# Classifying Student Performance as a Method for Setting Achievement Levels for NAEP Writing

Bradley A. Hanson
ACT, Inc.

Luz G. Bay
Advanced Systems

Revised April 16, 1999

Presented at the Annual Meeting of the National Council
on Measurement in Education

Montreal
April 1999

## Classifying Student Performance as a Method for Setting Achievement Levels for NAEP Writing

One way standard setting methods can be distinguished is by whether they are test-centered or examinee-centered (Jaeger, 1989; Kane, 1998). In test-centered methods judges decide on a level of performance that is considered just adequate for the standard to be met on each item or task in the test. The item-level judgments are combined to obtain a set of performance level cutpoints for the test as a whole. In examinee-centered methods judges categorize examinees directly into performance levels using definitions of adequate performance for each level and information about the level of achievement of each examinee. Test-centered methods have been used for setting NAEP standards. This paper examines an examinee-centered standard setting method (Booklet Classification) as a possible standard setting method for the 1998 NAEP Writing assessment. The study described in this paper is the second NAEP 1998 Writing field trial for achievement levels-setting.

There are three NAEP achievement levels: 1) basic, 2) proficient, and 3) advanced. Achievement Level Descriptions (ALD) describe what students should know and be able to do at the basic, proficient and advanced achievement levels. Achievement that does not meet the requirements for the basic level is classified as below basic. The Booklet Classification method involves panelists using their interpretation of what constitutes acceptable performance at each achievement level (based on the ALD and other information) to classify completed NAEP booklets in achievement levels. The task for the panelists in the Booklet Classification method is to make a holistic judgment about a student's level of achievement based on the panelists' understanding of the achievement levels and a sample of the student's work as represented by the responses of that student to the items in a NAEP booklet.

### Method

Booklet classification was one of two standard setting methods evaluated in the second NAEP 1998 Writing achievement levels-setting (ALS) field trial. Only grade 8 was included in the second field trial. NAEP is also administered in grades 4 and 12, and standards will also be set for those grades in the operational ALS. Panelists in the field trial were assigned to one of four groups labeled A, B, C, and D. Booklet classification was used in groups A and B, and a test-centered method (the Reckase method — for details see Loomis, Bay, Yang, and Hanick, 1999) was used in groups C and D. Groups A and C received

consequences data after round 1 of the standard setting process, and groups B and C did not receive consequences data after round 1 (all groups received consequences data after round 2). Consequences data consists of information on the proportion of students nationally that would be classified at or above each achievement level using the cutpoints set by the panelists. This paper discusses procedures and results for groups A and B only.

To implement the booklet classification method booklets needed to be selected and assigned to panelists to be classified. In addition, a method was needed to to compute cut-points on the NAEP scale using the booklet classifications of the panelists. The following sections discuss how these issues were addressed for this study.

**Panelists**

There were 10 panelists each in groups A and B. Panelists included teachers, non-teacher educators, and members of the general public. In group A there were 7 teacher panelists, 1 non-teacher educator panelist, and 2 general public panelists. In group B there were 8 teacher panelists, 1 non-teacher educator panelist, and 1 general public panelist. Eight of the ten panelists in group A were female, and nine of the ten panelists in group B were female. Panelists were recruited from eastern Iowa and western Illinois. They were paid $300 for their participation in the two-day study.

**Booklets**

Booklets from the 1992 NAEP Writing assessment were used because materials from the 1998 Writing assessment were not available in time for the field trial. There were ten forms of the 1992 NAEP Writing assessment used in the study. Each of the ten forms consisted of two unique 25-minute writing prompts. Each prompt was one of three types: narrative, informative, and persuasive. The types of the two prompts in each of the 10 forms used in the study are given in Table 1. The order of the prompt types was balanced across the forms so that narrative, informative, and persuasive prompts appeared as the first or second prompt in the form about the same number of times. Table 1 also contains prompt numbers identifying the unique prompts used on the forms. Some prompts were used on multiple forms. For example, prompt 9 is used as the second prompt of form 5, the first prompt of form 6, and the first prompt of form 8.

A booklet contains the responses of an examinee to the two prompts in a form. Thirty booklets were selected for each form. The NAEP holistic scoring of each prompt is on a scale of 1 to 6 Items for which there was no response or an off task response were coded zero (no booklet with a zero on either prompt was used in the study). In scaling the

data, Persuasive prompts were recoded so that a few "6's" were collapsed into the "5" category. Thirty completed booklets were chosen for each form representing a wide range of performance on each prompt. Table 2 gives the score combinations on the two prompts for all 30 booklets selected for each of the 10 forms. For each form the score combinations for the booklets in Table 2 are divided into five categories based on the sum of the scores on each prompt (total score on booklet): 1) total scores 2 and 3, 2) total scores 4 and 5, 3) total scores 6 and 7, 4) total scores 8 and 9, and 5) total scores 10, 11, and 12.

Each of the 20 panelists in the study classified 20 booklets from each of two forms (a total of 40 booklets per panelist). It was important that panelists classify some common booklets in order to be able to discuss their booklet classifications. To arrange for panelists to read common booklets, 30 booklets for each form were organized into 3 groups, 10 booklets per group. The three groups of booklets for each form were labeled X, Y, and Z (within each form there were three sets of 10 booklets labeled X, Y, and Z).

The system for assigning booklets to panelists in illustrated in Figure 1. In Figure 1, F1 through F10 stand for forms 1 through 10, and within each form the three booklet groups are indicated by X, Y, and Z. Panelists are indicated by P1 through P10 (the design was replicated for panelist groups A and B). As an example of how to interpret Figure 1 consider panelist 3 (labeled P3). Panelist 3 classified the 20 booklets from booklet groups Y and Z of form 3 (labeled F3), and 20 booklets from booklet groups X and Y of form 4 (labeled F4). Note that panelist 3 has the ten booklets in group Y of form 3 in common with panelist 2, and the ten booklets in group Y of form 4 in common with panelist 4. Each panelist was able to discuss 20 duplicate booklets with two other panelists, 10 booklets for each partner.

As can be seen from the information in Table 1 and Figure 1, every panelist reviewed booklets from four distinct prompts (two per form), and at least one each of the three types of prompts (narrative, informative, and persuasive). The fourth prompt type was evenly distributed across the three types of prompts. An exception was panelist 5 who only reviewed booklets from 3 prompts because the second prompt of form 5 was the same as the first prompt of form 6 (see Table 1).

The generalized partial credit model (Muraki, 1992) was used to scale the NAEP Writing prompts for the purposes of reporting NAEP results. For each booklet the scores on the two prompts were used along with the item parameter estimates for the prompts from the generalized partial credit model to produce a maximum likelihood $\theta$ estimate for

the booklet (the $\theta$ estimate is a value on the latent variable scale of the generalized partial credit model which represents examinee proficiency). Maximum likelihood estimates of $\theta$ were constrained to the interval from -4 to 4. The maximum likelihood $\theta$ estimate for each booklet was converted to a value on the ACT NAEP-Like scale (the ACT NAEP-Like scale is similar to the scale used to report NAEP results, but not identical).

The 30 booklets within each form were ordered by ACT NAEP-Like scale scores. Booklets were evenly distributed to the three booklet groups by rank on the ACT NAEP-Like scores. The sum of the rankings across each set of 20 booklets classified by a panelist was approximately equal. Table 3 shows how the 30 booklets in any of the ten forms are assigned to the three booklet groups and the two panelists who classify booklets from that form. The first column in Table 3 indicates the booklets ranked by ACT NAEP-Like score. The next three columns indicate which 10 booklets were assigned to each of the three booklet groups (X, Y, and Z). The last two columns show which 20 booklets were assigned to the two panelists who classified booklets from the form.

**Booklet Classification Task**

As noted above each panelist classified 20 booklets from each of 2 forms (a total of 40 booklets). Each booklet was classified in one of seven levels: 1) below basic, 2) borderline basic, 3) basic, 4) borderline proficient, 5) proficient, 6) borderline advanced, 7) advanced. Within each of the three NAEP achievement levels (basic, proficient, and advanced) panelists were asked to distinguish booklets that were borderline (represented achievement just barely meeting the achievement level description) from those that were non-borderline (representing behavior that solidly meets the achievement level description, but does not even minimally meet the description for the next higher achievement level). For example, a booklet classified as borderline proficient would be clearly distinguished from basic performance, but just barely meet the requirements for proficient performance. A booklet classified as proficient would meet the definition of proficient performance in more than a minimal way, but not even minimally meet the requirements to be considered advanced.

As noted in the previous section an ACT NAEP-Like scale score estimate was computed for each booklet based on a maximum likelihood $\theta$ estimate. Within each form the 20 booklets were presented to panelists ordered in terms of ACT NAEP-Like scale scores. Only the ranking of the booklets in terms of ACT NAEP-Like scale scores were presented to panelists, not the actual ACT NAEP-Like scale scores for the booklets. The panelists

were told that the ordering of the booklets was only one of several ways to order the booklets with respect to student performance. The panelists were advised that the booklets were ordered as an aid in performing the classification task, and that they were free to classify the booklets "out of order."

