulted in panelists' classifying student performance at a lower level than plausible values scores indicate. These results indicate that cutpoints computed from booklet classification data would be higher than cutpoints based on the item-by-item rating methods that were used operationally. Procedures using the proportional odds model and nonparametric discriminant analysis were developed as a way to compute Achievement Level cutpoints using booklet classification data. Further refinements to these procedures, especially the nonparametric discriminant analysis, are needed before they could be used operationally to set cutpoints. (Contains 15 tables, 3 figures, and 13 references.) (SLD)

Booklet Classification Study[1]

Bradley A. Hanson
Luz Bay
Susan Cooper Loomis

ACT, Inc.

[1]Paper presented in M.L. Bourque (Moderator), *Setting Standards for the NAEP*, Related-Papers Session at the meeting of the National Council in Measurement in Education, April 14-16, 1998, San Diego, CA.

The research reported here was supported by contract #ZA97001001 from the National Assessment Governing Board.

Booklet Classification Study

Introduction

Two of the biggest criticisms of achievement levels set for the NAEP are: 1) they were set using analytic (i.e., item-by-item) methods, and 2) the achievement levels set were too high. A report by the National Academy of Education (1993) criticized item-by-item ratings as being too cognitively complex for panelists to provide ratings that will result in valid standards. Moreover, the proportion of students scoring at or above each achievement level, especially the Advanced level, was considered too small, implying the achievement levels were set too high.

Research studies using booklet classification were implemented by ACT to investigate the linkage between the NAEP Achievement Levels Descriptions (ALDs) and the cutpoints set to represent student performance with respect to the achievement levels. In 1995, a Booklet Classification Study (BCS) was implemented for each of the 1994 NAEP in Geography and U.S. History achievement levels. Panels composed of teachers, nonteachers, and members of the general public judge performance exhibited in student booklets. Classifications were compared to the empirical classifications of the booklets based on plausible values. Results of the two studies were reported in ACT (1995), Bay and Loomis (1995), and Kane and Bay (1996). The panelists generally judged the booklets in lower achievement level classifications than the empirical classifications. "These findings certainly do not suggest that the NAGB cutscores were set too high" (Kane and Bay, 1996, p. 22).

A Booklet Classification Study was also conducted for the 1996 NAEP Science Achievement Levels for grades 4 and 8. As in the booklet classification studies for geography and U.S. history, this study aimed to examine the extent to which students with scores in the intervals defined by the cutpoints demonstrated knowledge and skills corresponding to the ALDs. A complete description of the study is included in *Setting Achievement Levels on the 1996 National Assessment of Educational Progress in Science Final Report, Volume IV: Validation Studies.*

This paper describes the process reports the results of the BCS implemented for the science achievement levels. Secondly, it explores the possibility of using booklet classification as a way to set achievement levels by investigating methodologies for computing achievement level cutpoints using booklet classification data. The methodologies for computing achievement levels cutpoints are described and applied to the data from the Booklet Classification studies for Science, Geography, and U.S. History. Achievement level cutpoints computed using the data from the Booklet Classification Studies are compared to achievement level cutpoints obtained using item-by-item methods. Technical issues in computing cutpoints based on booklet classification data are discussed.

Method

The BCS for science was implemented for grades 4 and 8, but not for grade 12. Because of the concern regarding the unusually low cutscore for grade 4 Basic, coupled with the unusually small percentage of students above the grade 4 Advanced level, grade 4 was selected for the BCS. The State NAEP in Science was administered at grade 8. That increased the potential to identify booklets to represent all levels of achievement. Thus, the study was planned to include grades 4 and 8.

The Panelists. Thirteen panelists for each grade level participated in the study. There were eight teachers, three nonteacher educators, and two general public members at grade 4; and there were seven teachers, two nonteacher educators, and four general public members at grade 8. Five males and eight females on the grade 4 panel were mirrored by eight males and five females on the grade 8 panel.

The panelists for this study were selected on the same basis as the Achievement Levels-Setting (ALS) panelists. The pool of nominees remaining from the two pilot studies and the ALS process was used as the pool from which BCS panelists were selected.

The Booklets. The process of selecting the booklets for this study was in three stages: (1) selecting the blocks; (2) selecting the forms containing those blocks; and (3) selecting the booklets.

Two forms of the NAEP were used for each of the grade levels. For grade 4, these two forms contain four distinct blocks: one hands-on, one theme-based, and two concept/problem-solving blocks. Each of the forms contains each type of block. The two forms for grade 8 contain five distinct blocks: one hands-on, one theme-based, and three concept/problem-solving blocks. The blocks were fairly representative of the grade level item pool in terms of the percentages of items in each subscale and each item type, and the average difficulty of the items.

ETS provided a data file containing five plausible values for each booklet copy of the forms selected for the BCS. The *composite* plausible values[2] were used for classifying the booklets into one of the four levels of achievement. Except at the Advanced level[3], all booklets used in the study had all five plausible values within the range of the achievement level cutscores. The forty booklets used were distributed across the levels such that seven booklets were at the Below Basic level and thirteen at the Basic level. Two booklets were at the Advanced level for

[2]Five plausible values are randomly drawn for each student for each of the three science assessment subscales. The weighted average of the corresponding plausible values is the composite plausible value; e.g., weighted average of the first plausible values is the first composite plausible value. The weights of the subscales are specified in the assessment framework.

[3]Booklets classified at the Advanced level had three of the five plausible values within the range and the average of the five plausible values was within the range of a particular achievement level.

grade 8 and one for grade 4; and the remaining booklets were at the Proficient level.[4]

For each of the grades, 22 booklets were from one form, and 18 from another. Booklets selected for each achievement level were about evenly distributed across forms. For grade 8, only three of the 40 booklets were from the national assessment and the remainder were from the state assessment sample.

Training. To the extent practicable, panelists were provided the same orientation and training provided to the science ALS panelists, including taking the NAEP exam. For the most part, the BCS agenda paralleled the ALS agenda up to the first round of ratings. BCS panelists did not modify the ALDs or write borderline descriptions. All item-by-item exercises related to the internalization of the ALDs were eliminated from the BCS training to promote the holistic approach to the task. Panelists were given time to review the grade-level item pool, and they were instructed to review the scoring rubrics for constructed response items.

The panelists were trained in the Science NAEP framework, the NAGB policy definitions of the achievement levels, and the ALDs. Content resource staff worked with the panelists to help them understand the frameworks and to gain a confident understanding of the ALDs. They examined the alignment of statements included in the ALDs as a means of gaining a better understanding of the ALDs and the relationships across the levels. Panelists also participated in exercises that provided them the opportunity to apply their understanding of the ALDs as a means of training for the booklet classification tasks.

