DOCUMENT RESUME

ED 389 736 TM 024 365

AUTHOR Reckase, Mark D.; And Others
TITLE Setting Standards on NAEP Performance Items.
PUB DATE Apr 95
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Educational Assessment; Elementary Secondary Education; Evaluation Methods; Geography; *Interrater Reliability; *Research Methodology; Scores; *Scoring; Selection; *Standards; *Test Items; United States History
IDENTIFIERS Dichotomous Scoring; *National Assessment of Educational Progress; *Performance Based Evaluation; Polytomous Variables; Standard Setting

ABSTRACT

The research reported in this paper was conducted to gain information to guide the selection of standard setting procedures for use with polytomous items to set achievement levels on the National Assessment of Educational Progress (NAEP) assessments in U.S. History and geography. Standard-setting procedures were evaluated to determine the relative level of the standards, the relationships of the level of the standard set by each method to the standard set using dichotomously scored items, and the practicality of the procedure for operational standard setting. The methods explored were: (1) estimated score point percentage (P method); (2) modified percentage estimate (PGE2); (3) estimated mean score method (M method); and (4) the hybrid method (H method), a combination of the previously used paper-selection method and the M method. Data were from the full NAEP pool for each content area with 20 panelists at each level for grades 4, 8, and 12. The P method tended to set the highest standard overall, and the PGE2 method, the lowest overall. The M method, which set standards between P and PGE2, but above the H method when they could be compared directly, was selected as the operational method for achievement level setting, with raters trained in paper-selection as a first step. (Contains two figures, seven tables, and nine references.) (SLD)

Setting Standards on NAEP Performance Items[1]

Mark D. Reckase, Susan Cooper Loomis, Tianyou Wang,
and Luz Bay

American College Testing (ACT)

While there is a wealth of information about methods for setting standards on educational tests in the literature (see Jaeger (1989) or Berk (1986), for a summary), there is relatively little information about the setting of standards on tests composed of tasks that are not scored dichotomously (i.e., correct or incorrect). Most of the literature that does exist was produced in response to the need for the National Assessment Governing Board (NAGB) to set performance standards on the National Assessment of Educational Progress (NAEP). For example, ACT proposed a method called the paper-selection method for the setting of achievement levels on the NAEP in Mathematics, Reading, and Writing Assessments (ACT, 1992). That method required standard setting panelists to identify actual examples of students' work that matched their conception of how the minimally qualified student would need to perform to be considered to have met the standard. Standards were set using the scores assigned through the NAEP scoring process to the selected papers.

That paper-selection method was found to be easily impiemented for the NAEP Mathematics Assessment, a test with a small number of performance tasks, but it was also found to be very time consuming when large numbers of performance tasks were included on a test. such as was the case for the reading and writing assessments.

[1] Paper presented at the national meeting of the American Educational Research Association and the National Council of Measurement in Education, San Francisco, April 1995.

Further, the standard set using the paper-selection method was found to be consistently higher than that set using the modified Angoff method for multiple-choice items, raising a number of questions about the technical characteristics of the procedure (National Academy of Education, 1993).

Since the initial experience with NAEP polytomously-scored items, some additional work has been published on setting standards using performance assessment tasks (Hambleton & Plake, 1995; Reckase, 1994). But that work, while providing some foundation for addressing the issues, did not lead to any conclusive recommendations about procedures for setting standards on tests composed of polytomously-scored performance assessment tasks. Of course, there are many procedures that use the total scores on tests for standard setting, rather than information on individual tasks. Such procedures can be applied to tests composed of polytomously-scored tasks just as they can to tests composed of dichotomously-scored items. Those procedures will not be addressed here since they are beyond the scope of this research.

The motivation for the research reported in this paper was to gain information that would guide the selection of a standard setting procedure for use with polytomous items to be used to set achievement levels on the NAEP assessments in U. S. History and geography. Both of these assessments contain a substantial number of performance assessment tasks. Specifically, a number of procedures for setting standards were evaluated to determine: (1) the relative level of the standards that they set; (2) the relationships of the level of standard set by each method to the standard set using dichotomously-scored items; and (3) the practicality of the procedure for operational standard setting.