The Achievement Levels-Setting (ALS) process in the second Writing field trial took place over two days. Panelists first went through an orientation to provide a common understanding of the purposes for setting achievement levels and the procedures to follow in the ALS process. Panelists were then presented with the Writing framework and achievement levels descriptions. Panelists performed exercises in order to reach a common understanding of the the meaning of the achievement levels descriptions. The panelists then completed the round 1 classifications of booklets. After the round 1 classifications basic, proficient, and advanced cutpoints were computed for each panelist and cutpoints across all panelists were computed (see the next section for details on how the cutpoints were computed). The cutpoints were provided as feedback to the panelists. Group A was also provided with consequences data (the percentage of examinees nationally that would be at or above each achievement level based on the cutpoints set by the panelists). Group B did not receive consequences data after round 1. Before the round 2 classifications panelists discussed their round 1 classifications on common booklets. Each panelist had 10 booklets in common with 2 other panelists (these are the group Y booklets in Figure 1). After the discussion of the round 1 common booklet classifications the panelists completed round 2 of the booklet classification. Feedback was then provided for the round 2 classifications, including consequences data for both groups A and B. Finally, agreement was reached by panelists on final group cutpoints.

**Computing Cutpoints**

Each panelist classifies 40 booklets into the seven categories. The classifications for each panelist are used with the estimated ACT NAEP-Like scale scores for the booklets (converted from the maximum likelihood estimates of $\theta$ for each booklet) to compute four sets of three cutpoints (basic, proficient, and advanced) for the panelist in each round. The four sets of cutpoints for each panelist are computed using four methods: 1) Collapsed Categories, 2) Average Borderline, 3) Weighted Collapsed and Borderline, and 4) Cubic Regression. The Weighted Collapsed and Borderline method was used operationally in the standard setting process to compute the cutpoints used to provide feedback to the panelists. Results will be reported using all four sets of cutpoints in order to assess the

extent to which the cutpoints are affected by the method used to compute them.

To obtain overall cutpoints across panelists for a particular method mean cutpoints were computed across panelists. For example, the overall basic cutpoint using the Cubic Regression method is the mean of the basic cutpoints computed using the Cubic Regression method across panelists. Overall cutpoints were computed for groups A and B separately. Computational details for the four methods of computing cutpoints are given in the following subsections.

*Collapsed Categories.* For this method of computing cutpoints the seven categories are collapsed into four categories by assigning booklets classified in a borderline category to the corresponding main category (i.e., borderline basic is collapsed with basic, borderline proficient is collapsed with proficient, and borderline advanced is collapsed with advanced). The booklets as classified into the four categories are used to set cutpoints.

A decision rule is defined by basic, proficient and advanced cutpoints on the ACT NAEP-Like proficiency scale: $c_1$, $c_2$, and $c_3$, respectively. For a decision rule defined by the cutpoints $\mathbf{c} = (c_1, c_2, c_3)$ there can be booklets that are classified in the same category by the decision rule and the panelist (consistent classifications), and booklets that are classified in a different category by the decision rule and the panelist (inconsistent classifications). If the cost of a consistent classification is zero and the cost of an inconsistent classification is one (no matter how different the classifications of the panelist and decision rule are) then the Bayes risk of a decision rule is given by

$$\begin{aligned} r(\boldsymbol{\pi}, \mathbf{c}) = \pi_0 p(\eta \geq c_1 \mid l = 0) + \pi_1 [p(\eta < c_1 \mid l = 1) + p(\eta \geq c_2 \mid l = 1)] \\ + \pi_2 [p(\eta < c_2 \mid l = 2) + p(\eta \geq c_3 \mid l = 2)] + \pi_3 p(\eta < c_3 \mid l = 3)\,, \end{aligned} \tag{1}$$

where $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2, \pi_3)$ are the prior probabilities of a booklet being classified as below basic, basic, proficient, and advanced, respectively, and $p(\eta \geq c_1 \mid l = 0)$ is the probability of the ACT NAEP-Like proficiency ($\eta$) for a booklet being greater than equal to the basic cutpoint given the booklet is classified by the panelist as below basic (the other probabilities in Equation 1 are similarly interpreted). The Bayes rule is given by the cutpoints that minimize Equation 1.

The observed proportions classified in each category are used as the prior probabilities ($\pi_j = n_j/n$, where $n_j$ is the number booklets classified at level $j$ and $n = n_0 + n_1 + n_2 + n_3$) in computing the Bayes rule. The other probabilities in Equation 1 are estimated by their observed values. For example, $p(\eta \geq c_1 \mid l = 0)$ is estimated by the proportion of booklets classified as below basic whose $\eta$ is greater than or equal to $c_1$.

Since there are a finite number of booklets classified by the panelists (and consequently a finite number of observed values of $\eta$) there are only a finite number of decision rules that will result in different values of the Bayes risk. One way to compute the Bayes rule would be to compute the Bayes risk of each possible decision rule and choose the rule with the smallest Bayes risk. To simplify this procedure each cutpoint was computed separately by computing the Bayes rule for a two category problem. For cutpoint $j, j = 1, 2, 3$, the Bayes risk for classifying a booklet given the booklet is in category $j-1$ or $j$ is

$$r(\boldsymbol{\pi}, c_j) = \pi_{j-1} p(\eta \geq c_j \mid l = j-1) + \pi_j p(\eta < c_j \mid l = j)\,. \tag{2}$$

In this case $\pi_j = n_j/(n_{j-1} + n_j)$ and $\pi_{j-1} = n_{j-1}/(n_{j-1} + n_j)$. Each of the cutpoints is chosen such that Bayes risk of the cutpoint given by Equation 2 is minimized.

An example of cutpoints computed using the Collapsed Category method is given in the top plot of Figure 2. This plot contains the collapsed round 2 classifications of 40 booklets for panelist 8 of group B (panelist B08). Each point represents a booklet, with the vertical axis giving the classification of the booklet by panelist B08 and the horizontal axis giving the maximum likelihood estimate of the ACT NAEP-Like scale score for the booklet (there are some cases where two or more booklets classified at a given level have the same NAEP ACT-Like scale score, so there are a few points in the plot that represent more than one booklet). The basic cutpoint was computing using the booklets classified as below basic and basic. The basic cutpoint results in two below basic booklets being above the cutpoint, and one basic booklet being below the cutpoint (these misclassifications are associated with the two misclassification probabilities involved in the Bayes risk as given in Equation 2). If the basic cutpoint was moved up to eliminate the below basic misclassifications, then the corrected below basic misclassifications would be more than offset by additional basic misclassifications. The analogous situation occurs if the basic cutpoint were moved down to eliminate the basic misclassifications. This illustrates how the procedure is minimizing the Bayes risk (which is a function of misclassification probabilities).

*Average Borderline.* This method of setting cutpoints uses only booklets classified in a borderline category (non-borderline booklets are not used at all in this method). The cutpoint for each category is the mean $\eta$ of the borderline booklets for that category.

Let $m_1, m_2$, and $m_3$ be the number of booklets classified as borderline basic, borderline proficient, and borderline advanced. Cutpoints $c_j$, $j = 1, 2, 3$, are given by

$$c_j = \frac{1}{m_j} \sum_{i \in B_j} \eta_i\,, \tag{3}$$

where $B_j$ is the set containing the indices of the booklets classified in category $j$ ($j = 1, 2, 3$ for borderline basic, borderline proficient, and borderline advanced), and $\eta_i$ is the estimated $\eta$ for booklet $i$.

An example of cutpoints computed using the Average Borderline method are given in the bottom plot of Figure 2. The bottom plot in Figure 2 contains the round 2 classifications of 40 booklets for panelist B08 (the top portion of Figure 2 gives the collapsed versions of these classifications). Only the borderline booklets (indicated by the hollow symbols) are involved in the computation of the the cutpoints using the Average Borderline method. For example, the basic cutpoint is the average of the ACT NAEP-Line scale scores of the five borderline basic booklets (only four borderline basic points appear in the plot because two of the booklets have the same ACT NAEP-Like scale score). Similarly, the proficient cutpoint is the average of the ACT NAEP-Like scale scores for the borderline proficient booklets. Note that for this panelist the proficient cutpoint is *below* the basic cutpoint.

*Weighted Collapsed and Borderline.* Weighted Collapsed and Borderline cutpoints were computed by taking a weighted average of the Collapsed Categories and Average Borderline cutpoints for each panelist. The weight given to the Collapsed Category cutpoint $j$ ($j = 1, 2, 3$ for basic, proficient, and advanced, respectively) is

$$\frac{n_{j-1} + n_j}{n_{j-1} + n_j + m_j}\,, \tag{4}$$

and the weight given to the Average Borderline cutpoint is

$$\frac{m_j}{n_{j-1} + n_j + m_j}\,. \tag{5}$$

The value of the Weighted Collapsed and Borderline cutpoint for a particular achievement level is the Collapsed Category cutpoint at that achievement level multiplied by Equation 4 plus the Average Borderline cutpoint at that achievement level multiplied by Equation 5.