[4] The intended distribution was 7-13-13-7 for the Below Basic, Basic, Proficient, and Advanced levels, respectively. This distribution was used for the geography and U.S. history studies, but it could not be used here. Very few students scored at or above the Advanced level for any grade in science, and there were not enough booklets meeting the criteria—even the relaxed criteria—to select more at the Advanced level for these particular test forms.

Booklet Classification Task. The panelists were instructed to classify each booklet into one of the three achievement levels—or the Below Basic level—on the basis of the content framework, the policy definitions, and the ALDs. Forms were provided for panelists to record their classifications. The classification task was performed independently.

Panelists generally completed their classifications within the allotted time of approximately four hours. They were told that they could spend more time, if needed. It was suggested to them, however, that taking as much time as they desired would change the task they performed. They were urged to try to complete the classification of 40 booklets within the time allocated.

Correspondence Between Judgmental and Empirical Classifications

The "hit rate" of a panelist is the percentage of booklets that he/she classified the same as the empirical classification. The statistics on the hit rates relative to the plausible values classifications are presented in Table 1. The overall hit rates were 49% for grade 4 and 56% for grade 8. These hit rates were not very different from the results of the geography and U.S. history BCS.

One panelist in the grade 4 group classified 36 booklets as Below Basic and four as Basic. She was asked if she were certain about those classifications, and she was. The next day, however, she was certain that she had been unfair and too demanding in her classifications. During group discussions of the booklets, she seemed comfortable with suggesting the level at which she would then classify the booklets. If this panelist's ratings were deleted from the grade 4 group, the hit rate for grade 4 would increase to 52%.

To determine whether there were significant differences in the hit rates according to types of panelists, a *Kruskal-Wallis One-Way Analysis of Variance by Ranks* was performed. A significant difference was found in grade 4 but not in grade 8. (Please see Table 2.) Teachers had the lowest hit rates in grade 4, and the highest hit rates in grade 8.

Of the five grade 8 panelists who had the highest hit rates, four were females. Of the five grade 4 panelists who had the highest hit rates, four were males. No test was performed to determine the significance of these differences in hit rates by sex.

A 4x4 table of correspondence of judgmental classification and empirical classification based on plausible values was produced for each panelist. The within-cell percentages were averaged across panelists; these are presented in Tables 3 and 4. The numbers in parentheses are standard deviations. The quantity P_A is the proportion of matches in the judgmental and empirical classifications. Since the correspondence in classifications have been due to chance, the Kappa statistic (the proportion of agreement corrected for chance) was computed. The quantity P_E is the proportion of "chance" agreement; that is, the sum of the product of the corresponding marginal proportions. It is the expected value of the hit rates if the two classifications were done independently, keeping the distributions the same. The Kappa statistic is computed as $K = (P_A - P_E)/(1 - P_E)$, and it ranges from -1 to 1, with 1 indicating perfect agreement and zero indicating no agreement. Both Kappa values were significantly different from zero.[5]

As was the case for geography and U.S. history (ACT, 1995; Bay and Loomis, 1995; Kane and Bay, 1995), most panelists tended to classify the booklets at a level lower than the plausible values classifications. On average, grade 4 panelists classified 49% of the booklets at the empirical level indicated by the plausible values and 42% at one level lower. Grade 8 panelists classified about 56% of the booklets at the same level as the plausible values classification and about 36% at one level lower. (Please see Table 5.) Notice also that the judgmental classification was within one level of the empirical classification for an average of 93% of the booklets in grade

[5] Because the Kappa statistic is normally distributed for large sample sizes (Siegel and Castellan, 1988, p. 289) the z-statistic was used to test whether the K value is significantly higher than zero for each grade level.

4 and 99% of the booklets in grade 8.

Booklet Classification as a Method to Set Achievement Levels for the NAEP

Booklet classification studies for geography, U.S. history and science have all resulted in panelists classifying student performance at a lower level than the plausible value scores indicate. These results indicate that cutpoints computed from booklet classification data would have been higher that cutpoints based on the item-by-item rating methods that were used operationally. Booklet classification data from the three studies were used to explore methodologies to compute cutpoints, and to determine whether BCS panelists would have set higher cutpoints.

The reader is advised to interpret results of the computations with caution. The booklet classification studies were implemented without the intention of using the data to compute cutpoints. The computations presented here were performed *post hoc*.

Booklet Classification as a Method of Setting Achievement Levels

The Booklet Classification method involves panelists using the ALDs to classify completed NAEP booklets to achievement levels: Basic, Proficient, and Advanced, and the Below Basic level. The task for the panelists in the Booklet Classification method is to make a holistic judgement about a student's level of achievement based on the ALDs and a sample of the student's work as represented the responses of that student to the items in a NAEP booklet.

The judgements required by the panelists in the Booklet Classification method differs considerably from the judgements required by panelists in the modified Angoff method used in setting the NAEP Achievement Levels (ACT, 1997). The Booklet Classification method requires holistic judgements about actual student performance on a set of items. The modified Angoff method requires judgements about individual items for hypothetical students. Information about actual student performance on the items is not even needed for the modified Angoff method, although in the NAEP standard setting the panelists are given some information about overall performance on the items by a group of students and provided with samples of student responses.

It is possible that the differences in the types of judgements required by the Booklet Classification and modified Angoff methods could result in differences in the standards that are set. For instance, producing item-by-item judgements may lead the panelists to require a high level of performance on every item for students at the proficient and advanced levels which does not take into account the fact that even high performing students may not perform well on *every* item. The holistic judgements used in the Booklet Classification method could be more conducive to panelists producing more realistic standards that would allow a student to be at the Proficient or Advanced level without requiring exceptional performance on *every* item the student was administered.

Computing Achievement Levels Cutpoints

Booklet classifications provided by panelists were used to set Achievement Level cutpoints on the NAEP performance scale. This section discusses using panelists' judgements in the Booklet Classification method and plausible values for the booklets to set Achievement Level cutpoints. Approaches to modeling the data as a discrimination problem in order to set the Achievement Level cutpoints are presented next. The models will be presented under the assumption the NAEP scale score associated with a booklet is known. Following the presentation of the models the issue of the scale score associated with a booklet being unknown will be considered by describing the use of the plausible values associated with each booklet in setting the Achievement Level cutpoints.

Approaches to Modeling the Data

There are two approaches to modeling data in a classification problem: the sampling approach, and the diagnostic approach (Dawid, 1976; Titterington, Smith & Makov, 1985, pp. 168). Let $p_j(\theta, l)$ be the probability that judge j would classify a randomly selected booklet at level l (θ is the value of the scale score for this randomly selected booklet). The sampling approach uses the equality

$$p_j(\theta,\ l)\ =\ p(\theta | l,\ \phi_j)p(l | \pi_j) \tag{1}$$

where ϕ_j and π_j are parameter vectors. For the sampling approach the focus is on modeling the conditional distributions $p(\theta \mid l, \phi_j)$ and $p(l \mid \pi_j)$. Discriminant analysis fits into the sampling approach. In discriminant analysis estimates of $p(\theta \mid l, \phi_j)$ are used to determine the classification [the $p(l \mid \pi_j)$ are typically assumed known].

