DOCUMENT RESUME

ED 463 746 IR 021 168

AUTHOR Harvey, Anne L.

TITLE Comparing Onsite and Online Standard Setting Methods for Multiple Levels of Standards.

PUB DATE 2000-00-00

NOTE 30p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2000).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Comparative Analysis; *Computer Assisted Testing; Computer Mediated Communication; Computer Oriented Programs; Higher Education; Internet; Online Systems; Standards; *Tests; World Wide Web

IDENTIFIERS College Level Examination Program; *Standard Scores; Standard Setting

ABSTRACT

The Web-based standard setting (WBSS) system described in Harvey and Way (1999) provides an online, Internet-based alternative for testing programs that use onsite judgmental standard setting studies to set their cut scores. Standard setting studies using the WBSS system were compared to similar onsite studies of two exams in the College-Level Examination Program. The online and onsite groups felt similarly about the usefulness of the discussions in making the final ratings and about the overall experience of the study. Results indicated significant differences in the perceptions of group process and working conditions, with the online studies scoring lower on these aspects. Two methods of standard setting were implemented in both the online and onsite modes: an Angoff method and a no/yes method. No significant differences were found in the ratings for the no/yes method. Substantial differences in the average ratings for the Angoff method were found for one of the exams, but not for the other exam. Further research is needed in the aspects of these studies that cause a difference in one case but not the other. (Author/MES)

Comparing Onsite and Online Standard Setting Methods for Multiple Levels of Standards

Anne L. Harvey

Educational Testing Service

Paper presented at the annual meeting of the
National Council on Measurement in Education, New Orleans, April 2000

Comparing Onsite and Online Standard Setting Methods for Multiple Levels of Standards

Abstract

Standard setting studies using the online, web-based standard setting (WBSS) system were compared to similar onsite studies of two exams in the College-Level Examination Program. The online and onsite groups felt similarly about the usefulness of the discussions in making the final ratings and about the overall experience of the study. Results indicated significant differences in the perceptions of group process and working conditions, with the online studies scoring lower on these aspects.

Two methods of standard setting were implemented in both the online and onsite modes: An Angoff method and a no/yes method. No significant differences were found in the ratings for the no/yes method. Substantial differences in the average ratings for the Angoff method were found for one of the exams, but not for the other exam. Further research is needed in the aspects of these studies that cause a difference in one case, but not the other.

Comparing Onsite and Online Standard Setting Methods for Multiple Levels of Standards

Introduction

The web-based standard setting (WBSS) system described in Harvey and Way (1999) provides an online, Internet-based, alternative for testing programs that use onsite judgmental standard setting studies to set their cut scores. Using the WBSS to conduct standard setting studies provides advantages in addition to the saving of travel and housing costs for judges. These advantages include the opportunity to standardize and review training materials, manage paperwork, and produce data files without manual entry or scanning.

There are, however, some potential disadvantages with using an online system that were explored in the pilot study (Harvey & Way, 1999). One potential disadvantage is the quality of the discussion of characteristics of the target group due to judges not being face-to-face. This discussion is intended to reduce the variability of the judgments by providing a common definition of the group for whom the judgments are being made. The pilot study showed no significant differences in effect of the discussion on the variability of ratings for the Angoff[1] method (1971) and benchmark methods (Faggen, 1994). The previous study, however, noted that the judges participating in the online discussion were not as satisfied with the discussion as were judges participating in an onsite discussion. In addition, judges participating in an online system were less likely to feel comfortable asking questions. Based on these results, the prototype system was revised to allow easier navigation of previous portions of the study during the discussion and to improve the general interface.

The first program to consider using the WBSS operationally is the College-Level Examination Program (CLEP). As the CLEP planned for the transition to an all computer-delivered examination program, one of the challenges was the updating of the recommended credit-granting scores (RCGS) for several of the examinations.

At least 14 CLEP examinations will require a new RCGS for the introduction of computer-delivered testing in July 2001. In the past, the CLEP has set a RCGS by testing students in corresponding courses. For example, the RCGS for the Western Civilization I Examination would be set by testing students at the end of a Western Civilization I college course. The average test score for students receiving a grade of C would become the RCGS for that exam. This approach, while scientifically satisfying, has become harder to administer as instructors become more reluctant to give up teaching time to outside activities such as the CLEP exams. In addition, it is often difficult to assess the motivation of students and the representation of colleges when obtained on a volunteer basis over an increasingly long time. The problem of obtaining an adequate sample was

[1] Angoff ascribes the original idea for his method to L. R Tucker, c. 1952.

especially felt for the less popular subject areas, sometimes taking as long as five years to set the RCGS for a new edition of an exam.

The program determined that the traditional method of obtaining the RCGS would not be feasible and chose judgmental standard setting panels as the alternative. A concern, however, was the cost of hosting 14 panels of 15 to 20 judges for a two-day study. The WBSS was considered as a cost-saving alternative. It was important to the program however, to have results that would be a sound alternative to the onsite panels that have traditionally been used for standard setting studies.

Another concern of the CLEP was the number of grade levels for which average scores would be reported. A benefit of gathering information on grades is that information is collected on the performance of students at all grade levels. The average scores for students with a grade of B were often reported for the CLEP exams, with some colleges using the average score associated with a grade of B as the credit-granting score, rather than the RCGS associated with a grade of C. This prompted an exploration of standard setting methods that would allow multiple grade levels to be reported, specifically, both the B and C grade levels.

The first standard setting method considered was a multiple Angoff method. This method asked judges to estimate for each question the percent of typical students who would know the correct answer, first at the B level and then at the C level. The judges were asked to choose from percents rounded to the nearest 10 percent: 10, 20, 30...90 percent.