The process of combining the two sets of cutpoints results in more weight being given to booklets classified as borderline than those not classified as borderline in determining the cutpoints.

*Cubic Regression.* The Cubic Regression method is an implementation of a method described in Plake and Hambleton (1998). The Cubic Regression method is illustrated in

Figure 3, which contains round 2 ratings for B08. There are 40 points in Figure 3 corresponding to the 40 booklets rated by panelist B08. These are the same points represented in the bottom plot in Figure 1, although the axes have been switched in Figure 3 to more naturally illustrate the regression being performed (ACT NAEP-Like scale score is the dependent variable, and achievement level is the independent variable). The value of each point on the horizontal axis is the achievement level in which the booklet was classified by panelist B08. The achievement level categories are assigned numerical values as follows (Plake and Hambleton, 1998): below basic = 1; borderline basic = 2, basic = 3; borderline proficient = 4; proficient = 5; borderline advanced = 6; advanced = 7. The value of each point on the vertical axis is the estimated ACT NAEP-like scale score for the booklet.

Cutpoints using the Cubic Regression method are obtained by fitting a cubic regression model to the data in Figure 3 with achievement level as the independent variable (the regression uses the numerical values associated with the achievement level categories described above) and ACT NAEP-like scale score as the dependent variable. The solid line in Figure 3 gives the cubic regression curve fitted to these data. The curve represents the conditional mean ACT NAEP-like scale score as a function of achievement level. Cutpoints are given by the values of the cubic regression curve corresponding to borderline basic, borderline proficient, and borderline advanced levels (i.e., values of the regression curve corresponding to numerical achievement levels of 2, 4 and 6).

The Average Borderline method and the Cubic Regression method both define the cutpoints as the conditional mean ACT NAEP-like scale score corresponding to borderline categories. The methods differ in how this mean is calculated. For the Average Borderline method the mean is directly computed from the booklets classified in the borderline categories. For the Cubic Regression method a regression curve is computed giving the conditional mean ACT NAEP-like scale score at all levels, and the cutpoints are computed as the value of the regression curve at achievement levels of 2, 4, and 6.

## Results

Tables 4 through 7 give four sets of cutpoints for each panelist in each round (one set for each of the four methods of computing cutpoints). Tables 4 and 5 give the round 1 and round 2 cutpoints for panelists in group A, and tables 6 and 7 give the round 1 and round 2 cutpoints for panelists in group B. Levels 1, 2 and 3 in the tables correspond to cutpoints for the basic, proficient, and advanced levels, respectively. No round 1 basic cutpoint using the Average Borderline method is reported for panelist 3 in group B (panelist B03) due

to the fact that this panelist did not classify any booklets as borderline basic in round 2. Rows labeled "B08" in Table 7 contain the cutpoints presented Figures 2 and 3. Table 8 gives the average cutpoints over panelists for each round and panelist group.

There is considerable variability in the cutpoints across panelists. There are also some considerable differences among the cutpoints computed using the four methods. The cutpoints using the Average Borderline and Cubic Regression methods tend to be similar to one another. This is not surprising since they use the same cutpoint definition and just use different methods to compute the cutpoints. The average cutpoints for the Collapsed Categories method tend to be lower than the cutpoints for the other methods (especially the basic and proficient cutpoints).

Table 9 gives the distribution of round 2 category classification by sorted booklet order. The counts are combined across the two groups of panelists and the two forms. There is a relationship between the order of the booklets and the category in which the booklets were classified. Lower ranked booklets tended to be classified in lower achievement levels, and higher ranked booklets classified in higher achievement levels.

Table 10 gives classifications of the booklets for each panelist in round 2 by booklet order. This shows that panelists did not classify booklets strictly by rank (e.g., classifying booklets ranked 1-4 as below basic, booklets ranked 5 and 6 as borderline basic, booklets 7-10 as basic, booklets 11-12 as borderline proficient, booklets 13-17 as proficient, booklets 18 and 19 as borderline advanced, and booklet 20 as advanced). This alleviates the concern that panelists would only use booklet ranks as the basis for their classifications rather than deciding on the level of a booklet from their interpretation of how well the performance on the prompts in the booklet represents the achievement levels as described by the achievement level descriptions.

Table 11 gives overall round 2 cutpoints computed using the Weighted Collapsed and Borderline method for several alternative proficiency estimates (these are alternative $\theta$ estimates that are converted to the ACT NAEP-Like proficiency scale). The cutpoints in the first row labeled "ML" are the maximum likelihood proficient estimates computed from the two prompt scores that were used as proficient estimates in this study (i.e., the same cutpoints reported in Tables 6 and 8). The other cutpoints are based on using plausible values for the booklets that were computed for the purpose of reporting NAEP results (e.g., distributions of proficiency and proportions above the proficiency level cutpoints). The plausible values are computed using the responses of the examinee to the prompts (the

same data used to compute the maximum likelihood estimates used in this study) along with responses of the examinee to NAEP background and attitude questions (Mislevy, Johnson, and Muraki, 1992). There were five plausible values generated for each booklet. The rows labeled "PV1" through "PV5" give cutpoints computed using each of the five plausible values. The last row (labeled "Mean PV") gives the mean cutpoints across the five plausible values. Besides the mean cutpoints across panelists (overall cutpoints), Table 11 gives the standard deviation of the cutpoints across panelists, and estimates of the percentages of students nationally whose Writing proficiency is greater than each cutpoint.

The basic cutpoint is lower using the maximum likelihood estimate than any of the basic cutpoints computed using plausible values. The advanced cutpoint using maximum likelihood estimate is higher than the advanced cutpoints computed using plausible values. The proficient cutpoints are about the same using maximum likelihood estimates or plausible values.

## Discussion

This paper described the procedures and results from a field trial that examined using booklet classification as a standard setting method to set achievement levels for the 1998 NAEP Writing assessment. There were differences among cutpoints set using different methods of computing the cutpoints, and different methods of computing $\theta$ estimates for the booklets.

Figures 2 and 3 present an example in which the basic and proficient cutpoints computed using the different methods differed (the cutpoints for panelist B08 given in Table 7). One reason for the disparity among the basic and proficient cutpoints computed using the different methods in this example is that the booklets classified as below basic, borderline basic, basic, and borderline proficient are not well distinguished in terms of ACT NAEP-Like scale scores (see the bottom plot in Figure 2). If the booklets at various achievement levels are not well distinguished in terms of their ACT NAEP-Like scale scores there is no clear way to set the cutpoints, and different methods are likely to come up with different answers. Examination of additional plots like Figures 2 and 3 indicated that the cutpoints computed by the various methods were closer when the conditional distributions of ACT NAEP-like scale scores for the levels were better separated.

Table 11 shows a large difference between cutpoints computed using maximum likelihood $\theta$ estimates and plausible values. More students would be classified at or above basic and fewer students classified as advanced using the cutpoints computed with maximum

likelihood estimates than using cutpoints computed using plausible values. In addition, there is some reasonable variability in the cutpoints computed using different plausible values (from 43 to 56 percent of students would be classified at or above proficient using the group A cutpoints for the various plausible values). Fairly accurate $\theta$ estimates for booklets are needed for the booklet classification method to perform adequately. Since each booklet only contains the responses to two prompts it is debatable whether the $\theta$ estimates computed for booklets can be accurate enough to provide stable cutpoints using booklet classification. NAEP was not designed to provide accurate estimates of proficiency for individual students, but such individual estimates are needed in order to use booklet classification as a standard setting method. This characteristic of NAEP may imply that use of the booklet classification method in the achievement levels-setting process is problematic.

The variability of the cutpoints across the methods of computing cutpoint and across the different ways of estimating $\theta$ for booklets raised concerns about whether booklet classification could provide achievement level cutpoints that were stable and defensible. Based partly on concerns raised by the results of this study, the decision was made not to use booklet classification to determine achievement level cutpoints for the 1998 NAEP Writing assessment. Instead, a test-centered standard setting procedure was used (Loomis, Bay, Yang, and Hanick, 1999).

## References


Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). New York: American Council on Education & Macmillan.

Kane, M. (1998). Criterion bias in examinee-centered standard setting: Some thought experiments. *Educational Measurement: Issues and Practice*, *17*, 23-30.

Mislevy, R. J., Johnson, E. G., Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*, 131-154.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Plake, B. S., & Hambleton, R. K. (1998). *A standard setting method designed for complex performance assessments with multiple performance categories: Categorical assignments of student work.* Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, April)

Loomis, S. C., Bay, L. G., Yang, W., & Hanick, P. (1999). *Field trials to determine which standard setting method to use.* Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, April).