The diagnostic approach uses the equality

$$p_j(\theta,\ l)\ =\ p(l | \theta,\ \eta_j)p(\theta | \gamma) \tag{2}$$

where η_j and γ are parameter vectors. For the diagnostic approach the focus in on modeling the conditional distributions $p(l \mid \theta, \eta_j)$ and $p(\theta \mid \gamma)$.

The difference between the sampling and diagnostic approaches centers on the different conditional distributions that are modeled. That is, the conditional probability $p(\theta \mid l, \phi_j)$[6] for the sampling approach, and the conditional probability $p(l \mid \theta, \eta_j)$[7] for the diagnostic approach. Procedures developed for the two approaches from similar assumptions may not be equally

---

[6]The distribution of θ for booklets classified in a given level.

[7]The probability of a booklet being classified at a given level as a function of θ.

efficient or robust (Efron, 1975). Procedures will be considered based on both the sampling approach and the diagnostic approach. The next two sections discuss suggested procedures using the diagnostic and sampling approaches to compute cutpoints using BCS data.

Diagnostic Approach. The proportional odds model (Agresti, 1990, pp. 322; McCullagh & Nelder, 1989, pp. 153) can be used to model the conditional distributions $p(l \mid \theta, \eta_j)$. The proportional odds model takes into account that the categories are ordered. Define $c(l \mid \theta, \eta_j)$ as

$$c(l|\theta,\ \eta_j) = \sum_{l'=1}^{l} p(l'|\theta,\ \eta_j) \tag{3}$$

so $c(l \mid \theta, \eta_j)$ is the probability that rater j will classify a booklet at scale score θ in level l or less. The proportional odds model is

$$\log\left(\frac{c(l|\theta,\ \eta_j)}{1 - c(l|\theta,\ \eta_j)}\right) = \alpha_{jl} + \beta_j\theta \tag{4}$$

for $l = 1, 2, 3$, where $\eta_j = (\alpha_{j1}, \alpha_{j2}, \alpha_{j3}, \beta_j)$. The parameters in Equation 4 can be estimated for each rater. The $p(l \mid \theta, \eta_j)$ can be computed from the $c(l \mid \theta, \eta_j)$. The cutpoint between categories l and $l+1$, for $l = 1, 2, 3$, is the point were $p(l+1 \mid \theta, \eta_j)$ and $p(l \mid \theta, \eta_j)$ intersect. A single set of cutpoints over all raters can be obtained by averaging the cutpoints across raters. Another approach to producing a single set of cutpoints is to pool the data for all the raters and fit the proportional odds model to the pooled data.

Sampling Approach. The decision rules of interest consists of four intervals R_l, $l = 1, 2, 3, 4$, of the real line. The intervals are defined by three cutpoints ($t_1 < t_2 < t_3$) such that all points less than or equal to t_1 are in R_1 and are assigned to level 1, points greater than t_1 and less than or equal to

t_2 are in R_2 and are assigned to level 2, points greater than t_2 and less than or equal to t_3 are in R_3 and are assigned to level 3, and points greater than t_3 are in R_4 and are assigned to level 4. Let $C(l' \mid l)$ be the cost when the value of θ for a booklet is in interval $R_{l'}$ (the booklet would be classified by the decision rule to level l'), but the booklet is classified by the judge in level l. The Bayes rule in discriminant analysis (Anderson, 1958) minimizes the loss function

$$\sum_{l=1}^{4} p_j(l) \left[\sum_{\substack{l'=1 \\ l' \neq l}}^{4} C(l'|l) p(R_{l'}|, \phi_j) \right], \tag{5}$$

where $p_j(l)$ is the marginal probability of a booklet being rated in class l by judge j, and $p(R_{l'} \mid l, \phi_j)$ is given by

$$p(R_{l'}|,\ \phi_j) = \int_{R_{l'}} p(\theta | l, \phi_j) d\theta. \tag{6}$$

There are many methods of discriminant analysis that could be applied to produce cutpoints. In this paper, a nonparametric method similar to the procedure presented by Berk (1976) is used. For any set of cutpoints the values $p(R_{l'} \mid l, \phi_j)$, $l = 1, \ldots, 4$, are estimated by the proportion of the booklets classified by judge j at level l that are classified by the decision rule at level l' (in this case the parameter ϕ_j would just be an indicator for judge j). The value of $p_j(l)$ is estimated by the proportion of booklets classified at level l by judge j. Thus, for any decision rule the loss function can be computed.

There are a finite number of decision rules that will produce a unique value of the loss function. The number of possible decision rules is determined by the number of distinct θ values corresponding to the booklets used. The decision rule that minimizes the loss function is chosen.

If the number of decision rules is large, heuristic procedures could be used that minimize the loss over a smaller set of reasonable decision rules. For example, since the order of the cutpoints is known, one possible procedure would be to find a cutpoint between levels l and $l+1$, separately for $l = 1, 2, 3$. The problem for each pair of levels would be considered as a two level classification problem. For each pair of levels only data classified by judges at those two levels would be used in computing the cutpoint between the two levels. Finding cutpoints that minimize the loss for three separate two level classification procedures would be simpler than finding the cutpoints that minimized the loss for the four level problem where the number of possible decision rules could be very large.

As with the proportional odds model, the nonparametric discriminant analysis method can be applied to individual raters and average cutpoints calculated, or the method can be applied to a pooled data set of all raters to produce one overall set of cutpoints.

Use of Plausible Values

The statistical procedures for determining cutpoints presented above are functions of the unknown θ. The procedures as presented cannot be directly implemented since the θs are not directly observed. While the θs are not directly observed, information about the θs is available through the observed item responses of the examinees to the items in the booklets. In this situation Mislevy, Beaton, Kaplan, and Sheehan (1992) suggest in place of a statistic that is a function of individual θs, the expected value of the statistic over the conditional distribution of the θs for the booklets (the predictive distribution of the θs) be used. An approximation to this expected value can be computed using the five plausible values available for each booklet. This approximation is given by computing the cutpoints five times using the five plausible values for each booklet (the cutpoints are computed using the first plausible value for each booklet, computed again using the second plausible value for each booklet, etc.), and averaging the five results to obtain a final estimate.

Individual Panelist Versus Group Cutpoints

Using the procedures described above the cutpoints can be computed for an individual panelist using that panelist's classifications of the booklets to Achievement Levels and the plausible values for the booklets. Overall cutpoints across panelists can be obtained by computing the mean of the cutpoints over the individual panelists. Another possibility for computing a set of cutpoint across panelists is to pool the data for all panelists and compute cutpoints on the pooled data. Both of these procedures will be used in computing Achievement Level cutpoints in this paper.