Method

ACT staff and the Technical Advisory Committee on Standard Setting worked jointly to devise a number of methods for setting standards using polytomously-scored test tasks that would overcome the practical problem encountered with the paper-selection method. That problem was that panelists had to read through large numbers of papers during each round of the process to select the papers. The paper readings were found to take a very long time and the process required the management of large amounts of paper. While the procedure had strong face validity and the panelists were able to perform the task, more efficient procedures were desired. There was also some concern about the difference in levels set by the dichotomous and polytomous procedures. The four methods described below were the result of efforts to devise more efficient procedures for the polytomous standard setting process.

Proposed Standard-Setting Procedures

Estimated Score Point Percentage Method. For this method, panelists are asked to estimate the score distribution for each polytomously scored test item for 100 persons that just meet the qualifications for passing. Thus, if the performance task is scored using a four-point rubric, the panelists are asked to specify the number out of 100 individuals that will receive scores of 1, 2, 3, or 4 if all of the individuals are from the minimally qualified group. The frequencies assigned to each score category should, of course, add to 100.

For the purposes of this paper, this method will be referred to as the **Percentage** method and abbreviated as the **P**-method.

Modified Percentage Estimate Method. Because there was a concern that estimating all of the frequencies for the distribution of scores on an assessment task might be too difficult, a simplified procedure was proposed. This method required the panelists to estimate the number out of 100 examinees who would receive a rating of 2 or greater for the performance task based on the group of examinees that was just above the qualifying level. A score of 2 or better was selected because a rating of 2 on the general rubric was believed to indicate an acceptable level of competence.

If panelists were using both this method and the **P**-method, to be consistent, the frequency of scores estimated for 2 or better should be equal to the sum of the frequencies in categories 2 and higher for the P-method. That is, if the task is scored on a 4-point rubric, this method would result in estimates of the sum of frequencies assigned to 2, 3, and 4. For this paper, the **Modified Percentage Estimate Method** will be abbreviated the **PGE2** (percentage greater than or equal to 2).

Both the **P** and the **PGE2** methods can be considered as generalizations of the Angoff method. For that method (Angoff, 1971), panelists are asked to estimate the percentage of minimally qualified examinees who will get a dichotomously scored item correct. The percentage incorrect can be obtained by subtracting the percent correct from 100. Thus the Angoff method asks for the score distribution on dichotomous items as does method **P**. When polytomous items are dichotomized, the estimate of the percentage above the point of dichotomy (method **PGE2**) is also similar to the Angoff method.

Estimated Mean Score Method. For the Estimated Mean Score Method, panelists were asked to estimate the arithmetic average (mean) for the performance task for the group that would just meet the qualification level. The mean was estimated on the score point metric defined by the scoring rubric. In this paper, this method will be called the **Mean Method** and will be abbreviated as **M** method.

Hybrid Method. The paper-selection method was believed to have high face validity and to have built-in reality checks in that panelists had to read actual student papers. There was a desire to have a method that used the best parts of the paper-selection method, but that was more efficient. The Hybrid Method was an attempt to achieve that goal. For this method, the first round of ratings was done using the Paper-Selection Method. The panelists were asked to select three papers that matched the criteria of being just above the level of performance needed to be acceptable. After the first round, the mean score on the three papers each panelist selected was computed and returned to each panelist. They then used the Mean Method to adjust the estimated mean in subsequent rounds of standard setting. Since both the paper-selection method and mean method are used in this procedure, it was labeled the **Hybrid Method**. It is abbreviated as the **H**-method for this paper.

Implementation of Methods

The four standard-setting methods were implemented in two pilot studies, one using the U. S. History NAEP and the second using the Geography NAEP. The first pilot study used the Geography NAEP. The **H**, **M**, and **P** methods were used in that study. It was conducted from July 14 to 18, 1994. The second pilot study used U. S. History NAEP and the **H**, **M**, and **PGE2** methods. This study was conducted from August 11 to 15, 1994.