**Table 1:** Prompt Types and Numbers in Each Form.

| | First Prompt | | Second Prompt | |
|---|---|---|---|---|
| **Form Number** | **Number** | **Type** | **Number** | **Type** |
| 1 | 7 | N | 8 | N |
| 2 | 4 | I | 11 | P |
| 3 | 10 | P | 8 | N |
| 4 | 3 | I | 4 | I |
| 5 | 7 | N | 9 | P |
| 6 | 9 | P | 5 | I |
| 7 | 6 | I | 4 | N |
| 8 | 9 | P | 10 | P |
| 9 | 5 | N | 7 | I |
| 10 | 3 | I | 10 | P |

N = narrative
I = informative
P = persuasive

**Table 2**: Scores on the Two Prompts for the 30 Booklets Used for Each Form.

| | Total Score | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2,3 | | 4,5 | | 6,7 | | 8,9 | | 10,11,12 | |
| Form | Prompt 1 | Prompt 2 | Prompt 1 | Prompt 2 | Prompt 1 | Prompt 2 | Prompt 1 | Prompt 2 | Prompt 1 | Prompt 2 |
| 1 | 2 | 1 | 3 | 1 | 5 | 1 | 6 | 2 | 6 | 4 |
| | 1 | 2 | 2 | 2 | 4 | 2 | 6 | 2 | 6 | 4 |
| | | | 1 | 3 | 3 | 3 | 5 | 3 | | |
| | | | 4 | 1 | 3 | 3 | 5 | 3 | | |
| | | | 3 | 2 | 5 | 2 | 4 | 4 | | |
| | | | 2 | 3 | 5 | 2 | 4 | 4 | | |
| | | | 1 | 4 | 5 | 2 | 6 | 3 | | |
| | | | | | 4 | 3 | 5 | 4 | | |
| | | | | | 4 | 3 | | | | |
| | | | | | 3 | 4 | | | | |
| | | | | | 3 | 4 | | | | |
| 2 | 2 | 1 | 3 | 1 | 5 | 1 | 6 | 2 | 6 | 4 |
| | 1 | 2 | 2 | 2 | 4 | 2 | 6 | 2 | 6 | 4 |
| | | | 1 | 3 | 3 | 3 | 5 | 3 | | |
| | | | 4 | 1 | 3 | 3 | 5 | 3 | | |
| | | | 3 | 2 | 5 | 2 | 4 | 4 | | |
| | | | 2 | 3 | 5 | 2 | 4 | 4 | | |
| | | | 1 | 4 | 5 | 2 | 6 | 3 | | |
| | | | | | 4 | 3 | 5 | 4 | | |
| | | | | | 4 | 3 | | | | |
| | | | | | 3 | 4 | | | | |
| | | | | | 3 | 4 | | | | |
| 3 | 1 | 2 | 1 | 3 | 2 | 4 | 2 | 6 | 4 | 6 |
| | 2 | 1 | 2 | 2 | 3 | 3 | 3 | 5 | 5 | 5 |
| | | | 3 | 1 | 3 | 3 | 4 | 4 | 5 | 6 |
| | | | 3 | 1 | 4 | 2 | 4 | 4 | | |
| | | | 2 | 3 | 4 | 2 | 5 | 3 | | |
| | | | 2 | 3 | 2 | 5 | 4 | 5 | | |
| | | | 3 | 2 | 3 | 4 | | | | |
| | | | 3 | 2 | 4 | 3 | | | | |
| | | | 4 | 1 | 4 | 3 | | | | |
| | | | | | 5 | 2 | | | | |
| 4 | 1 | 1 | 3 | 1 | 5 | 1 | 5 | 3 | 5 | 5 |
| | 1 | 2 | 2 | 2 | 4 | 2 | 5 | 3 | 4 | 6 |
| | | | 4 | 1 | 3 | 3 | 4 | 4 | 5 | 6 |
| | | | 3 | 2 | 2 | 4 | 4 | 4 | 6 | 5 |
| | | | 1 | 4 | 1 | 5 | 3 | 5 | | |
| | | | | | 4 | 3 | 5 | 4 | | |
| | | | | | 4 | 3 | 4 | 5 | | |
| | | | | | 3 | 4 | 4 | 5 | | |
| | | | | | 3 | 4 | 3 | 6 | | |
| | | | | | 3 | 4 | | | | |

**Table 2**: Scores on the Two Prompts for the 30 Booklets Used for Each Form.

| | Total Score | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2,3 | | 4,5 | | 6,7 | | 8,9 | | 10,11,12 | |
| Form | Prompt 1 | Prompt 2 | Prompt 1 | Prompt 2 | Prompt 1 | Prompt 2 | Prompt 1 | Prompt 2 | Prompt 1 | Prompt 2 |
| 5 | 1 | 2 | 1 | 3 | 2 | 4 | 4 | 4 | | |
| | 2 | 1 | 2 | 2 | 3 | 3 | 5 | 3 | | |
| | | | 2 | 3 | 3 | 3 | 6 | 2 | | |
| | | | 2 | 3 | 4 | 2 | 6 | 2 | | |
| | | | 3 | 2 | 4 | 2 | 5 | 4 | | |
| | | | 3 | 2 | 4 | 2 | 5 | 4 | | |
| | | | 4 | 1 | 4 | 2 | 6 | 3 | | |
| | | | | | 5 | 1 | 6 | 3 | | |
| | | | | | 3 | 4 | | | | |
| | | | | | 3 | 4 | | | | |
| | | | | | 4 | 3 | | | | |
| | | | | | 4 | 3 | | | | |
| | | | | | 5 | 2 | | | | |
| 6 | 1 | 1 | 1 | 3 | 2 | 4 | 3 | 5 | | |
| | 1 | 2 | 2 | 2 | 2 | 4 | 4 | 4 | | |
| | 2 | 1 | 2 | 2 | 3 | 3 | 5 | 3 | | |
| | | | 3 | 1 | 4 | 2 | 4 | 5 | | |
| | | | 1 | 4 | 4 | 2 | | | | |
| | | | 2 | 3 | 2 | 5 | | | | |
| | | | 2 | 3 | 2 | 5 | | | | |
| | | | 3 | 2 | 2 | 5 | | | | |
| | | | 4 | 1 | 3 | 4 | | | | |
| | | | | | 3 | 4 | | | | |
| | | | | | 4 | 3 | | | | |
| | | | | | 4 | 3 | | | | |
| | | | | | 4 | 3 | | | | |
| | | | | | 4 | 3 | | | | |
| 7 | 1 | 2 | 1 | 3 | 2 | 4 | 3 | 5 | 4 | 6 |
| | 2 | 1 | 2 | 2 | 3 | 3 | 3 | 5 | 5 | 5 |
| | | | 3 | 1 | 4 | 2 | 4 | 4 | 6 | 4 |
| | | | 1 | 4 | 5 | 1 | 5 | 3 | 5 | 6 |
| | | | 2 | 3 | 2 | 5 | 5 | 3 | | |
| | | | 3 | 2 | 3 | 4 | 3 | 6 | | |
| | | | | | 3 | 4 | 4 | 5 | | |
| | | | | | 4 | 3 | 5 | 4 | | |
| | | | | | 4 | 3 | 5 | 4 | | |
| 8 | 2 | 1 | 3 | 1 | 4 | 2 | 4 | 4 | 4 | 6 |
| | 1 | 2 | 2 | 2 | 3 | 3 | 3 | 5 | | |
| | | | 1 | 3 | 3 | 3 | 5 | 4 | | |
| | | | 1 | 3 | 2 | 4 | | | | |
| | | | 1 | 3 | 2 | 4 | | | | |
| | | | 1 | 3 | 4 | 3 | | | | |
| | | | 4 | 1 | 4 | 3 | | | | |
| | | | 3 | 2 | 4 | 3 | | | | |
| | | | 2 | 3 | 4 | 3 | | | | |
| | | | 1 | 4 | 4 | 3 | | | | |
| | | | | | 3 | 4 | | | | |
| | | | | | 3 | 4 | | | | |
| | | | | | 3 | 4 | | | | |
| | | | | | 3 | 4 | | | | |

**Table 2**: Scores on the Two Prompts for the 30 Booklets Used for Each Form.

| | Total Score | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2,3 | | 4,5 | | 6,7 | | 8,9 | | 10,11,12 | |
| Form | Prompt 1 | Prompt 2 | Prompt 1 | Prompt 2 | Prompt 1 | Prompt 2 | Prompt 1 | Prompt 2 | Prompt 1 | Prompt 2 |
| 9 | 2 | 1 | 3 | 1 | 4 | 2 | 5 | 3 | 5 | 5 |
| | 1 | 2 | 2 | 2 | 3 | 3 | 4 | 4 | 4 | 6 |
| | | | 1 | 3 | 2 | 4 | 4 | 4 | 4 | 6 |
| | | | 4 | 1 | 5 | 2 | 3 | 5 | | |
| | | | 3 | 2 | 1 | 5 | 3 | 5 | | |
| | | | 2 | 3 | 4 | 3 | 5 | 4 | | |
| | | | 1 | 4 | 4 | 3 | 5 | 4 | | |
| | | | | | 3 | 4 | 4 | 5 | | |
| | | | | | 2 | 5 | | | | |
| | | | | | 2 | 5 | | | | |
| 10 | 1 | 1 | 3 | 1 | 4 | 2 | 5 | 3 | 6 | 4 |
| | 1 | 2 | 2 | 2 | 3 | 3 | 5 | 3 | 6 | 4 |
| | 1 | 2 | 1 | 3 | 2 | 4 | 4 | 4 | | |
| | | | 4 | 1 | 5 | 2 | 3 | 5 | | |
| | | | 3 | 2 | 5 | 2 | 3 | 5 | | |
| | | | 2 | 3 | 4 | 3 | 6 | 3 | | |
| | | | 1 | 4 | 4 | 3 | 5 | 4 | | |
| | | | | | 3 | 4 | 4 | 5 | | |
| | | | | | 3 | 4 | | | | |
| | | | | | 3 | 4 | | | | |