The second standard setting method considered is a modification[2] of the yes/no method first suggested by Angoff (1971) and described by Impara and Plake (1997). This method asks judges to determine for each question whether a typical A, B, C, and D level student would know the correct answer. Judges answer yes or no for each grade level. Since this method is less cognitively demanding than the Angoff method, the A and D levels were included. Including all four levels is intended to 'anchor' the judges, resulting in more reasonable results for the B and C levels.

This study compares the results from the two methods, Angoff and yes/no, for an onsite standard setting panel and an online panel. Of specific concern was the judges' perception of the discussion as helpful in making the final judgments, with the revisions to the WBSS expected to produce results more like that of the onsite judges.

[2] Krishna Tateneni and Neil Dorans, personal communication, August 1999.

Methodology

Standard setting panels were convened for two tests, a United States History exam and a Natural Science exam. For each of the tests, one panel participated in the training onsite at ETS, while the other panel participated online via the Internet using the WBSS system.

Participants

Judges were recruited from lists of current college-level teachers of the relevant subject. They either taught at a college that currently uses at least one of the CLEP examinations, served on previous College Board committees or panels, or were recommended by a colleague. Judges were paid $300 for their participation in the study. Characteristics of the judges are presented in Tables 1 and 2. Nineteen judges were recruited for each of the online studies and 20 judges were recruited for the onsite panels.

Different facilitators were used for the four studies, online and onsite for United States History and Natural Science. The four facilitators worked together to ensure common materials and training techniques, and were overseen by an experienced standard setting panel leader.

Materials

The tests. The CLEP Examination in History of the United States I: Early Colonizations to 1877 is composed of 120 multiple-choice questions to be answered in two separately timed 45-minute sections. It covers material usually taught in the first semester of what is often a two-semester course in United States History.

The CLEP General Examination in Natural Sciences is composed of 120 multiple-choice questions to be answered in two separately timed 45-minute sections. The first section covers biological science and the second section covers physical science. The exam covers a wide range of topics frequently taught in introductory courses surveying both biological and physical sciences at the freshman or sophomore level.

Biographical questionnaire. Participants in both studies filled out a biographical questionnaire. The questionnaire asked for information such as the judges' gender, racial/ethnic background, and years teaching the subject.

Final questionnaire. Participants in both studies filled out a final evaluation at the end of the study. The questionnaire asked judges to rate, on a scale of one (strongly disagree or too slow) to five (strongly agree or too fast), several aspects of the study. The statements the judges were asked to rate can be grouped into six categories: training, group process, general process, working environment, time spent on training, and navigation. There were small differences between

the two questionnaires, such as changing "main menu" for the online study to "agenda" for the monitored study. Both groups also answered the question "Please rate your overall experience for the standard setting study (1=poor... ...5=very good)."

The web-based standard setting system. The WBSS system has three modules:

- A study editor, which prepares text for the web and creates the structure (steps, substeps, and pages) for the study
- The judge interface which is used by the panel members
- The facilitator interface which is used to monitor the judges' progress and responses

The facilitator has the following capabilities in the WBSS system:

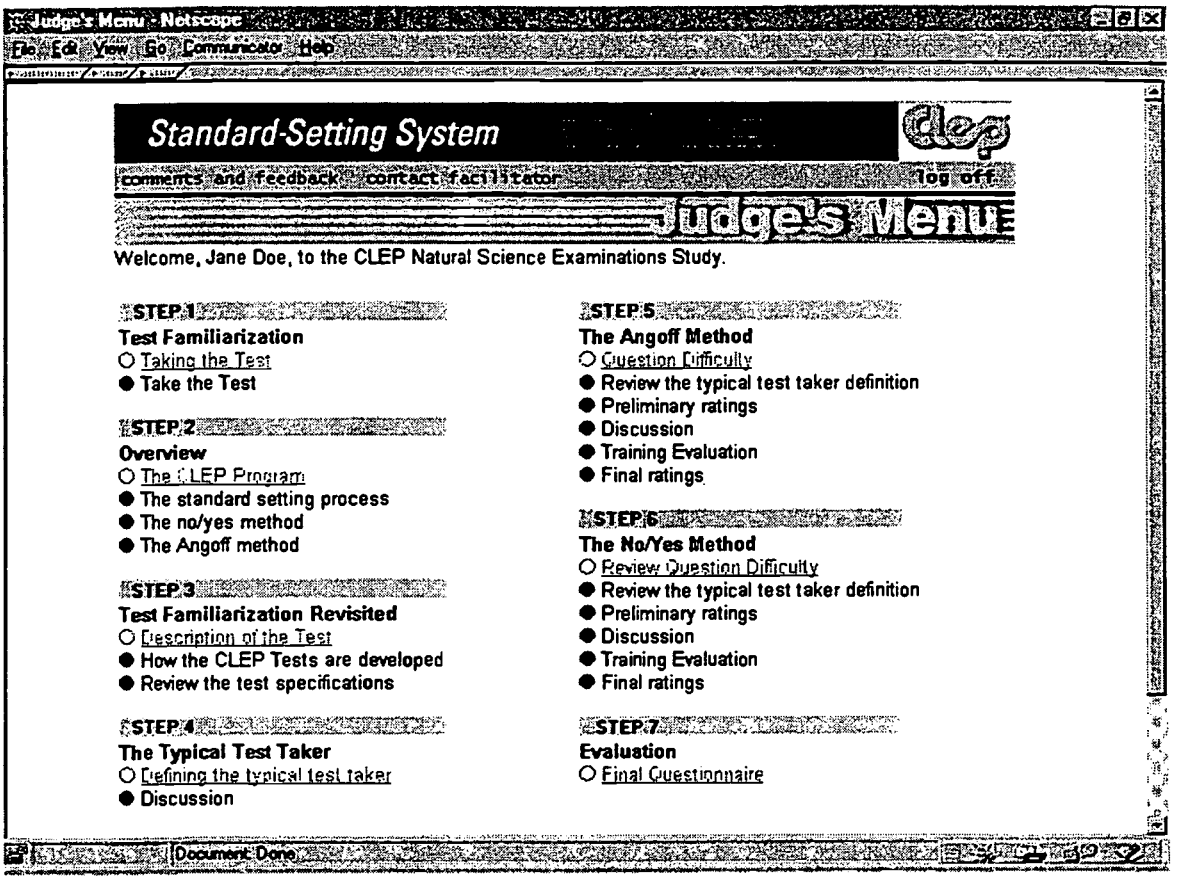
- Register and retire judges, allowing and disallowing access to the system
- View a particular judge's answers to the demographics questionnaire, time spent on each training page and item, judgments for each item, and comments made
- Allow access or deny access to each step in the study
- Close the study, disallowing any further changes to the data
- Participate in the discussions
- Create summaries of the judgments for display to the judges