For each of the studies, panelists were selected using the two-stage sampling method described in ACT (1994). The first stage of this model consisted of sampling school districts using a stratified random sampling plan. For each district that was sampled, nominators were identified that held specific positions (school board president, mayor, superintendent). The nominators were asked to nominate as standard-setting panelists individuals who were knowledgeable about the subject matter and the capabilities of students at the target grade level.

Nominated individuals were screened to insure that they met the necessary qualifications and to insure a representative panel with respect to demographic and geographic classifications. Panelists were selected from among those who met the requirements and they were invited to participate. To the extent possible, panelists were balanced on race/ethnicity, gender, type of school, district size, community type, regions, and whether they were a teacher, non-teacher educator, or a member of the general public.

Achievement levels were set at grades 4, 8, and 12. The goal was to have 20 panelists at each grade level for each content area. The actual numbers of panelists that participated are given in Table 1.

Insert Table 1 about here

The standard-setting meetings were approximately five days in length. The first two days were dedicated to informing the panelists about the purposes of the studies, the NAEP content frameworks and tests, and the standard-setting process. The rest of the time was devoted to three rounds of standard setting with feedback of various types between rounds one and two, and rounds two and three. In addition, the panelists were asked to evaluate the achievement levels-setting process.

The pilot study had the dual purposes of evaluating methods to be used with the performance assessment tasks, and refining the procedures to be used with the dichotomously-scored items. Therefore, the full NAEP item pool for the appropriate content area was used for each study, and the achievement levels set were based on both the dichotomously- and polytomously-scored items.

Data Collection Design.

To evaluate several standard-setting methods during the two pilot studies and to minimize the complexity of the panelists task, a balanced incomplete blocks (BIB) design was used for each pilot study. Each group of grade-level panelists was divided into matched halves, and standard setting procedures were assigned to the halves in a balanced way. The number of panelists assigned to each procedure by grade level and content area is given in Table 2.

Insert Table 2 about here

At each round of the standard-setting process, the panelists were asked to indicate how examinees who were just inside the lower bound of an achievement level category would perform on the tasks. Three achievement level categories were used: Basic, Proficient, and Advanced. The panelists received extensive training in the meaning of the achievement levels before any standard-setting activities were performed. The definitions of the three achievement levels are given in ACT (1994). At each round, panelists were asked to provide estimates of performance for the lower bound of each level -- that is, three ratings for each performance task.

After each round of the standard-setting process, the ratings of the panelists were averaged to arrive at an individual standard for the lower bound of each achievement level, and the individual results were averaged for each cell of the design to get a group standard. These averages were mapped to the NAEP IRT scale using the functional relationship specified by the generalized partial credit model (Muraki, 1992). Using this procedure, a standard was determined for each individual, and group standards were also available. The standards for each individual were analyzed using ANOVA procedures to determine if there were significant differences in the levels set by the different procedures. Since the BIB design was used, only main effects could be tested for significance.

Results

Polytomous Standards

The results for the various polytomous-standard setting procedures are presented in Tables 3, 4, and 5 using a pseudo NAEP scale that was created for reporting the results to the panelists. The scale is a linear transformation of the NAEP theta scale. However, since the actual NAEP results for the areas of geography and U. S. History have not yet been made public, the details of the transformation can not be presented here.

The ANOVA performed on the standards set by the panelists yielded a significant main effects for method for both the geography and the U. S. History analyses. The main effect for grade was not significant for the final round of ratings. The grade effect is not particularly interesting because the IRT calibration was done separately by grade. No effect was expected, however, the lack of an effect supported summarizing the results by collapsing over grade levels.

Inset Tables 4, 5, and 6 about here

The overall results show that the four different standard-setting methods resulted in standards at somewhat different levels. The Percentage (P) Method tended to set the highest standard overall. Only for U. S. History at the Advanced level for 4th grade does the Hybrid (**H**) Method set a higher standard. The judges' ratings for the 4th grade Hybrid Method at the Advanced level had an unusually large standard deviation, suggesting that the ratings at that level had an unusual distribution.

The **PGE2** Method resulted in the lowest standards overall. The standards set using that procedure were notably lower than those set by the other methods. The Hybrid Method and the Mean Method resulted in standards that were roughly the same. Those two methods were closely related in that the Hybrid Method used the Mean Method for the second and third rounds of ratings. Note, however, that the Mean Method had larger standard deviations for the panelists' individual standards than the Hybrid Method.