**Table 3**: Assignment of Booklets to Booklet Groups and Panelists for Any Form

| For any form f, where f = 1, 2,…,10 | | | | | |
|---|---|---|---|---|---|
| | | | | | |
| Form ID (by rank) | Form Group ID | | | Panelist 1 | Panelist 2 |
| f.01 | fX1 | | | 1 | |
| f.02 | | fY1 | | 2 | 1 |
| f.03 | | | fZ1 | | 2 |
| f.04 | | | fZ2 | | 3 |
| f.05 | | fY2 | | 3 | 4 |
| f.06 | fX2 | | | 4 | |
| f.07 | fX3 | | | 5 | |
| f.08 | | fY3 | | 6 | 5 |
| f.09 | | | fZ3 | | 6 |
| f.10 | | | fZ4 | | 7 |
| f.11 | | fY4 | | 7 | 8 |
| f.12 | fX4 | | | 8 | |
| f.13 | fX5 | | | 9 | |
| f.14 | | fY5 | | 10 | 9 |
| f.15 | | | fZ5 | | 10 |
| f.16 | | | fZ6 | | 11 |
| f.17 | | fY6 | | 11 | 12 |
| f.18 | fX6 | | | 12 | |
| f.19 | fX7 | | | 13 | |
| f.20 | | fY7 | | 14 | 13 |
| f.21 | | | fZ7 | | 14 |
| f.22 | | | fZ8 | | 15 |
| f.23 | | fY8 | | 15 | 16 |
| f.24 | fX8 | | | 16 | |
| f.25 | fX9 | | | 17 | |
| f.26 | | fY9 | | 18 | 17 |
| f.27 | | | fZ9 | | 18 |
| f.28 | | | fZ10 | | 19 |
| f.29 | | fY10 | | 19 | 20 |
| f.30 | fX10 | | | 20 | |

**Table 4:** Cutpoints for Group A in Round 1.

| | | Cutpoints | | | |
|---|---|---|---|---|---|
| Panelist | Level | Collapsed Categories | Average Borderline | Weighted | Cubic Regression |
| A01 | 1 | 119.45 | 133.38 | 122.77 | 138.27 |
| A02 | 1 | 128.71 | 135.47 | 130.27 | 147.00 |
| A03 | 1 | 111.54 | 134.02 | 118.46 | 130.81 |
| A04 | 1 | 119.47 | 139.12 | 127.11 | 130.65 |
| A05 | 1 | 115.31 | 136.70 | 123.86 | 136.48 |
| A06 | 1 | 132.78 | 141.06 | 135.54 | 136.76 |
| A07 | 1 | 133.11 | 129.09 | 132.16 | 131.57 |
| A08 | 1 | 116.55 | 149.29 | 120.40 | 128.95 |
| A09 | 1 | 126.24 | 130.08 | 126.96 | 131.65 |
| A10 | 1 | 133.89 | 156.42 | 143.54 | 155.38 |
| A01 | 2 | 164.23 | 167.23 | 164.65 | 170.07 |
| A02 | 2 | 150.21 | 174.33 | 151.63 | 159.72 |
| A03 | 2 | 152.62 | 169.19 | 155.38 | 159.17 |
| A04 | 2 | 166.01 | 163.33 | 165.17 | 168.26 |
| A05 | 2 | 141.79 | 161.77 | 144.75 | 162.33 |
| A06 | 2 | 165.02 | 180.29 | 170.27 | 178.65 |
| A07 | 2 | 148.75 | 172.93 | 156.19 | 168.49 |
| A08 | 2 | 163.97 | 155.37 | 162.03 | 163.24 |
| A09 | 2 | 141.80 | 158.38 | 146.32 | 157.87 |
| A10 | 2 | 148.17 | 169.80 | 154.66 | 171.76 |
| A01 | 3 | 193.56 | 196.90 | 194.08 | 192.86 |
| A02 | 3 | 170.88 | 167.44 | 169.99 | 171.06 |
| A03 | 3 | 195.96 | 194.45 | 195.68 | 191.10 |
| A04 | 3 | 205.96 | 209.89 | 206.32 | 206.96 |
| A05 | 3 | 200.28 | 184.85 | 196.58 | 183.70 |
| A06 | 3 | 210.60 | 190.87 | 208.52 | 192.07 |
| A07 | 3 | 187.89 | 198.01 | 190.09 | 194.45 |
| A08 | 3 | 207.72 | 185.51 | 203.86 | 193.86 |
| A09 | 3 | 177.33 | 188.95 | 180.72 | 189.94 |
| A10 | 3 | 203.36 | 186.22 | 198.85 | 187.30 |

**Table 5:** Cutpoints for Group A in Round 2.

| | | Cutpoints | | | |
|---|---|---|---|---|---|
| Panelist | Level | Collapsed Categories | Average Borderline | Weighted | Cubic Regression |
| A01 | 1 | 141.43 | 135.55 | 140.59 | 144.74 |
| A02 | 1 | 127.58 | 128.71 | 127.75 | 138.50 |
| A03 | 1 | 111.54 | 133.73 | 120.08 | 134.10 |
| A04 | 1 | 119.47 | 138.37 | 126.14 | 133.57 |
| A05 | 1 | 132.03 | 146.59 | 136.63 | 150.78 |
| A06 | 1 | 132.78 | 142.54 | 135.83 | 134.53 |
| A07 | 1 | 148.75 | 123.36 | 143.67 | 129.30 |
| A08 | 1 | 116.55 | 132.57 | 121.00 | 130.33 |
| A09 | 1 | 126.24 | 135.68 | 127.29 | 130.89 |
| A10 | 1 | 133.89 | 141.61 | 136.61 | 148.77 |
| A01 | 2 | 185.63 | 185.51 | 185.60 | 183.67 |
| A02 | 2 | 150.21 | 174.33 | 151.31 | 157.68 |
| A03 | 2 | 152.62 | 159.34 | 153.67 | 161.40 |
| A04 | 2 | 151.80 | 165.91 | 155.43 | 165.11 |
| A05 | 2 | 141.79 | 171.71 | 149.85 | 168.48 |
| A06 | 2 | 143.77 | 160.55 | 148.98 | 165.99 |
| A07 | 2 | 133.11 | 165.70 | 146.69 | 165.50 |
| A08 | 2 | 156.67 | 171.10 | 160.86 | 168.07 |
| A09 | 2 | 155.02 | 166.93 | 157.89 | 166.10 |
| A10 | 2 | 148.17 | 173.53 | 158.13 | 172.00 |
| A01 | 3 | 211.00 | 199.65 | 209.11 | 207.13 |
| A02 | 3 | 177.60 | 185.28 | 180.45 | 181.74 |
| A03 | 3 | 195.96 | 183.46 | 194.10 | 186.62 |
| A04 | 3 | 205.96 | 208.79 | 206.08 | 201.66 |
| A05 | 3 | 200.28 | 184.28 | 198.00 | 183.55 |
| A06 | 3 | 205.97 | 199.25 | 204.85 | 194.85 |
| A07 | 3 | 183.35 | 197.58 | 185.63 | 195.19 |
| A08 | 3 | 207.72 | 186.51 | 204.83 | 196.76 |
| A09 | 3 | 194.68 | 191.17 | 194.04 | 198.84 |
| A10 | 3 | 190.71 | 180.25 | 187.98 | 183.05 |