Achievement Level Cutpoints

Achievement level cutpoints were calculated for Science at grades 4 and 8, and Geography and U.S. History in grades 4, 8, and 12. Cutpoints were computed using both the proportional odds model (diagnostic approach) and nonparametric discriminant analysis (sampling approach) described previously. Cutpoints were calculated for each panelist and mean cutpoints were computed across the panelists. Cutpoints were also computed using the pooled data for all panelists. Thus, for each grade and subject combination there were four sets of cutpoints computed (two models by individual/pooled).

The grade 8 science data is used to illustrate the computational procedures used to produce the cutpoints. In the booklet classification study for grade 8 science, 13 panelists rated 40 booklets (see Table 6). Due to the very limited number of classifications of booklets at the Advanced level by the panelists, the Proficient and Advanced levels are combined into a single level for the analyses reported. Thus, decision rules will be produced that classify booklets into three levels (with two cutpoints), rather than four levels (with three cutpoints).

The proportional odds model was fit to the data for each panelist for each of the five plausible values. Cutpoints were produced for each panelist and each set of plausible values based on estimates from the proportional odds model as described above. The cutpoint for a panelist is the average of the cutpoints computed using the five plausible values. The cutpoints for the 13 judges are presented in columns 2 and 3 of Table 7, along with the mean cutpoints over judges. In addition, the proportional odds model was estimated using data pooled over all panelists to produce five sets of cutpoints, one for each set of plausible values. The five sets of cutpoints were averaged to produce overall cutpoints. These cutpoints are presented in Table 7 (labeled "Pooled"). Figure 1 is a graph of conditional probabilities $p(l|\theta)$ obtained from the proportional odds model using the pooled data from all panelists where the scaled scores are the first plausible values. Figure 2 plots the pairs of cutpoints for all panelists, as well as the mean pair of cutpoints, and the pair of cutpoints produced by pooling the data for all panelists.

For the nonparametric discriminant analysis the focus is on the distribution of θ for booklets classified at each level by the panelists. The heuristic procedure described above was used to solve for the two cutpoints from two separate two-category classification problems. For each of the separate two-category classification problem the loss for all cutpoints were calculated. There were a finite number of cutpoints that produced a unique value for the loss.

The costs of misclassification ($C(l'|l)$ in Equation 5) were all set equal to 1. For each panelist five sets of cutpoints were produced corresponding to the five sets of plausible values. The cutpoint for each panelist was the average over the five values.

The cutpoints for the 13 panelists are given in the last two columns of Table 7, along with the mean of the cutpoints across panelists. The procedure was also applied to the pooled data for all panelists. The cutpoints based on the pooled data are presented in Table 7 (labeled "Pooled"). Figure 3 plots the pair of cutpoints for all panelists, as well as the mean pair of cutpoints, and the pair of cutpoints produce by pooling the data for all panelists.

Cutpoints were computed for the other grades and subjects using the same procedures as described above for grade 8 Science. Cutpoints for the other grades and subjects are given in Tables 8 through 14. As for grade 8 Science, the Proficient and Advanced levels were combined into a single level for grade 4 Science (Table 8). For Geography (Tables 9-11) and U.S. History (Tables 12-14) all four achievement levels were used, so for those subjects three cutpoints were calculated. Advanced cutpoints could not be computed in grade 12 Geography for three panelists due to these panelists not classifying any of the booklets at the Advanced level. (See Table 11.)

For a few panelists consecutive cutpoints are reversed. For example, for panelist 2 in Geography grade 4 (Table 9) the cutpoint for the Basic level is greater than the cutpoint for the Proficient level for nonparametric discriminant analysis, and for panelist 5 in U.S. History grade 12 (Table 14) the cutpoint for the Proficient level is greater than the cutpoint for the Advanced level for the proportional odds model. Reversals can occur for nonparametric discriminant analysis due to the fact that cutpoints are computed for the consecutive levels independently, so there is nothing to constrain the cutpoint to be in the proper order. For the proportional odds model it is not possible for a panelist's cutpoints to be reversed when computed using one of the plausible values. For both the proportional odds model and nonparametric discriminant analysis it is possible for cutpoints to be revered due to the fact that the cutpoints reported in the tables are the mean cutpoints computed using the five plausible values. Even if the cutpoints for each of the five plausible values are in the correct order, it is not necessarily the case the mean cutpoints will be correctly ordered.

There is considerable variation in the cutpoints among the panelists. The cutpoints for the individual panelists are more disparate for the nonparametric discriminant analysis than for the proportional odds model. In most cases the mean of the cutpoints across panelists is very near the cutpoint computed by pooling panelists. The mean and pooled cutpoints tend to be closer for the proportional odds model than they are for nonparametric discriminant analysis. One difference between the procedures that could partially account for differences in results is that the proportional odds model is a parametric model (the cutpoints are based on the intersection of estimated smooth curves), whereas in nonparametric discriminant analysis there is no parametric smoothing of the data.

Table 15 presents the cutpoints presented in Tables 7 through 11 as computed using the pooled data along with percentages of students at or above each Achievement level cutpoint. Percentages at or above the cutpoints are only presented for Science and Geography, they were not available for U.S. History. The results in Table 15 show that the cutpoints from the Booklet Classification Studies are generally higher, and in many cases much higher, than the cutpoints from the Achievement Levels Studies which used a modified Angoff procedure. Consequently, the percentages at or above the cutpoints tend to be lower for the Booklet Classification Studies. The booklet classification method has not resulted in a greater percentage of students being classified at or above each Achievement Level than the item-by-item method used in Achievement Level Studies.

Technical Issues in Computing Cutpoints from Booklet Classification Data

The procedures using the proportional odds model and nonparametric discriminant analysis were developed as a way to compute Achievement Level cutpoints using booklet classification data. Further refinements to the procedures, especially the nonparametric discriminant analysis, are needed before they should be used operationally to set cutpoints. Two areas for improvement in the procedures for calculating cutpoints from booklet classification data are presented below.

Hierarchical Models

A more appropriate formulation of the problem could be to use hierarchical models. For example, using a proportional odds model for each judge, the α parameters with associated hyper-parameters, would be distributed across judges according to some distribution. The β parameter might be assumed to be constant across judges. The hyper-parameters could be estimated and the mean of the resulting distribution of α parameters could be used to provide an overall cutpoint. A hierarchical approach would be more difficult with the nonparametric discriminant analysis approach suggested.

Standard Errors

In computing standard errors of the cutpoints, it is necessary to be clear about the sources of random errors to be incorporated. Judges and booklets are sampled, as are the θ values assigned to the booklets. If a procedure is specified for computing standard errors of the cutpoints for a fixed set of θs associated with the booklets, then the methods described in Mislevy, Beaton, Kaplan, and Sheehan (1992) could be used to incorporate uncertainty about the θs in the standard errors (using the plausible values).