Since the Mean Method resulted in higher standards than the Hybrid Method in the one case where they can be directly compared (12th grade Geography), the results suggest the following ranking of the standards that might be expected from these methods if the results replicate in other situations: **P** > **M** > **H** > **PGE2**.

Comparison to Dichotomous Results

Previous experience with achievement level-setting using both dichotomous and polytomous items resulted in standards based on polytomous items that were set higher than those based on dichotomous items (ACT, 1992; National Academy of Education, 1993). Therefore, the relationship between the achievement levels set using dichotomous items and those using polytomous items was of interest.

The results of the comparison of the achievement levels set using dichotomous and polytomous items are given in Table 6. The table presents the mean difference computed from the polytomous achievement level minus the dichotomous achievement level for each panelist. All panelists set standards using both item types. The modified Angoff method was used to set the achievement levels for the dichotomous items as described in ACT (1994).

Insert Table 6 about here

The results for the comparison of the achievement levels set using the different types of items are not easily interpreted. For geography, the achievement levels set using the two different item types were fairly similar. For U. S. History, differences were sometimes fairly great. There also seemed to be an interaction between the achievement levels set using the dichotomous items and the method used for the polytomous items. For example, the achievement levels set using the dichotomous items were lower when the polytomous method was the Hybrid Method rather than the Percentage Method. The dichotomous-based achievement levels for the panelists using each of those polytomous methods are given in Table 7.

Insert Table 7 about here

Panelists Reactions

After each round of the achievement level-setting process, the panelists were asked to evaluate the process they used as to whether it was conceptually clear and easy to apply (5 = totally agree, 1 = totally disagree). The results for the geography and U. S. History pilot studies are presented for each round in Figures 1 and 2.

Insert Figures 1 and 2 about here

For the Geography pilot, there was little difference in the evaluation of the panelists of the various methods by Round 3. The Hybrid Method was rated harder to apply initially because the panelists had to read through a large number of papers in the first round. The Hybrid Method received a generally lower rating than the other methods from the U. S. History panelists. Again, this was probably because of the effort required to select papers in the first round of the achievement levels-setting process.

Discussion

Setting standards is a judgmental process that yields a numerical result, with no well defined correct answer that can be used as a criterion of accuracy. At best, the results of the application of a method can be supported as being consistent and reasonable. However, if there is a "truth" in the standard setting literature, it is that different methods yield different results. The results of these pilot studies are consistent with the general "truth" from that literature. As expected, the different methods yield different standards. However, the purpose for the pilot studies was to determine if there were anything in the results that tended to support or to eliminate from consideration any of the methods.

One anecdotal observation related to the Percentage Method was that panelists often submitted ratings that did not make sense given the rules of the application of the method. For example, sometimes the percentages for each score category did not add to 100. Or, when they were trying to raise their standard, the panelists did not know how to change the estimates of percentages to get the results they wanted. Despite the fact that the panelists rated the procedure as fairly easy to apply, the data analysis staff noted a number of problems with the data when the ratings were being converted to points on the NAEP theta scale. These problems might suggest that this method is too complex for general application.

The **PGE2** method resulted in much lower standards than the other procedures. While there is nothing that can logically be used to indicate that these achievement levels are incorrect, they are inconsistent with those provided by the other procedures.

If these observations cast doubts about the value of the Percentage and **PGE2** Methods, the Mean Method and the Hybrid Method are still available for consideration. There is some evidence that the Mean Method may set somewhat higher standards than the Hybrid Method. This may be because the Hybrid Method requires that panelists read many student papers, giving them a stronger "reality check" than the Mean Method. This is further suggested by the lower achievement levels set using the dichotomous items when the Hybrid Method was used for the polytomous items.

However, the Hybrid Method is more difficult to apply because of the need to read many papers.