**Table 6:** Cutpoints for Group B in Round 1.

| | | Cutpoints | | | |
|---|---|---|---|---|---|
| Panelist | Level | Collapsed Categories | Average Borderline | Weighted | Cubic Regression |
| A01 | 1 | 113.97 | 132.04 | 122.97 | 129.70 |
| B02 | 1 | 159.15 | 136.31 | 152.43 | 143.18 |
| B03 | 1 | 111.54 | 120.61 | 115.17 | 116.72 |
| B04 | 1 | 128.56 | 148.86 | 135.33 | 144.38 |
| B05 | 1 | 127.80 | 141.68 | 132.21 | 146.02 |
| B06 | 1 | 132.78 | 131.45 | 132.50 | 138.69 |
| B07 | 1 | 118.56 | 154.44 | 134.71 | 146.80 |
| B08 | 1 | 127.28 | 142.51 | 133.20 | 134.08 |
| B09 | 1 | 107.30 | 132.09 | 115.56 | 131.29 |
| B10 | 1 | 120.21 | 133.32 | 122.67 | 140.61 |
| B01 | 2 | 156.87 | 169.33 | 159.86 | 164.45 |
| B02 | 2 | 134.24 | 168.88 | 152.80 | 171.20 |
| B03 | 2 | 141.63 | 156.39 | 147.82 | 156.01 |
| B04 | 2 | 166.01 | 169.31 | 166.41 | 166.78 |
| B05 | 2 | 147.66 | 177.70 | 156.24 | 171.47 |
| B06 | 2 | 165.02 | 179.12 | 167.94 | 175.63 |
| B07 | 2 | 163.30 | 164.54 | 163.47 | 171.20 |
| B08 | 2 | 163.97 | 166.24 | 164.34 | 161.08 |
| B09 | 2 | 160.82 | 158.58 | 160.40 | 158.46 |
| B10 | 2 | 163.05 | 173.93 | 166.97 | 168.04 |
| B01 | 3 | 178.20 | 183.66 | 179.88 | 184.80 |
| B02 | 3 | 195.15 | 208.00 | 196.83 | 200.48 |
| B03 | 3 | 187.49 | 199.23 | 189.84 | 194.46 |
| B04 | 3 | 188.99 | 189.27 | 189.04 | 189.24 |
| B05 | 3 | 191.79 | 193.77 | 192.12 | 191.48 |
| B06 | 3 | 205.97 | 188.57 | 203.79 | 198.16 |
| B07 | 3 | 143.05 | 175.50 | 154.41 | 178.53 |
| B08 | 3 | 193.47 | 192.03 | 193.14 | 190.55 |
| B09 | 3 | 170.92 | 184.96 | 176.54 | 184.51 |
| B10 | 3 | 198.04 | 191.48 | 195.85 | 190.46 |

**Table 7:** Cutpoints for Group B in Round 2.

| | | Cutpoints | | | |
|---|---|---|---|---|---|
| Panelist | Level | Collapsed Categories | Average Borderline | Weighted | Cubic Regression |
| A01 | 1 | 119.45 | 137.50 | 127.79 | 132.97 |
| B02 | 1 | 127.58 | 133.58 | 129.83 | 134.92 |
| B03 | 1 | 134.09 | - | 134.09 | 131.70 |
| B04 | 1 | 119.47 | 144.52 | 127.18 | 136.92 |
| B05 | 1 | 127.80 | 141.38 | 132.74 | 146.99 |
| B06 | 1 | 132.78 | 130.00 | 131.96 | 132.49 |
| B07 | 1 | 118.56 | 136.41 | 124.51 | 142.57 |
| B08 | 1 | 127.28 | 153.44 | 136.62 | 133.67 |
| B09 | 1 | 148.45 | 148.83 | 148.50 | 134.06 |
| B10 | 1 | 120.21 | 124.74 | 121.60 | 130.99 |
| B01 | 2 | 141.43 | 165.54 | 148.32 | 167.96 |
| B02 | 2 | 168.22 | 166.81 | 167.76 | 167.67 |
| B03 | 2 | 152.62 | 168.52 | 158.74 | 168.36 |
| B04 | 2 | 128.56 | 149.14 | 134.66 | 156.11 |
| B05 | 2 | 147.66 | 179.81 | 153.84 | 170.20 |
| B06 | 2 | 151.63 | 169.72 | 158.33 | 169.09 |
| B07 | 2 | 151.16 | 162.71 | 152.54 | 163.09 |
| B08 | 2 | 163.97 | 140.88 | 157.24 | 154.07 |
| B09 | 2 | 160.82 | 151.41 | 158.08 | 161.56 |
| B10 | 2 | 148.17 | 166.81 | 153.39 | 160.54 |
| B01 | 3 | 206.55 | 188.41 | 202.20 | 188.94 |
| B02 | 3 | 195.15 | 206.51 | 196.10 | 193.54 |
| B03 | 3 | 187.49 | 200.63 | 190.62 | 202.02 |
| B04 | 3 | 188.99 | 185.26 | 188.44 | 186.38 |
| B05 | 3 | 182.59 | 191.89 | 184.66 | 186.96 |
| B06 | 3 | 183.71 | 176.15 | 182.72 | 193.14 |
| B07 | 3 | 162.12 | 185.19 | 170.62 | 179.76 |
| B08 | 3 | 193.47 | 190.73 | 192.84 | 186.37 |
| B09 | 3 | 194.68 | 204.43 | 196.18 | 189.14 |
| B10 | 3 | 198.04 | 187.21 | 195.23 | 190.66 |

**Table 8:** Cutpoints Averaged Over Panelists.

| | | | Cutpoints | | | |
|---|---|---|---|---|---|---|
| Round | Group | Level | Collapsed Categories | Average Borderline | Weighted | Cubic Regression |
| 1 | A | Basic | 123.70 | 138.46 | 128.11 | 136.75 |
| | | Proficient | 154.26 | 167.26 | 157.10 | 165.96 |
| | | Advanced | 195.36 | 190.31 | 194.47 | 190.33 |
| 1 | B | Basic | 124.71 | **137.33** | 129.67 | 137.15 |
| | | Proficient | 156.26 | 168.40 | 160.62 | 166.43 |
| | | Advanced | 185.31 | 190.65 | 187.14 | 190.27 |
| 2 | A | Basic | 129.03 | 135.87 | 131.56 | 137.55 |
| | | Proficient | 151.88 | 169.46 | 156.84 | 167.40 |
| | | Advanced | 197.32 | 191.62 | 196.51 | 192.94 |
| 2 | B | Basic | 127.57 | **138.93** | 131.48 | 135.73 |
| | | Proficient | 151.42 | 162.13 | 154.29 | 163.86 |
| | | Advanced | 189.28 | 191.64 | 189.96 | 189.69 |

Note. Bold-faced numbers were averages from less than ten judges.

**Table 9**: Classification Pattern of Panelists on Pre-sorted Booklets
Round 2, Groups A and B

| Pre-sorted Booklet Order | Classification Outcome (Frequency) of 20 Raters on Two Forms | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|
| | Below Basic | Borderline Basic | Basic | Borderline Proficient | Proficient | Borderline Advanced | Advanced | |
| 1 | 28 | 5 | 4 | 2 | 1 | 0 | 0 | 40 |
| 2 | 24 | 8 | 6 | 2 | 0 | 0 | 0 | 40 |
| 3 | 19 | 14 | 5 | 1 | 1 | 0 | 0 | 40 |
| 4 | 14 | 16 | 6 | 3 | 1 | 0 | 0 | 40 |
| 5 | 6 | 12 | 12 | 6 | 3 | 1 | 0 | 40 |
| 6 | 2 | 10 | 14 | 9 | 4 | 1 | 0 | 40 |
| 7 | 7 | 9 | 12 | 8 | 3 | 1 | 0 | 40 |
| 8 | 6 | 7 | 13 | 5 | 7 | 2 | 0 | 40 |
| 9 | 1 | 4 | 11 | 13 | 10 | 0 | 1 | 40 |
| 10 | 3 | 2 | 10 | 12 | 11 | 2 | 0 | 40 |
| 11 | 1 | 2 | 9 | 13 | 9 | 6 | 0 | 40 |
| 12 | 1 | 2 | 8 | 14 | 11 | 2 | 2 | 40 |
| 13 | 0 | 1 | 8 | 16 | 12 | 3 | 0 | 40 |
| 14 | 0 | 0 | 5 | 15 | 16 | 2 | 2 | 40 |
| 15 | 0 | 0 | 2 | 8 | 17 | 10 | 3 | 40 |
| 16 | 0 | 0 | 1 | 7 | 17 | 10 | 5 | 40 |
| 17 | 0 | 0 | 4 | 6 | 12 | 11 | 7 | 40 |
| 18 | 0 | 1 | 2 | 6 | 13 | 12 | 6 | 40 |
| 19 | 0 | 0 | 0 | 2 | 12 | 15 | 11 | 40 |
| 20 | 0 | 0 | 1 | 2 | 11 | 11 | 15 | 40 |
| % | 14.0 | 11.6 | 16.6 | 18.8 | 21.4 | 11.1 | 6.5 | 800 |

Note. Frequencies less than 2 are shaded.