References

ACT (1997). Achievement levels-setting study. In *Setting Achievement Levels on the 1996 National Assessment of Education Progress in Science Final Report.* Iowa City, IA: ACT.

ACT (1995). *Research studies on the achievement levels set for the 1994 NAEP in Geography and U.S. History.* (Unpublished).

Agresti, A. (1990). *Categorical Data Analysis.* New York: John Wiley & Sons.

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis.* New York: John Wiley & Sons.

Bay, L. & Loomis, S.C. (1995, June). *Validation of the 1994 NAEP achievement levels: Booklet classification study.* Paper presented at the National Conference on Large Scale Assessment, Phoenix, AZ.

Berk, R. A. (1976). Determination of optional cutting scores in criterion- referenced measurement. *Journal of Experimental Education, 15*, 4-9.

Dawid, A. P. (1976). Properties of diagnostic data distributions. *Biometrics, 32*, 647-658.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association, 70*, 892-898.

Kane, M. & Bay, L. (1996, April). The intersection of score interpretation, assessment design, and standard-setting methodology on NAEP. In R. L. Brennan (Moderator), *Psychometric Issues in Setting Standards on the NAEP*. Symposium conducted at the meeting of the National Council on Measurement in Education, New York.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2nd Ed.). London: Chapman and Hall.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133-161.

National Academy of Education (1993). *Setting performance standards for student achievement.* A report of the National Academy of Education Panel on the evaluation of the NAEP trial state assessment: An evaluation of the 1992 achievement levels. Stanford, CA: Stanford University, National Academy of Education.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions.* New York: John Wiley & Sons.

Table 1

Distribution Statistics of Science BCS Panelists' Hit Rates

Statistics	Grade 4 (n=13)	Grade 8 (n=13)
Minimum	18	26
Maximum	72	77
Median	44	56
Average	49	56
S.D.	15	16

Table 2

Average Rank of Panelists Based on Their Hit Rates (Lower Average Rank=Higher Hit Rate)

Grade	Panelist Type			H Statistics
	Teacher	Nonteacher	General Public	
4	4.88	9.67	11.50	6.57 (p=0.037)
8	8.71	5.00	5.00	2.97 (p=0.226)

Note: The value p is the probability of the observed rankings given that there are no true differences among average rankings by panelist type.

Table 3

Average Correspondence of Judgmental Classifications and Empirical Classifications Based on Plausible Values: Grade 4

Empirical	Judgmental				
	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	17.6 (0.9)	0.2 (0.7)	0.2 (0.7)	0.0 (0.0)	17.9
Basic	17.4 (10.2)	15.6 (10.3)	0.4 (0.9)	0.0 (0.0)	33.3
Proficient	5.7 (9.8)	23.1 (6.1)	15.8 (5.9)	1.6 (2.4)	46.2
Advanced	0.0 (0.0)	0.4 (0.9)	1.8 (1.2)	0.4 (0.9)	2.6
Total	40.6 (18.8)	39.3 (13.3)	18.1 (6.2)	2.0 (3.2)	P_A=.49 P_E=.29 K=.29

22

Table 4

Average Correspondence of Judgmental Classifications and Empirical Classifications Based on Plausible Values: Grade 8

Empirical	Judgmental				
	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	17.4 (1.1)	0.6 (1.1)	0.0 (0.0)	0.0 (0.0)	17.9
Basic	13.4 (8.7)	18.3 (8.3)	1.6 (2.8)	0.0 (0.0)	33.3
Proficient	1.0 (1.2)	19.5 (9.2)	18.3 (8.8)	4.7 (6.3)	43.6
Advanced	0.0 (0.0)	0.4 (0.9)	3.2 (2.0)	1.6 (2.1)	5.1
Total	31.8 (10.5)	38.9 (10.7)	23.1 (10.0)	6.3 (8.1)	P_A=.56 P_E=.29 K=.37

Table 5

Percentages of Judgmental Classifications Within One Level of Empirical Classifications Based on Plausible Values

Grade		E < J	E = J	E > J
4	Minimum	0	18	21
	Maximum	8	72	54
	Average	2	49	42
	S.D.	3	15	12
8	Minimum	0	26	8
	Maximum	33	77	62
	Average	7	56	36
	S.D.	9	15	16

Note:
E < J: Judgmental classification is higher than empirical classification.
E = J: Judgmental classification is the same as empirical classification.
E > J: Judgmental classification is lower than empirical classification.

Table 6

Numbers of Panelists and Booklets in Each Booklet Classifications Study

Study	Grade	Number of Panelists	Number of Booklets
1994 U.S. History	4	10	34
	8	10	39
	12	10	31
1994 Geography	4	10	37
	8	10	40
	12	10	40
1996 Science	4	13	40
	8	13	40

Table 7

BCS Cutpoints[1] Computed Using the Proportional Odds Model and Discriminant Analysis Science Grade 8

Panelist	Proportional Odds Model		Discriminant Analysis	
	Basic	Proficient	Basic	Proficient
1	158.24	184.05	158.50	182.32
2	153.07	172.61	149.82	172.49
3	168.29	178.23	170.49	180.11
4	164.95	183.19	168.53	187.57
5	161.05	173.46	161.53	172.00
6	166.39	180.06	167.67	183.85
7	150.07	163.80	146.61	161.80
8	148.42	179.00	149.82	174.84
9	153.58	178.05	153.32	175.13
10	151.72	175.82	148.98	173.53
11	152.25	173.45	154.46	171.46
12	166.24	178.37	168.16	178.05
13	158.64	180.32	157.40	180.77
Mean	157.92	176.96	158.10	176.45
Pooled	158.06	177.08	152.74	175.04

[1]Cutpoints are in the ACT NAEP-Like scale.

Table 8

BCS Cutpoints[2] Computed Using the Proportional Odds Model and Discriminant Analysis Science Grade 4

Panelist	Proportional Odds Model		Discriminant Analysis	
	Basic	Proficient	Basic	Proficient
1	161.552	177.413	165.658	178.385
2	159.678	175.233	159.645	175.670
3	143.843	175.964	143.784	173.037
4	158.836	177.025	159.085	179.102
5	157.883	182.858	159.891	182.348
6	165.038	180.902	163.212	180.844
7	147.565	176.418	151.762	175.628
8	155.471	173.325	157.472	171.396
9	165.732	177.586	165.579	177.617
10	146.490	183.019	141.239	183.990
11	162.924	177.529	164.111	179.427
12	143.919	172.405	141.239	169.234
Mean	155.744	177.473	156.056	177.223
Pooled	156.081	177.812	157.428	180.021

[2]Cutpoints are in the ACT NAEP-Like scale.