The project staff selected the Mean Method for operational use for the achievement levels-setting process after consultation with the Technical Advisory Committee. Based on the belief that the experience gained from reading the papers better informed the judgements made by the panelists and based on the observation that panelists needed more focus on borderline performance, panelists were trained on the paper-selection process before the first round of ratings. Thus, before rating items, panelists reviewed several student responses to each polytomously scored item. Further results about the application of this method will be reported once data from the operational achievement levels setting studies are fully analyzed.

References

American College Testing (1992, April). Description of mathematics achievement levels-setting process and proposed achievement level definitions. Iowa City, IA: Author.

American College Testing (1994, April). *Design document: Setting achievement levels on the 1994 National Assessment of Educational Progress in Geography and U.S. History and the 1996 National Assessment of Educational Progress in Science.* Iowa City, IA: Author.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed). Washington, DC: american Council on Education.

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

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

Jaeger, Richard M. (1989). Certification of student competence. In Robert L. Linn (Ed.) *Educational measurement, 3rd edition.* New York: American Council on Education/Macmillan.

Muraki, Eiji (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159-176.

National Academy of Education (1993). *Setting performance standards for student achievement -- A report of the National Academy of Education Panel on the on the Evaluation of the NAEP Trial State Assessment: An evaluation of the 1992 achievement levels.* Stanford, CA: Author.

Reckase, Mark D. (1994, June). Standard setting on performance assessments: a comparison between the paper selection method and the contrasting groups method. Paper presented at the National Conference on Large Scale Assessment, Albuquerque, NM.

Table 1

Number of Panelists Used in Pilot Studies

	Study	
Grade	Geography	U.S. History
4	17	20
8	20	19
12	18	21

Table 2

Balanced Incomplete Block Design for Each Content Area

		Method		
Content Area	Grade	P	H	M
Geography	4	9	8	*
	8	10	*	10
	12	*	9	9
		P	H	PGE2
U.S. History	4	10	10	*
	8	10	*	9
	12	*	10	11

* No panelists at this grade level used the method indicated by this column.

Table 3

Mean Achievement Levels Set by Each Method for Each Grade

Geography

		Method		
Grade	Achievement Level	P	H	M
4	Basic	158.7	150.6	—
	Proficient	173.0	164.2	—
	Advanced	180.4	176.5	—
8	Basic	156.7	—	144.1
	Proficient	170.7	—	157.8
	Advanced	181.2	—	172.5
12	Basic	—	149.4	155.5
	Proficient	—	163.5	168.4
	Advanced	—	177.5	183.5

Table 4

Mean Achievement Levels Set by Each Method for Each Grade

U.S. History

		Method		
Grade	Achievement Level	P	H	PGE2
4	Basic	162.6	152.4	—
	Proficient	169.9	166.8	—
	Advanced	177.1	187.9	—
8	Basic	157.4	—	140.9
	Proficient	171.5	—	156.7
	Advanced	180.5	—	169.8
12	Basic	—	160.6	134.9
	Proficient	—	172.4	156.1
	Advanced	—	186.3	171.4

Table 5

Mean Achievement Levels Set by Each Method Collapsed Over Grade

Content Area	Achievement Level	Method		
		P	H	M
Geography	Basic	157.8(8.8)	149.9(5.9)	149.5(9.4)
	Proficient	172.9(10.9)	164.0(5.0)	163.1(8.1)
	Advanced	183.9(12.6)	178.0(7.0)	178.3(8.3)
		P	H	PGE2
U.S. History	Basic	160.2(5.9)	156.4(6.4)	137.5(5.5)
	Proficient	170.9(4.8)	169.9(5.7)	156.3(5.2)
	Advanced	179.5(6.7)	189.6(11.2)	171.6(6.3)

Note: Numbers in parentheses are the standard deviations of the panelists standards.

Table 6

Mean Difference in Achievement Levels Set Using Polytomous and Dichotomous Items Collapsed over Grades

Content Area	Achievement Level	Method		
		P	H	M
Geography	Basic	9.4	2.2	-6.3
	Proficient	2.9	.1	-6.7
	Advanced	.1	-1.1	-7.5
		P	H	PGE2
U.S. History	Basic	36.2	44.7	17.6
	Proficient	7.2	19.1	-3.5
	Advanced	3.5	20.9	-6.1

Note: Mean difference is given by mean polytomous-mean dichotomous.