**Table 10**: Conformity of Panelists' Round Two Booklet Classifications with Pre-sorted Order on Both Forms.

| Pre-sorted Booklet Order | A10 | | A9 | | A8 | | A7 | | A6 | | A5 | | A4 | | A3 | | A2 | | A1 | | B10 | | B9 | | B8 | | B7 | | B6 | | B5 | | B4 | | B3 | | B2 | | B1 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 |
| 1 | 1 | 2 | 3 | 1 | 1 | 3 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 1 | 1 | 5 | 1 | 1 | 1 | 1 | 2 | 4 | 1 | 1 | 3 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 1 | 1 | 1 |
| 2 | 1 | 1 | 3 | 1 | 2 | 3 | 3 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 4 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 4 | 3 | 3 | 2 | 1 | 2 | 1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| 3 | 1 | 2 | 1 | 1 | 1 | 3 | 1 | 1 | 3 | 1 | 2 | 1 | 2 | 2 | 2 | 5 | 1 | 2 | 1 | 3 | 2 | 2 | 1 | 1 | 1 | 4 | 2 | 3 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 3 | 1 | 2 | 2 | 2 |
| 4 | 1 | 2 | 3 | 2 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 1 | 5 | 3 | 1 | 2 | 3 | 2 | 1 | 2 | 1 | 3 | 4 | 1 | 3 | 4 | 1 | 2 | 1 | 2 | 2 | 2 | 4 | 3 | 1 | 1 | 2 | 2 | 2 | 2 |
| 5 | 2 | 3 | 3 | 1 | 2 | 2 | 4 | 2 | 4 | 3 | 2 | 1 | 3 | 3 | 5 | 3 | 1 | 6 | 2 | 1 | 2 | 4 | 5 | 1 | 4 | 2 | 2 | 2 | 3 | 3 | 1 | 2 | 5 | 4 | 3 | 3 | 3 | 4 | 3 | 2 |
| 6 | 2 | 4 | 3 | 2 | 3 | 4 | 4 | 2 | 2 | 3 | 2 | 4 | 4 | 2 | 5 | 3 | 6 | 5 | 3 | 1 | 1 | 3 | 4 | 2 | 3 | 4 | 3 | 5 | 3 | 2 | 2 | 3 | 4 | 2 | 3 | 3 | 3 | 5 | 4 | 3 |
| 7 | 3 | 6 | 1 | 4 | 2 | 3 | 3 | 1 | 1 | 4 | 3 | 4 | 3 | 5 | 4 | 2 | 3 | 3 | 3 | 2 | 3 | 5 | 1 | 3 | 1 | 2 | 3 | 1 | 2 | 4 | 2 | 2 | 4 | 5 | 3 | 1 | 2 | 4 | 4 | 2 |
| 8 | 4 | 6 | 3 | 3 | 1 | 3 | 4 | 1 | 4 | 5 | 6 | 5 | 2 | 2 | 3 | 2 | 1 | 5 | 3 | 1 | 3 | 5 | 3 | 3 | 1 | 2 | 4 | 2 | 3 | 4 | 5 | 5 | 1 | 2 | 3 | 3 | 2 | 3 | 5 | 3 |
| 9 | 5 | 5 | 4 | 3 | 3 | 4 | 4 | 3 | 2 | 4 | 5 | 4 | 4 | 3 | 5 | 3 | 5 | 5 | 3 | 1 | 5 | 4 | 7 | 2 | 2 | 3 | 4 | 5 | 3 | 4 | 3 | 5 | 5 | 3 | 4 | 4 | 3 | 4 | 4 | 2 |
| 10 | 4 | 5 | 3 | 5 | 1 | 4 | 4 | 3 | 3 | 4 | 5 | 2 | 1 | 4 | 5 | 5 | 5 | 5 | 3 | 3 | 6 | 4 | 3 | 4 | 2 | 6 | 5 | 3 | 4 | 4 | 4 | 5 | 1 | 5 | 3 | 4 | 3 | 5 | 4 | 3 |
| 11 | 4 | 6 | 5 | 4 | 4 | 4 | 5 | 5 | 4 | 4 | 3 | 1 | 5 | 3 | 5 | 4 | 3 | 5 | 3 | 3 | 5 | 6 | 5 | 4 | 4 | 5 | 6 | 6 | 6 | 3 | 3 | 2 | 3 | 2 | 4 | 4 | 3 | 4 | 6 | 4 |
| 12 | 3 | 4 | 4 | 4 | 5 | 5 | 4 | 4 | 2 | 5 | 3 | 2 | 3 | 4 | 5 | 6 | 5 | 5 | 4 | 3 | 4 | 4 | 5 | 4 | 5 | 7 | 3 | 7 | 3 | 5 | 3 | 1 | 4 | 5 | 3 | 5 | 4 | 4 | 6 | 4 |
| 13 | 4 | 3 | 3 | 5 | 5 | 4 | 5 | 4 | 5 | 4 | 5 | 3 | 4 | 4 | 5 | 3 | 3 | 4 | 3 | 4 | 4 | 5 | 4 | 5 | 6 | 5 | 4 | 2 | 6 | 4 | 4 | 3 | 5 | 6 | 4 | 3 | 4 | 4 | 5 | 5 |
| 14 | 4 | 4 | 4 | 4 | 5 | 3 | 7 | 5 | 3 | 5 | 4 | 5 | 4 | 4 | 5 | 5 | 5 | 5 | 4 | 3 | 5 | 5 | 5 | 5 | 5 | 5 | 3 | 6 | 4 | 4 | 3 | 4 | 5 | 6 | 4 | 4 | 4 | 5 | 7 | 4 |
| 15 | 5 | 4 | 6 | 5 | 6 | 5 | 5 | 4 | 5 | 6 | 5 | 5 | 4 | 5 | 6 | 4 | 6 | 6 | 3 | 4 | 6 | 6 | 5 | 5 | 7 | 5 | 3 | 7 | 4 | 6 | 5 | 6 | 4 | 7 | 4 | 5 | 5 | 5 | 5 | 5 |
| 16 | 4 | 3 | 5 | 5 | 4 | 6 | 6 | 6 | 5 | 5 | 6 | 5 | 5 | 5 | 6 | 4 | 7 | 5 | 4 | 5 | 5 | 4 | 6 | 5 | 6 | 5 | 6 | 7 | 7 | 5 | 7 | 6 | 5 | 6 | 5 | 4 | 5 | 4 | 5 | 7 |
| 17 | 6 | 3 | 6 | 5 | 4 | 6 | 7 | 7 | 4 | 6 | 4 | 4 | 5 | 5 | 5 | 5 | 6 | 6 | 3 | 3 | 6 | 5 | 7 | 5 | 4 | 5 | 7 | 6 | 5 | 7 | 3 | 6 | 5 | 7 | 5 | 6 | 7 | 5 | 4 | 6 |
| 18 | 5 | 2 | 6 | 6 | 5 | 7 | 5 | 4 | 5 | 7 | 6 | 3 | 5 | 5 | 6 | 5 | 6 | 5 | 4 | 6 | 7 | 5 | 7 | 5 | 6 | 7 | 3 | 5 | 4 | 7 | 4 | 4 | 6 | 4 | 6 | 6 | 6 | 5 | 5 | 6 |
| 19 | 6 | 6 | 5 | 7 | 5 | 5 | 7 | 5 | 5 | 6 | 7 | 4 | 7 | 6 | 5 | 5 | 6 | 6 | 4 | 6 | 6 | 6 | 7 | 6 | 6 | 7 | 6 | 7 | 5 | 5 | 7 | 7 | 7 | 7 | 6 | 5 | 5 | 6 | 5 | 6 |
| 20 | 7 | 4 | 5 | 5 | 5 | 4 | 6 | 6 | 5 | 6 | 7 | 7 | 5 | 5 | 7 | 5 | 6 | 7 | 5 | 7 | 7 | 7 | 6 | 6 | 7 | 6 | 6 | 3 | 7 | 7 | 6 | 5 | 5 | 5 | 6 | 7 | 7 | 7 | 7 | 6 |

F1 = Form 1
F2 = Form 2

**Table 11**: Average Round 2 Cutpoints Across Panelists using Alternative Proficiency Estimates.

| Proficiency Estimate | Proficiency Level | Booklet Classification Outcomes | | | | | |
|---|---|---|---|---|---|---|---|
| | | Group A | | | Group B | | |
| | | Cutpoint | std. dev. | %>= | Cutpoint | std. dev. | %>= |
| ML | Basic | 131.56 | 8.19 | 94.9 | 131.48 | 7.50 | 94.9 |
| | Proficient | 156.84 | 11.07 | 44.8 | 154.29 | 8.60 | 52.3 |
| | Advanced | 196.51 | 9.73 | 0.1 | 189.96 | 8.94 | 0.4 |
| PV1 | Basic | 140.58 | 5.85 | 84.4 | 139.54 | 5.47 | 86.1 |
| | Proficient | 152.88 | 7.95 | 56.4 | 155.59 | 9.47 | 49.1 |
| | Advanced | 176.14 | 5.99 | 6.2 | 170.72 | 5.64 | 13.1 |
| PV2 | Basic | 144.31 | 5.25 | 77.7 | 140.31 | 5.73 | 85.0 |
| | Proficient | 157.75 | 6.95 | 42.8 | 155.58 | 8.46 | 49.1 |
| | Advanced | 176.11 | 4.31 | 6.2 | 174.45 | 3.26 | 8.2 |
| PV3 | Basic | 141.18 | 4.54 | 83.9 | 141.37 | 4.27 | 83.3 |
| | Proficient | 153.41 | 9.25 | 55.5 | 154.92 | 8.52 | 51.3 |
| | Advanced | 177.53 | 6.10 | 5.0 | 175.42 | 7.47 | 7.0 |
| PV4 | Basic | 139.69 | 3.34 | 86.1 | 141.61 | 6.67 | 83.3 |
| | Proficient | 155.72 | 7.90 | 47.9 | 153.68 | 12.16 | 54.5 |
| | Advanced | 176.47 | 5.21 | 5.8 | 171.59 | 4.50 | 11.9 |
| PV5 | Basic | 142.13 | 6.66 | 82.0 | 142.75 | 5.94 | 81.3 |
| | Proficient | 157.51 | 7.45 | 43.8 | 157.47 | 7.47 | 43.8 |
| | Advanced | 174.33 | 5.28 | 8.2 | 171.62 | 7.00 | 11.5 |
| Mean PV | Basic | 141.58 | - | 83.3 | 141.12 | - | 83.9 |
| | Proficient | 155.45 | - | 49.1 | 155.45 | - | 49.1 |
| | Advanced | 176.11 | - | 6.2 | 172.76 | - | 9.9 |