Table 9

BCS Cutpoints[3] Computed Using the Proportional Odds Model and Discriminant Analysis Geography Grade 4

Panelist	Proportional Odds Model			Discriminant Analysis		
	Basic	Proficient	Advanced	Basic	Proficient	Advanced
1	158.213	172.202	193.447	157.630	171.463	189.574
2	169.876	172.682	184.657	170.290	166.642	182.363
3	158.963	174.308	194.209	159.281	173.048	185.641
4	170.344	193.037	195.143	167.951	184.418	174.279
5	157.646	177.602	189.952	159.264	176.500	185.929
6	168.229	182.420	194.812	168.651	179.519	188.108
7	166.986	177.005	189.904	166.787	178.514	184.095
8	171.529	177.291	185.176	171.794	173.698	183.124
9	158.558	175.887	191.026	157.661	172.992	185.436
10	156.579	173.738	182.762	155.688	171.167	181.846
Mean	163.692	177.617	190.109	163.500	174.796	184.039
Pooled	163.804	177.357	190.360	165.188	175.171	185.830

[3] Cutpoints are in the ACT NAEP-Like scale.

Table 10

BCS Cutpoints[4] Computed Using the Proportional Odds Model and Discriminant Analysis Geography Grade 8

Panelist	Proportional Odds Model			Discriminant Analysis		
	Basic	Proficient	Advanced	Basic	Proficient	Advanced
1	143.276	163.810	181.104	142.564	161.037	179.839
2	149.077	170.994	189.925	145.761	170.869	185.133
3	155.645	171.348	181.298	156.615	167.025	176.396
4	142.758	172.215	177.739	142.495	167.931	174.620
5	143.231	165.798	180.619	142.564	163.017	177.529
6	145.981	165.745	174.511	146.742	166.018	175.429
7	153.381	167.026	177.842	152.325	161.261	176.854
8	145.235	173.321	179.849	144.603	171.357	178.261
9	149.555	163.705	177.574	146.932	162.373	175.454
10	149.174	167.195	178.646	146.620	162.435	176.279
Mean	147.731	168.116	179.911	146.722	165.332	177.579
Pooled	147.213	167.864	179.852	145.845	164.783	178.382

[4]Cutpoints are in the ACT NAEP-Like scale.

Table 11

BCS Cutpoints[5] Computed Using the Proportional Odds Model and Discriminant Analysis Geography Grade 12

Panelist	Proportional Odds Model			Discriminant Analysis		
	Basic	Proficient	Advanced	Basic	Proficient	Advanced
1	158.660	174.008	182.360	157.613	173.957	181.153
2	157.232	178.645	175.768	157.120	179.696	173.306
3	161.206	178.895		160.963	177.047	
4	167.300	192.745		166.670	184.232	
5	167.679	180.914	192.552	165.537	179.664	186.747
6	157.897	179.614	186.066	157.647	179.208	169.649
7	172.145	179.363	187.949	172.339	177.809	185.635
8	175.346	177.276		177.282	175.823	
9	148.027	175.062	171.316	147.396	175.967	170.293
10	152.949	174.657	180.147	151.095	174.949	179.927
Mean	161.844	179.118	182.308	161.366	177.835	178.101
Pooled	162.009	180.877	185.836	159.108	179.696	182.303

[5]Cutpoints are in the ACT NAEP-Like scale.

Table 12

BCS Cutpoints[6] Computed Using the Proportional Odds Model and Discriminant Analysis U.S. History Grade 4

Panelist	Proportional Odds Model			Discriminant Analysis		
	Basic	Proficient	Advanced	Basic	Proficient	Advanced
1	150.415	173.208	186.612	146.187	172.659	182.409
2	157.367	172.255	185.487	157.085	171.542	182.317
3	150.446	174.778	183.987	150.597	174.574	182.409
4	158.916	172.565	184.041	157.169	171.250	179.012
5	148.569	164.159	180.748	147.901	162.051	178.879
6	154.123	166.701	183.658	156.705	165.581	179.050
7	153.951	168.077	181.713	156.596	170.042	182.623
8	150.397	167.090	176.945	154.456	166.858	175.849
9	158.652	177.112	186.547	159.881	178.341	181.963
10	154.062	170.337	179.587	155.496	171.387	180.477
Mean	153.690	170.628	182.932	154.207	170.429	180.499
Pooled	153.661	170.796	182.922	154.739	171.793	181.632

[6]Cutpoints are in the ACT NAEP-Like scale.

Table 13

BCS Cutpoints[7] Computed Using the Proportional Odds Model and Discriminant Analysis U.S. History Grade 8

Panelist	Proportional Odds Model			Discriminant Analysis		
	Basic	Proficient	Advanced	Basic	Proficient	Advanced
1	168.116	182.389	192.551	169.892	184.069	188.840
2	162.637	176.414	189.383	164.094	175.640	186.146
3	171.944	180.530	184.646	173.641	178.501	182.261
4	168.568	175.295	188.464	171.269	167.901	183.293
5	163.776	182.323	186.182	164.416	180.644	180.577
6	163.898	177.315	198.100	159.989	173.074	189.531
7	163.723	183.210	188.471	163.659	187.108	180.350
8	166.973	181.637	184.561	169.285	181.749	180.795
9	160.086	186.460	199.205	161.000	188.980	192.283
10	165.705	178.007	183.275	165.960	174.352	181.287
Mean	165.543	180.358	189.484	166.321	179.202	184.536
Pooled	165.215	180.465	188.931	163.351	182.396	184.101

[7]Cutpoints are in the ACT NAEP-Like scale.

Table 14

BCS Cutpoints[8] Computed Using the Proportional Odds Model and Discriminant Analysis U.S. History Grade 12

Panelist	Proportional Odds Model			Discriminant Analysis		
	Basic	Proficient	Advanced	Basic	Proficient	Advanced
1	161.835	172.074	183.904	157.788	170.815	180.718
2	159.484	183.653	176.034	159.531	179.330	179.845
3	173.797	178.879	193.529	171.064	171.115	191.259
4	164.230	180.796	185.826	161.594	179.580	181.905
5	163.387	177.427	174.952	160.599	175.561	176.011
6	165.904	170.696	182.734	164.597	170.206	181.675
7	158.399	172.208	181.209	159.125	171.386	180.234
8	164.330	175.866	189.031	162.519	175.279	190.684
9	158.573	174.588	187.918	154.951	174.833	188.547
10	159.779	172.742	181.339	158.291	171.728	181.383
Mean	162.972	175.893	183.648	161.006	173.983	183.226
Pooled	162.749	175.319	183.906	161.594	172.336	180.972

[8] Cutpoints are in the ACT NAEP-Like scale.