The judges have the following capabilities:

- Proceed through the training materials in the order specified
- Participate in the discussions
- Send e-mail to the facilitator
- Record comments
- Record and revise judgments

The system begins with a welcome screen, which the facilitator can update with a new message as needed. Following the welcome screen, the judge must agree to a non-disclosure statement before proceeding to the main menu (see Figure 1). Beyond this point, there is a great deal of flexibility in the workflow of a particular standard setting study.

Figure 1
The Judges' Main Menu



Three types of screens make up the rest of the judges' interface:

- Text and graphics screens used for training (see Figure 2)
- Discussions, using a threaded discussion format, which organizes the responses by topic with responses to earlier messages indented under the original message (see Figure 3)
- Rating forms; in this case, an answer form, a no/yes rating form, and an Angoff rating form

Examples of the entire rating page are not given, as the questions are from active, secure, exam forms. However, the no/yes and Angoff portions of the rating pages are included in Figures 4 and 5.

Figure 2
Example of a text/graphics page

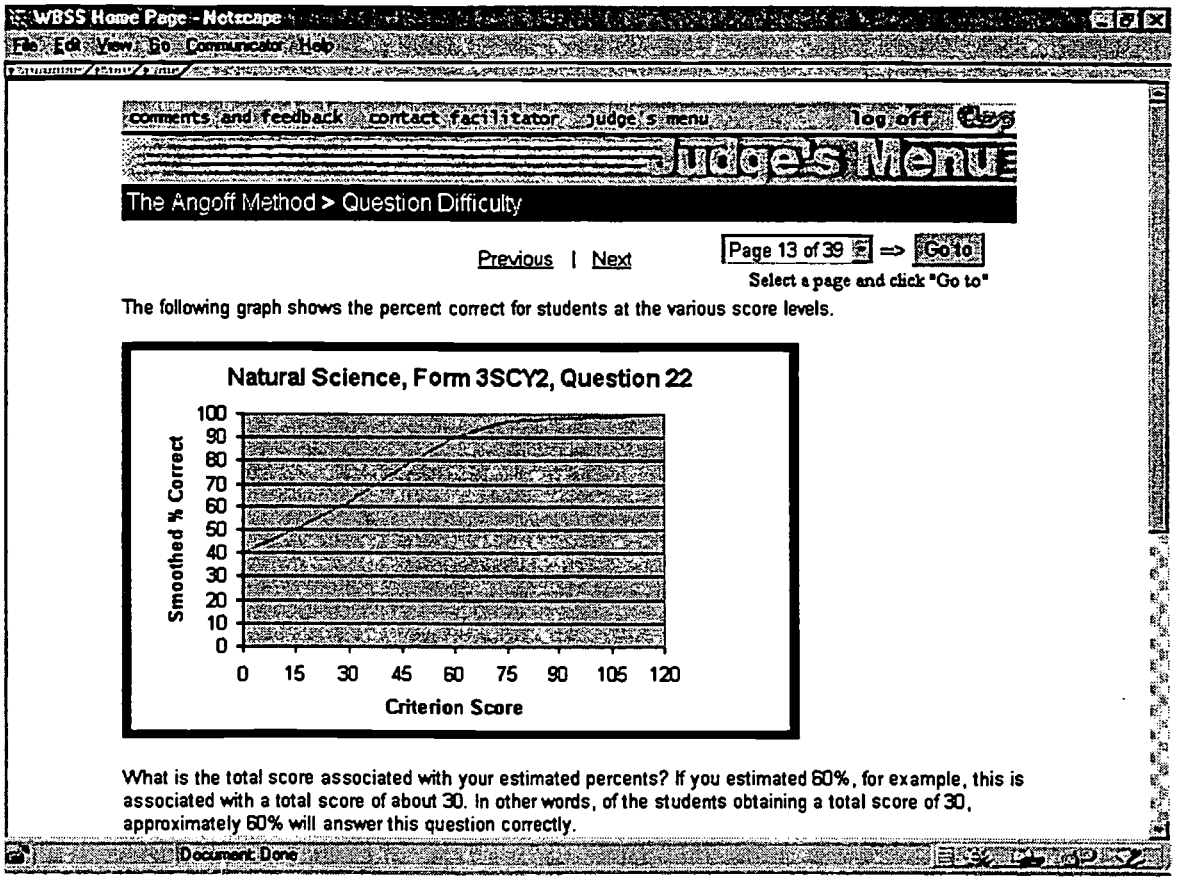


Figure 3
Example of a discussion page

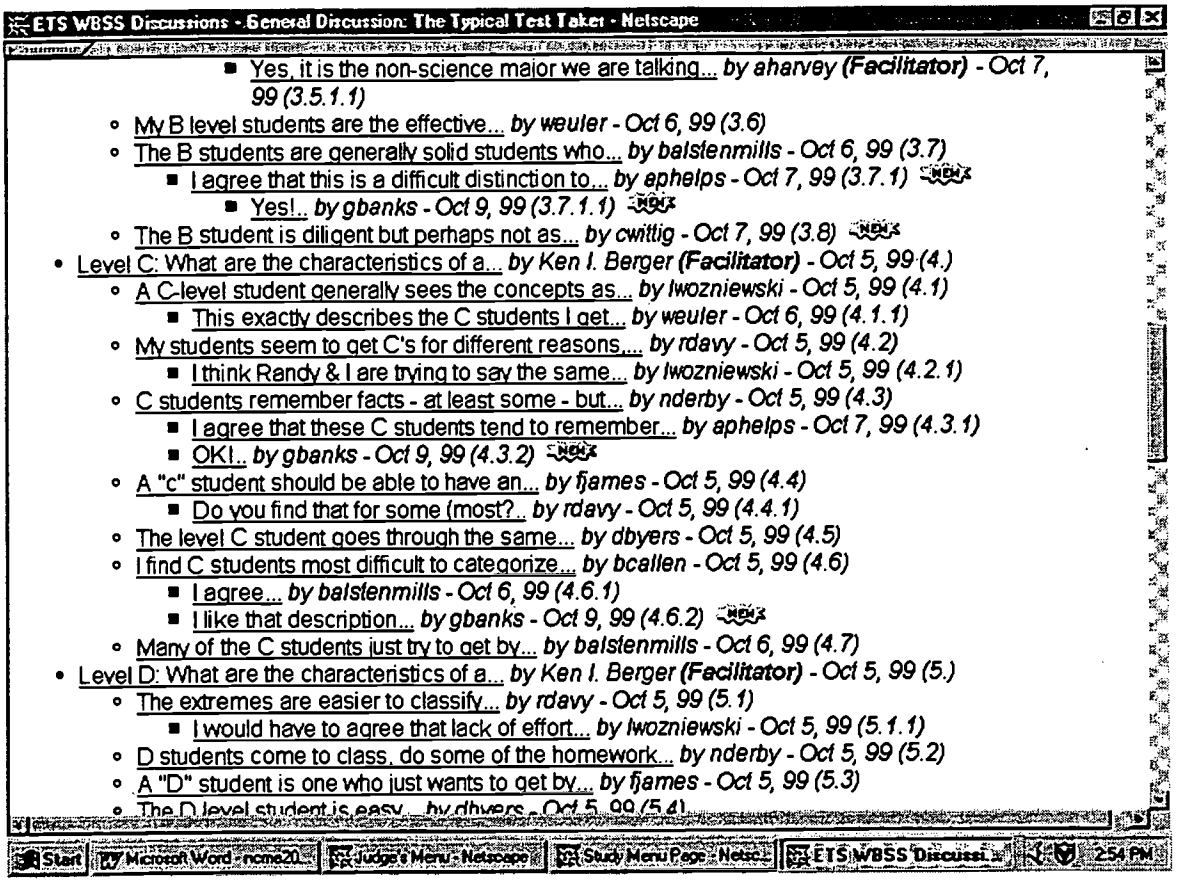


Figure 4
The no/yes rating form

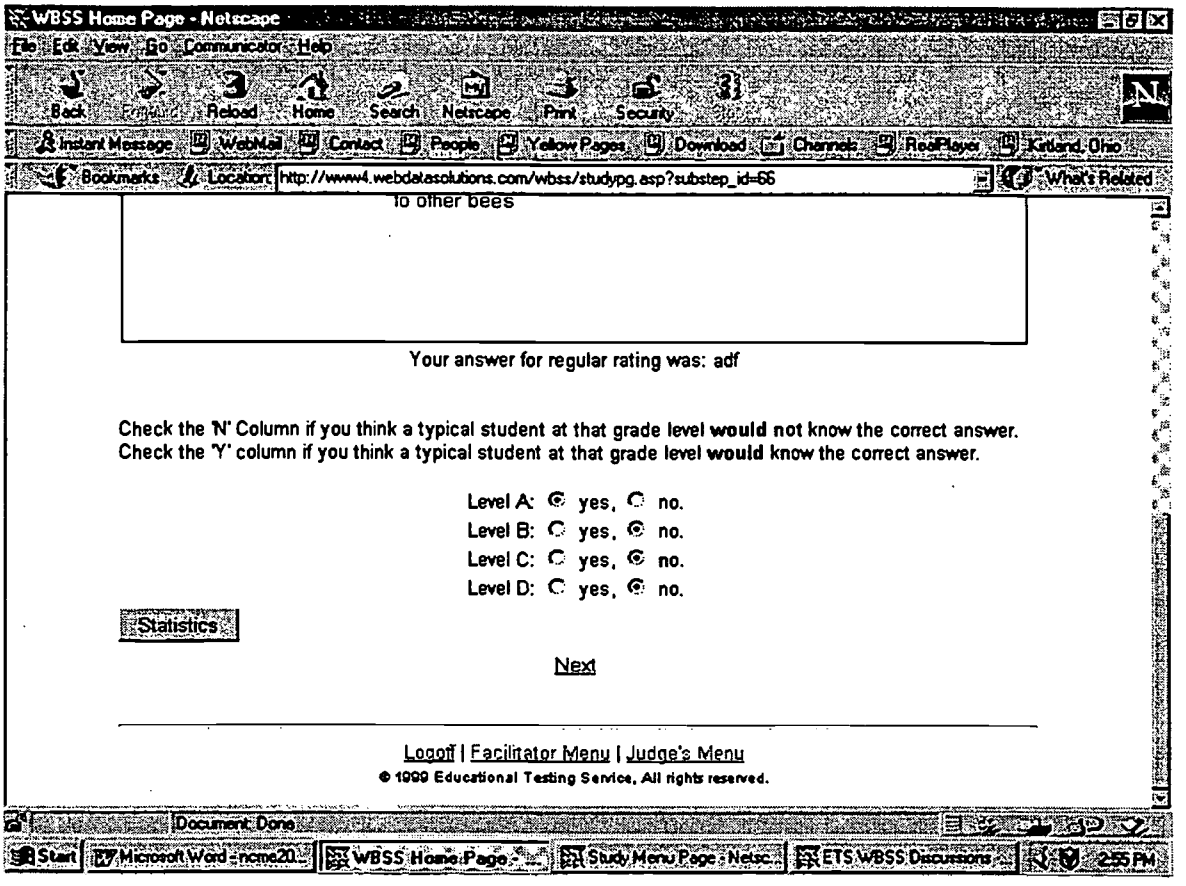
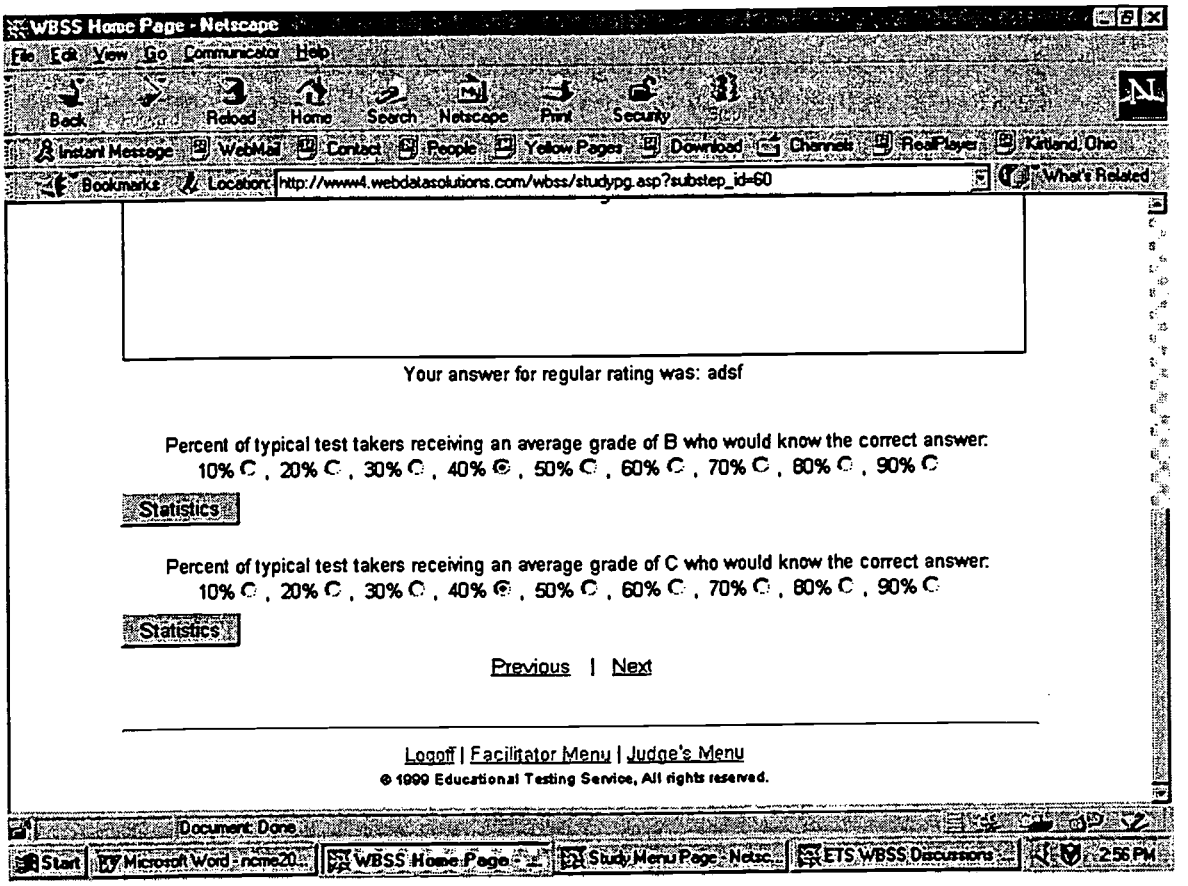


Figure 5
The Angoff rating form



Procedure

The onsite and online panels proceeded in the same sequence, as outlined in Figure 1.

1. The onsite panel was asked to fill out their biographical questionnaire and answer the exam questions before arriving for the study. The online panel filled out the questionnaire and answered the exam questions as their first task.
2. Both panels received information on the CLEP, the standard setting process, and specifics on their exam such as the content specifications.
3. The panels received a short description of the typical student and guidelines for a discussion.
4. A discussion took place in which the panels determined by consensus their definition of the typical student at the A, B, C, and D levels.