Table 7

Dichotomous Based Mean Achievement Levels for Groups Using the Hybrid and Proportion Polytomous Methods

Achievement Level	Polytomous Method	
	P	H
Basic	125.1	101.8
Proficient	162.8	150.0
Advanced	173.4	169.6

FIGURE 1
Geography

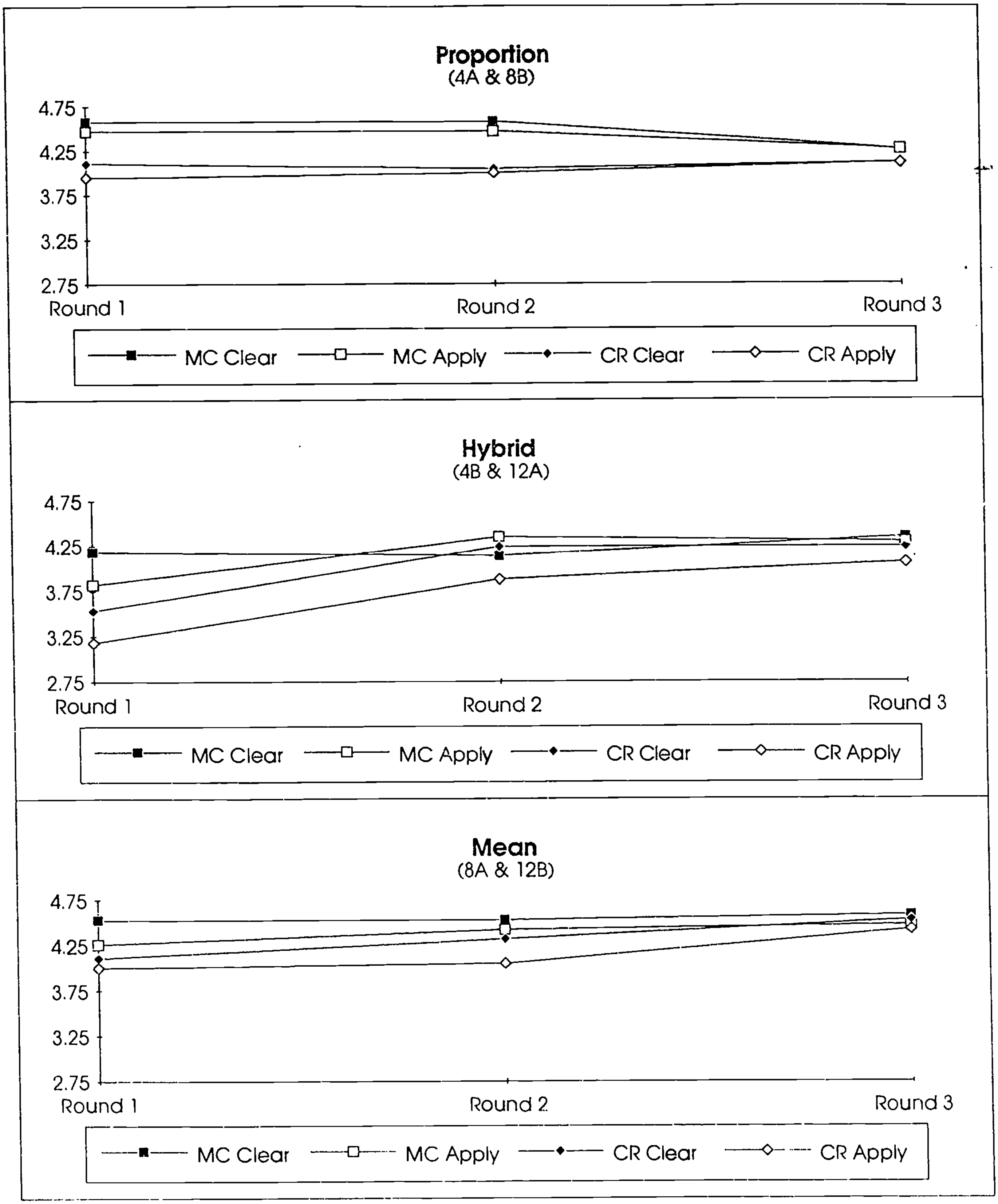
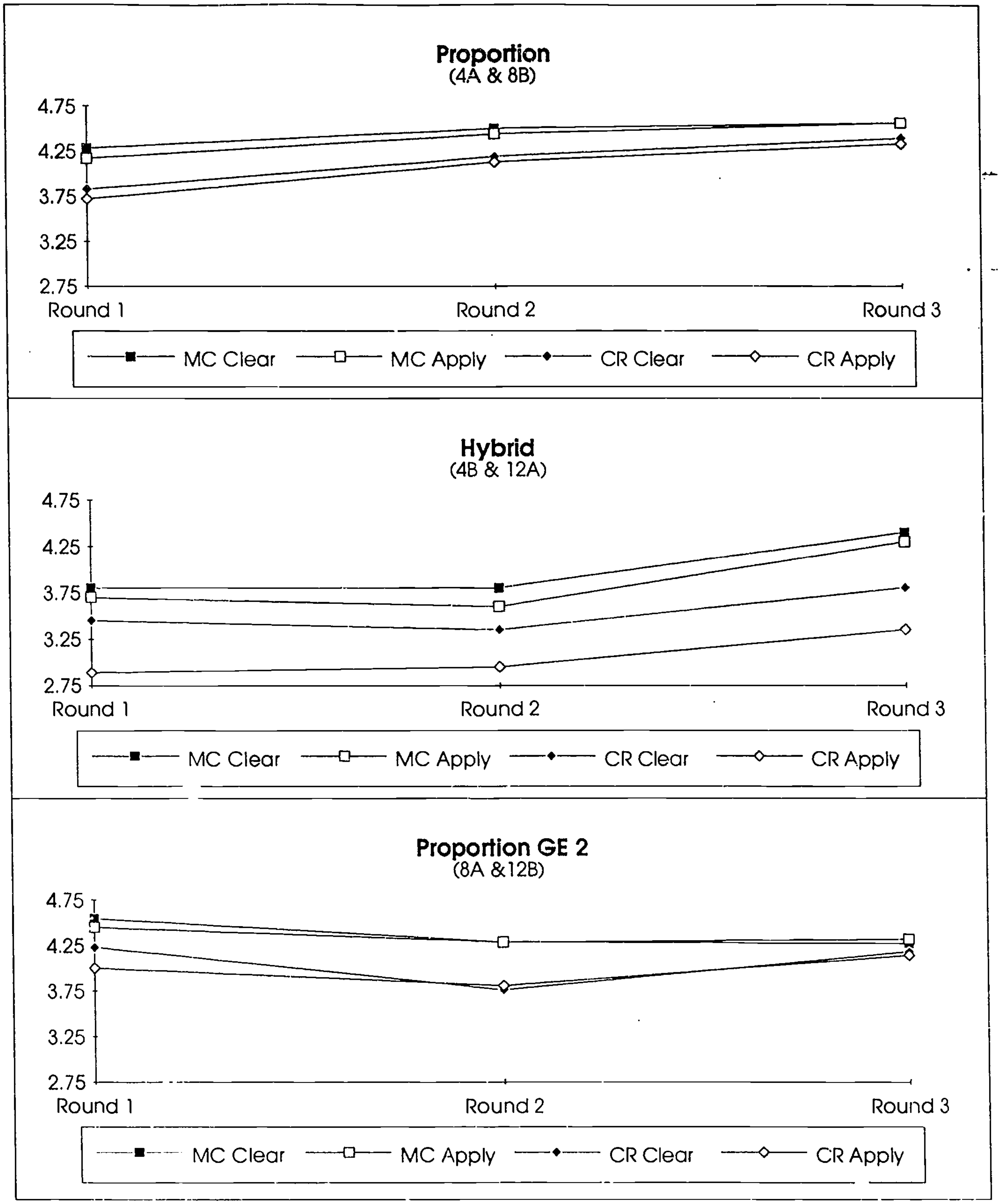


FIGURE 2

U.S. History



Proportion
(4A & 8B)

Hybrid
(4B & 12A)

Proportion GE 2
(8A &12B)