Table 15
Comparisons of Cutpoints and Percentages of Students Scoring At or Above Each Achievement Level

Subject	Grade	Data Source		Basic		Proficient		Advanced	
				Cutpoint	%≥	Cutpoint	%≥	Cutpoint	%≥
Geography	4	BCS	PO	163.8	**28.9**	177.4	**3.1**	190.4	**0.1**
			DA	165.2	**25.1**	175.2	**5.2**	185.8	**0.3**
		ALS		137.4	**71.0**	152.2	**22.8**	162.3	**3.0**
	8	BCS	PO	147.2	**72.6**	167.9	**18.7**	179.9	**2.3**
			DA	145.8	**77.4**	164.8	**25.5**	178.4	**3.1**
		ALS		152.8	**71.8**	164.0	**27.8**	173.2	**4.0**
	12	BCS	PO	162.0	**31.9**	180.9	**1.3**	185.8	**0.3**
			DA	159.1	**42.6**	179.7	**2.0**	182.3	**1.0**
		ALS		160.6	**71.8**	170.4	**26.5**	180.0	**1.6**
	4	BCS	PO	156.1	**49.8**	177.8	**2.6**		
			DA	157.4	**45.9**	180.0	**1.4**		
		ALS		142.6	**82.9**	166.9	**0.1**		
	8	BCS	PO	158.1	**43.5**	177.1	**3.9**		
			DA	152.7	**59.0**	175.0	**5.9**		
		ALS		154.2	**55.5**	176.7	**4.2**		
		Reconvention		150.1	**66.3**	171.3	**10.5**		

Figure 1
Conditional Probabilities from Proportional Odds Model Using the First Set of Plausible Values (Science Grade 8)

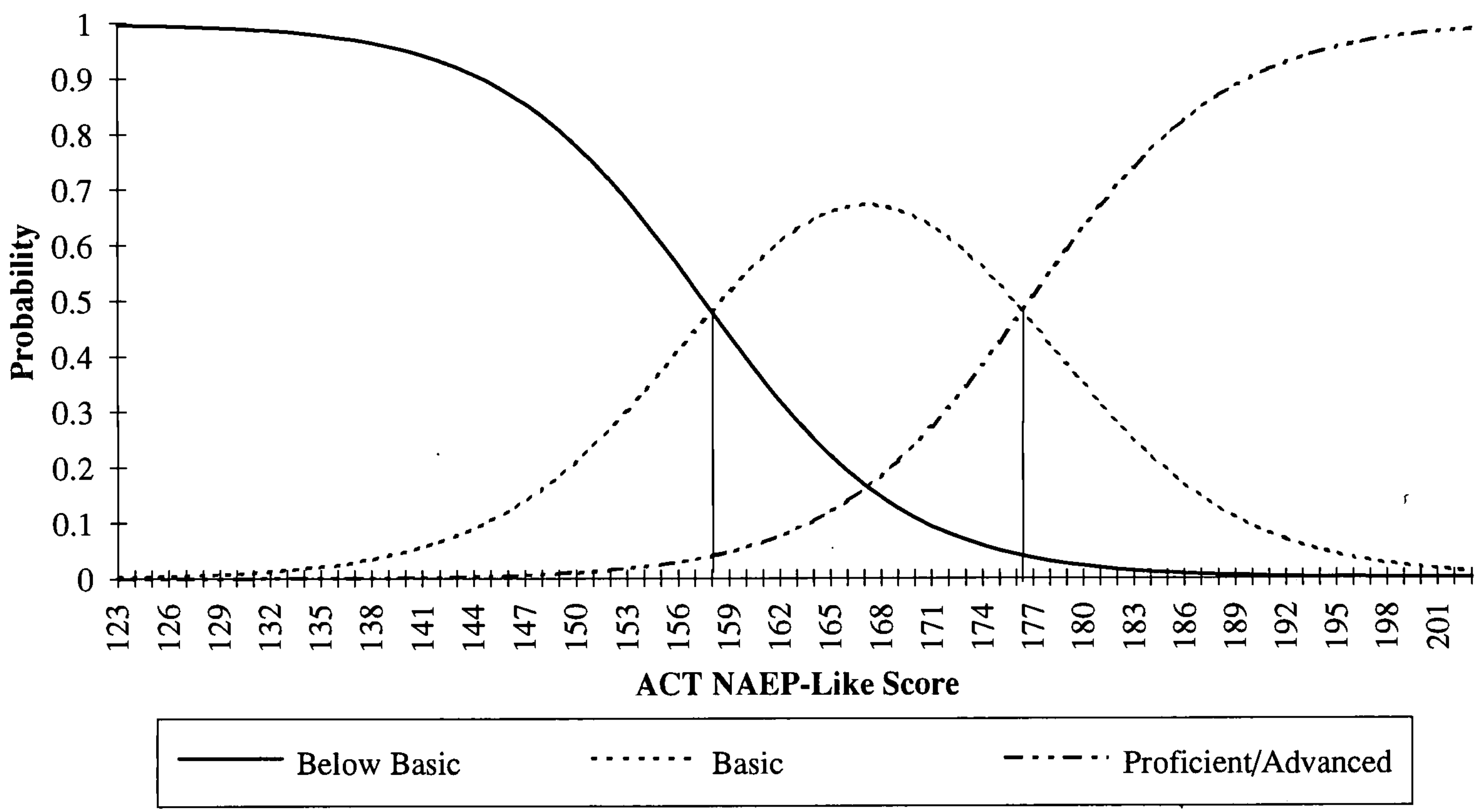


Figure 2
Cutpoints from Proportional Odds Model for Individual Panelists (Science Grade 8)