5. Panelists were instructed on characteristics of exam questions that would make them more or less difficult, regardless of topic. The effects of cognitive level, question format, complexity of phrasing, and the similarity of alternatives were discussed. At the same time, panelists practiced making judgments about questions and received feedback on the difficulty of the questions for CLEP examinees. In both the onsite and online studies, the United States History panels practiced using the no/yes method, while the Natural Sciences panels practiced using the Angoff method.
6. Panelists completed preliminary ratings of several questions intended to be characteristic of the tests. Again, the United States History panels used the no/yes method, while the Natural Sciences panels used the Angoff method.
7. The preliminary ratings were followed by a discussion of the ratings chosen.
8. Final ratings, no/yes ratings for the United States History panels and Angoff ratings for the Natural Sciences panels, were completed.
9. After a review of the question difficulty instructions given in step 5, panelists again gave preliminary ratings. For this step, the United States History panels used the Angoff method and the Natural Science panels used the no/yes method.
10. The preliminary ratings were discussed, followed by the final ratings for that method.
11. Panelists filled out the final evaluation.

The onsite study took approximately two days, or about 15 hours. The online study took approximately four weeks. The United States History panelists spent an average of 7.6 hours on the study (minimum=4.7 hours, maximum=14.2 hours), excluding the discussions, which could not be timed. The Natural Science panelists spent an average of 6.6 hours on the study (minimum=3.3 hours, maximum=10.2 hours), excluding the discussions.

Results

Samples

United States History panel. Of the 20 judges recruited for the onsite study, 19 attended the study and completed all ratings. Of the 19 judges recruited for the online study, 15 persisted for the entire study and completed all ratings. Biographical information is presented in Table 1.

Table 1
Background and Experience of the United States History Study Judges

	Online Sample	Onsite Sample
Number Recruited	19	20
Number Not Completing Study	4 (21% of 19)	1 (5% of 20)
Total Number Completing Study	15	19
Sex:		
Men	9 (60%)	12 (63%)
Women	6 (40%)	7 (37%)
Ethnic Group		
African American	2 (13%)	1 (6%)
Asian American	0 (0%)	1 (6%)
Hispanic	1 (7%)	0 (0%)
White	12 (80%)	16 (89%)
Missing	0	1
Years of Teaching Experience		
1 to 5	0 (0%)	3 (16%)
6 to 10	2 (13%)	1 (5%)
11 to 15	5 (33%)	3 (16%)
16 to 20	3 (20%)	5 (26%)
More than 20	5 (33%)	7 (37%)

Natural Sciences panel. Of the 20 judges recruited for the onsite study, 18 attended the study and completed all ratings. Of the 19 judges recruited for the online study, 15 persisted for the entire study and completed all ratings. Biographical information is presented in Table 2.

Table 2
Background and Experience of the Natural Science Study Judges

	Online Sample	Onsite Sample
Number Recruited	19	20
Number Not Completing Study	4 (21% of 19)	2 (10% of 20)
Total Number Completing Study	15	18
Sex:		
Men	10 (67%)	16 (89%)
Women	5 (33%)	2 (11%)
Ethnic Group		
African American	1 (7%)	0 (0%)
Asian American	0 (0%)	0 (0%)
Hispanic	0 (0%)	1 (6%)
White	14 (93%)	16 (94%)
Missing	0	1
Years of Teaching Experience		
1 to 5	1 (7%)	1 (6%)
6 to 10	3 (20%)	6 (33%)
11 to 15	3 (20%)	1 (6%)
16 to 20	2 (13%)	0 (0%)
More than 20	6 (40%)	10 (56%)

Final Questionnaire

The questionnaire was analyzed by averaging the responses for statements within the six categories, training, general process, group process, working environment, time spent on training, and navigation. These six scores were analyzed together with the overall experience question by completing a multivariate analysis of variance (MANOVA).

United States History study. The overall results for the MANOVA analyzing the final questionnaire results for the United States History study were statistically significant. Univariate follow-up tests indicated the results for the training, general process, time spent, navigation and overall statements were not significant. The training statements referred to the clarity and completeness of the training materials, comfort in asking the facilitator questions, promptness with which questions were answered, whether this was a good learning experience, and whether the directions for participating in the discussions was clear. The general process statements referred to comfort in filling out the forms, understanding the purpose of each exercise, confidence that the standard-setting process would produce a fair score, and understanding of the purpose for each

of the study exercises. The time-spent statements asked the judges to rate each of the steps in the study as too slow, too fast, or about right. The navigation statements referred to the progression of the topics, usefulness of the main menu or agenda, and whether the forms were easy to use. The results are presented in Table 3.

Table 3
Questionnaire Data for the United States History Study

	Online Sample (N = 14)	Onsite Sample (N = 19)
Training (9 questions)		
Average (S.D.)	4.2 (.4)	4.5 (.4)
Range	3.3 to 5.0	4.0 to 5.0
Group Process* (4 questions)		
Average (S.D.)	3.9 (.6)	4.5 (.4)
Range	2.8 to 5.0	3.5 to 5.0
General Process (5 questions)		
Average (S.D.)	4.2 (.6)	4.1 (.4)
Range	3.0 to 5.0	3.2 to 4.6
Working Environment * (4 questions)		
Average (S.D.)	3.7 (.8)	4.4 (.4)
Range	2.5 to 4.8	3.5 to 5.0
Time Spent (5 questions)		
Average (S.D.)	3.0 (.2)	3.0 (.4)
Range	2.6 to 3.2	1.8 to 4.0
Navigation (4 questions)		
Average (S.D.)	4.4 (.5)	4.2 (.5)
Range	3.8 to 5.0	3.0 to 5.0
Overall Experience (1 question)		
Average (S.D.)	4.4 (.5)	4.7 (.6)
Range	4.0 to 5.0	3.0 to 5.0

*Statistically significant ($p < .05$)
Wilks' Lambda = .39 ($F = 5.63$, $p < .0005$)
Note: One online panelist did not complete the final questionnaire.