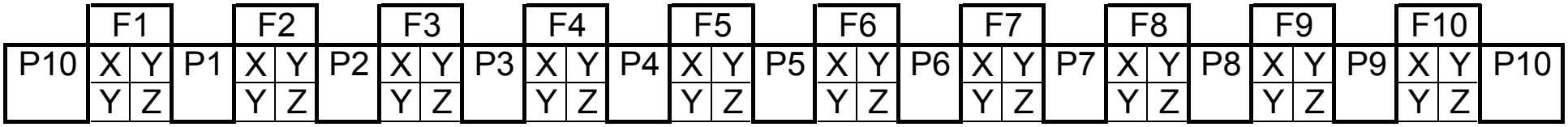


**Figure 1**: Form and Panelist Map

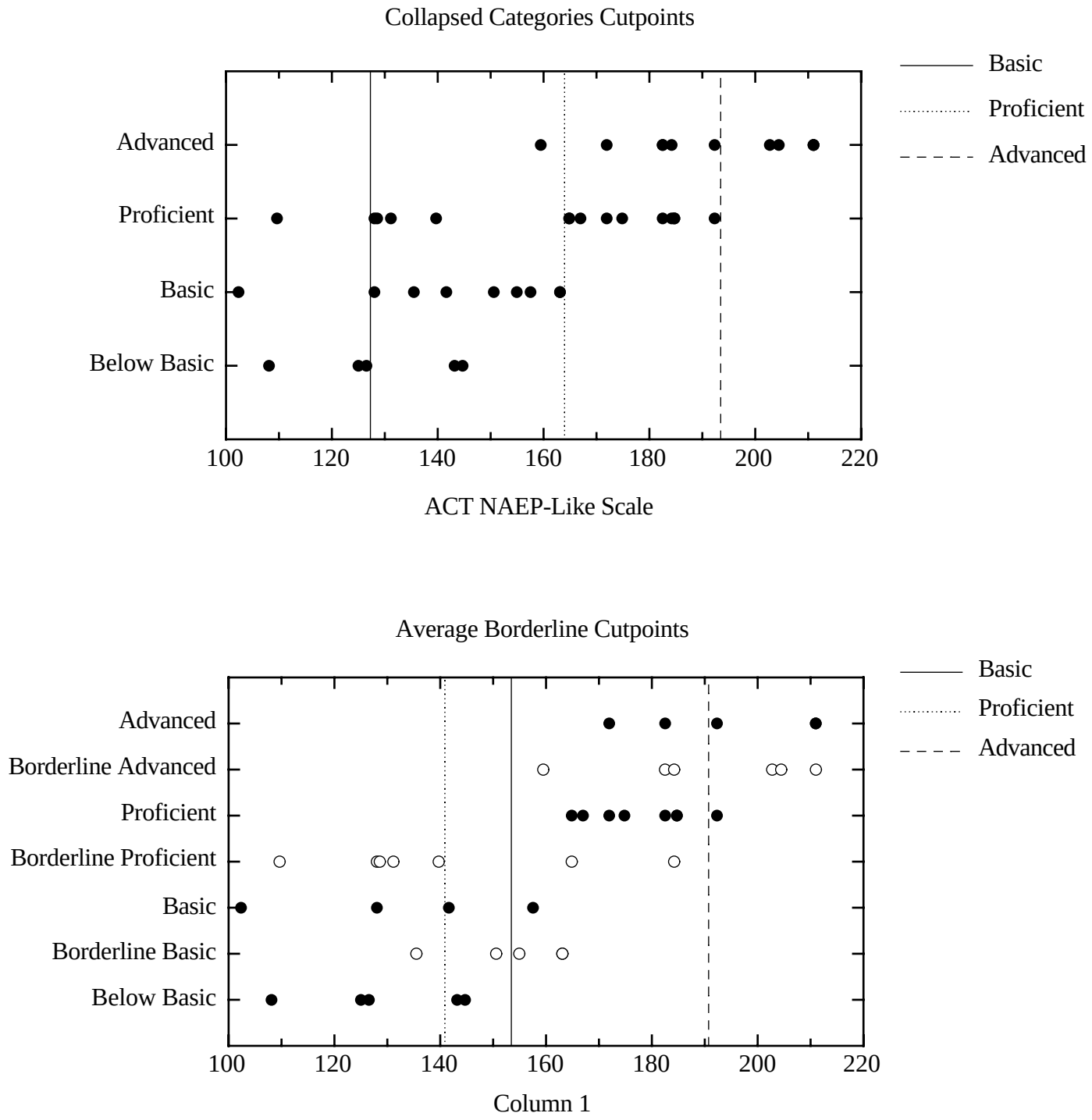


**Figure 2**: Round 2 Cutpoints for Panelist B08.

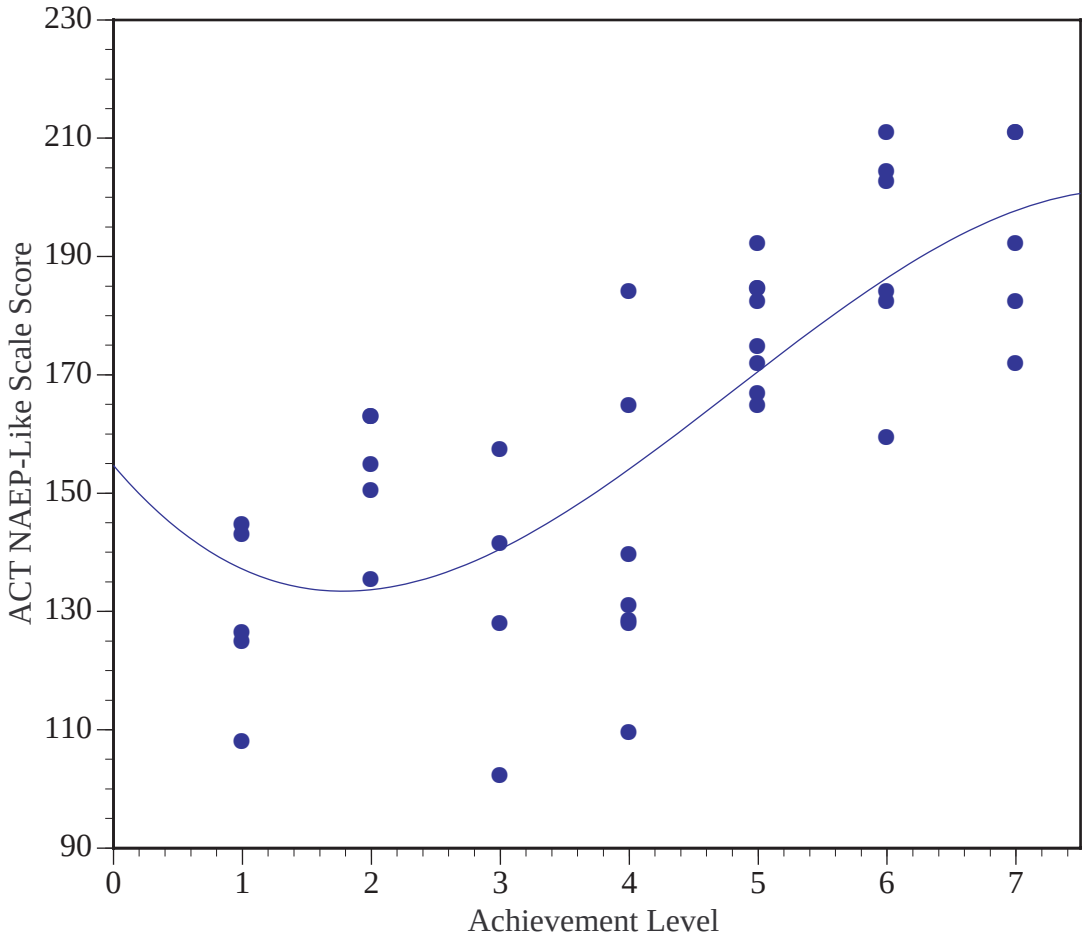


**Figure 3:** Level Assigned by Rater B08 on Round 2 Versus Estimated ACT NAEP-Like Scale Score for Booklets.