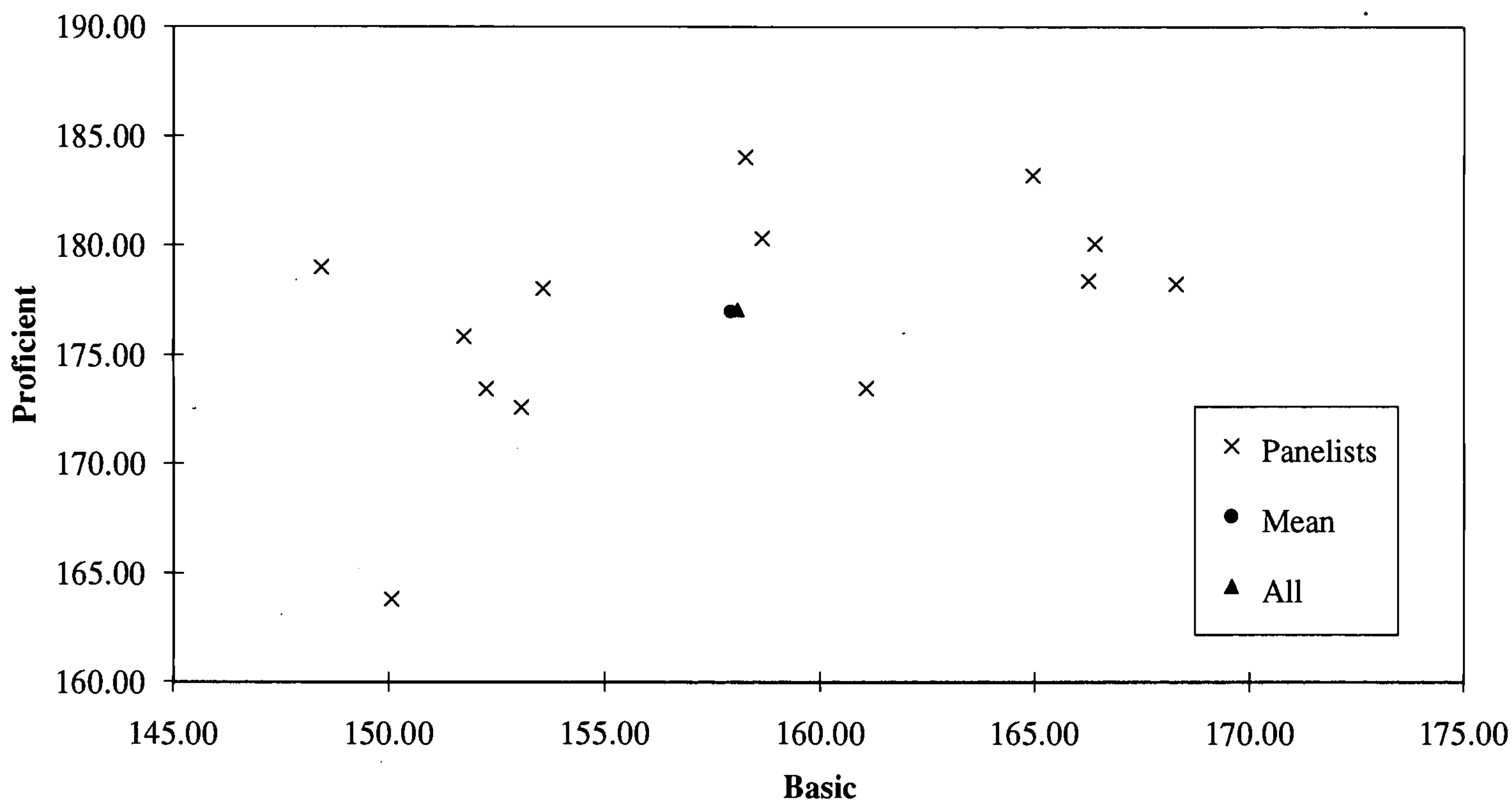
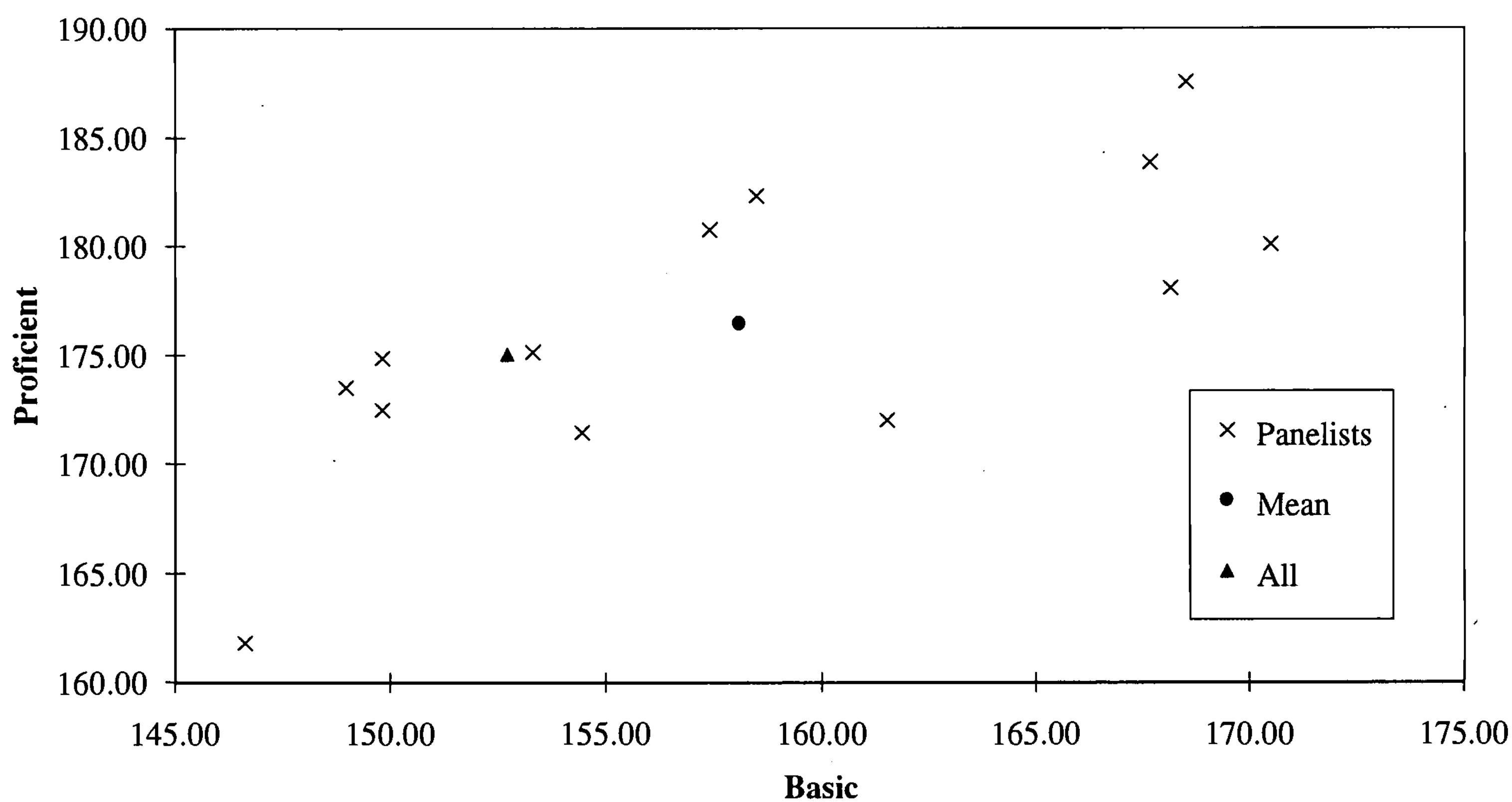


Figure 3
Cutpoints from Nonparametric Discriminant Analysis (Science Grade 8)



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

ERIC®

TM028903

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Booklet Classification Study	
Author(s): Luz Bay, Bradley Hanson, Susan Cooper Loomis	
Corporate Source: ACT, Inc.	Publication Date:

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	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	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	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
[X]	[ ]	[ ]
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	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: Luz Bay/Psychometrician	
Organization/Address: ACT, 2201 N. Dodge St., PO Box 168 Iowa City, IA 52243-0168	Telephone: 319-337-1638	FAX: 319-337-1497
	E-Mail Address: Bay@act.org	Date: 5-11-98

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com