The scale for the training, group process, general process, working environment, and navigation questions was 1 (strongly disagree) to 5 (strongly agree). The scale for the time spent questions was 1 (too slow) to 5 (too fast). The scale for the overall experience was 1 (poor) to 5 (very good).

Significant results were obtained for the group process and working environment scores. These two scores were further analyzed by completing a MANOVA for the statements within each of the categories. The results are presented in Tables 4 and 5.

Table 4
Questionnaire Data for the United States History Study: Group Process

	Online Sample (N = 14)	Onsite Sample (N = 19)
Discussions were helpful in rating questions		
Average (S.D.)	4.3 (.6)	4.3 (.7)
Range	3 to 5	3 to 5
Good opportunity to know colleagues and share ideas*		
Average (S.D.)	3.3 (1.3)	4.7. (.6)
Range	1 to 5	3 to 5
I was an active participant in the discussions*		
Average (S.D.)	3.9 (1.1)	4.5 (.5)
Range	1 to 5	4 to 5
I was comfortable sharing my ideas with other judges*		
Average (S.D.)	4.1 (.8)	4.6 (.5)
Range	3 to 5	4 to 5

*Statistically significant ($p < .05$)
Wilks' Lambda = .52 ($F = 6.39$, $p < .0009$)

Table 5
Questionnaire Data for the United States History Study: Working Environment

	Online Sample (N = 14)	Onsite Sample (N = 19)
Working conditions were pleasant*		
Average (S.D.)	3.9 (1.1)	4.8 (.4)
Range	2 to 5	4 to 5
There were few distractions*		
Average (S.D.)	3.5 (1.1)	4.5 (.7)
Range	2 to 5	3 to 5
Study location contributed positively*		
Average (S.D.)	3.7 (.8)	4.7 (.5)
Range	3 to 5	4 to 5
I was adequately compensated		
Average (S.D.)	3.5 (1.0)	3.8 (.6)
Range	2 to 5	3 to 5

*Statistically significant ($p < .05$)
Wilks' Lambda = .62 ($F = 4.30$, $p < .0077$)

The univariate follow-up analyses indicate that the online group was less likely to agree that the study was a good opportunity to get to know colleagues, less likely to agree they were an active participant in the discussions, and less comfortable sharing their ideas with other judges. There was no difference in whether or not the judges felt that the discussions were helpful in rating the questions.

The online group was also less positive about the working conditions. They were less likely to agree with statements about the working conditions being pleasant, that there were few distractions, and that the study location contributed positively. Both the online and onsite groups felt similarly about their compensation.

Natural Science study. The overall results for the MANOVA analyzing the final questionnaire results for the Natural Science study were statistically significant. Univariate follow-up tests indicated the results for the general process, time spent, navigation and overall statements were not significant. The results are presented in Table 6.

Table 6
Questionnaire Data for the Natural Science Study

	Online Sample (N = 15)	Onsite Sample (N = 18)
Training* (9 questions) Average (S.D.) Range	4.1 (.4) 2.9 to 4.6	4.5 (.4) 3.7 to 5.0
Group Process* (4 questions) Average (S.D.) Range	3.6 (.6) 2.5 to 4.5	4.6 (.4) 3.8 to 5.0
General Process (5 questions) Average (S.D.) Range	4.1 (.3) 3.6 to 4.6	4.1 (.7) 2.8 to 5.0
Working Environment * (4 questions) Average (S.D.) Range	3.7 (.7) 2.8 to 5.0	4.4 (.6) 3.3 to 5.0
Time Spent (5 questions) Average (S.D.) Range	3.0 (.3) 2.4 to 3.6	2.9 (.2) 2.4 to 3.2
Navigation (4 questions) Average (S.D.) Range	4.1 (.6) 3.0 to 5.0	4.3 (.6) 3.0 to 5.0
Overall Experience (1 question) Average (S.D.) Range	4.0 (.8) 2.0 to 5.0	4.4 (.5) 4.0 to 5.0

*Statistically significant ($p < .05$)
Wilks' Lambda = .32 ($F = 7.69$, $p < .0001$)

Significant results were obtained for the training, group process, and working environment scores. These three scores were further analyzed by completing a MANOVA for the statements within each of the categories. The results are presented in Tables 7, 8, and 9.

The univariate follow-up analyses were similar to those of the United States History study for the group process and working conditions statements. The online group was less likely to agree that the study was a good opportunity to get to know colleagues, less likely to agree they were an active participant in the discussions, and less comfortable sharing their ideas with other judges. There was no difference in whether or not the judges felt that the discussions were helpful in rating the questions.

As was true for the United States History study, the online group was less positive about the working conditions. They were less likely to agree with statements about the working conditions being pleasant, that there were few distractions, and that the study location contributed positively. Both the online and onsite groups felt similarly about their compensation.

For the training statements, the online group was less likely to agree that the training was clear and complete for the no/yes method. They were also less likely to feel comfortable asking questions and less likely to feel their questions were answered promptly.

Table 7
Questionnaire Data for the Natural Science Study: Training

	Online Sample (N = 15)	Onsite Sample (N = 18)
Training materials were clear and complete for:		
a. Overview and Introduction Average (S.D.) Range	4.4 (.6) 3 to 5	4.3 (.7) 3 to 5
b. The Typical Student Average (S.D.) Range	4.1 (1.0) 1 to 5	4.3 (.9) 2 to 5
c. Question Difficulty Average (S.D.) Range	4.0 (.4) 3 to 5	4.1 (1.0) 2 to 5
d. The No/Yes Method* Average (S.D.) Range	4.1 (.3) 4 to 5	4.4 (.5) 4 to 5
The Angoff Method Average (S.D.) Range	4.2 (.4) 4 to 5	4.3 (.8) 2 to 5
I was comfortable asking the facilitator questions* Average (S.D.) Range	4.1 (.8) 3 to 5	4.9 (.2) 4 to 5
All my questions were answered promptly* Average (S.D.) Range	3.4 (1.5) 1 to 5	4.6 (.5) 4 to 5
This was a good learning experience Average (S.D.) Range	4.3 (.7) 3 to 5	4.7 (.6) 3 to 5
The directions for participating in the discussions were clear Average (S.D.) Range	3.9 (1.0) 2 to 5	4.4 (.8) 2 to 5

*Statistically significant ($p < .05$)
Wilks' Lambda = .50 ($F = 2.45$, $p < .0414$)

Table 8
Questionnaire Data for the Natural Science Study: Group Process

	Online Sample (N = 15)	Onsite Sample (N = 18)
Discussions were helpful in rating questions Average (S.D.) Range	3.9 (.7) 2 to 5	4.1 (.8) 2 to 5
Good opportunity to know colleagues and share ideas* Average (S.D.) Range	3.1 (1.1) 1 to 5	4.7 (.5) 4 to 5
I was an active participant in the discussions* Average (S.D.) Range	3.5 (.7) 2 to 4	4.8 (.4) 4 to 5
I was comfortable sharing my ideas with other judges* Average (S.D.) Range	4.1 (.8) 2 to 5	4.8 (.4) 4 to 5

*Statistically significant ($p < .05$)
Wilks' Lambda = .34 ($F = 13.45$, $p < .0001$)

Table 9
Questionnaire Data for the Natural Science Study: Working Environment

	Online Sample (N = 15)	Onsite Sample (N = 18)
Working conditions were pleasant*		
Average (S.D.)	4.0 (.8)	4.8 (.4)
Range	3 to 5	4 to 5
There were few distractions*		
Average (S.D.)	3.5 (1.2)	4.6 (.5)
Range	1 to 5	4 to 5
Study location contributed positively*		
Average (S.D.)	3.5 (.9)	4.6 (.6)
Range	2 to 5	3 to 5
I was adequately compensated		
Average (S.D.)	3.6 (.7)	3.6 (1.4)
Range	2 to 5	1 to 5

*Statistically significant ($p < .05$)
Wilks' Lambda = .50 ($F = 7.01$, $p < .0005$)

Rating Results

Angoff method results and the no/yes method results were calculated by adding the ratings for each question and then averaging across judges, separately for each rating level. Although the A and D level results for the no/yes method would not be reported to colleges, they are included, for completeness, in the analysis comparing the results for the online and onsite samples.

The Angoff method and no/yes method results were analyzed using a MANOVA. Sample (online or onsite) is the independent variable and Angoff B and C level and no/yes A, B, C, and D level ratings are the dependent variables.

United States History study. The results for the analysis of the United States History study ratings indicate significant differences in the ratings of the two samples (see Table 9). Univariate follow-up analyses indicate that the Angoff B and C level ratings are significantly different for the online and onsite panels, with the online panel rating questions higher in both cases. The results for the no/yes method were not significantly different.

Table 9
Average Ratings for the United States History Study

	Online Sample (N = 15)	Onsite Sample (N = 19)	Difference
Angoff B*			
Average (S.D.)	83.3 (6.9)	67.8 (6.3)	15.5
Range	23.3 (67.3 to 90.6)	22.5 (54.4 to 76.9)	.8
Angoff C*			
Average (S.D.)	53.7 (7.9)	36.2 (8.6)	17.5
Range	31.3 (31.1 to 62.4)	28.5 (22.6 to 51.1)	2.8
No/Yes A			
Average (S.D.)	115.4 (8.0)	112.5 (10.1)	2.9
Range	27 (93 to 120)	32 (88 to 120)	-5
No/Yes B			
Average (S.D.)	90.1 (11.6)	85.4 (16.6)	4.7
Range	45 (66 to 111)	64 (52 to 116)	-19
No/Yes C			
Average (S.D.)	43.3 (13.0)	42.9 (20.6)	.4
Range	45 (22 to 67)	80 (14 to 94)	-35
No/Yes D			
Average (S.D.)	11.6 (10.5)	12.9 (18.7)	-1.3
Range	33 (0 to 33)	71 (0 to 71)	-38

*Statistically significant ($p < .05$)
Wilks' Lambda = .27 ($F = 12.47$, $p < .0001$)

Natural Science study. The results for the analysis of the Natural Science study ratings indicate no significant differences in the ratings of the two samples (see Table 10).

Table 10
Average Ratings for the Natural Science Study

	Online Sample (N = 15)	Onsite Sample (N = 18)	Difference
Angoff B			
Average (S.D.)	72.8 (9.0)	70.2 (9.0)	2.6
Range	33.2 (51.3 to 84.5)	27.0 (55.3 to 82.3)	6.2
Angoff C			
Average (S.D.)	50.6 (7.3)	50.0 (7.5)	.6
Range	22.2 (38.3 to 60.5)	27.3 (33.9 to 61.2)	-5.1
No/Yes A			
Average (S.D.)	113.3 (6.4)	113.2 (7.9)	.1
Range	24 (96 to 120)	32 (88 to 120)	-8
No/Yes B			
Average (S.D.)	79.1 (12.4)	85.6 (19.0)	-6.5
Range	43 (55 to 98)	57 (57 to 114)	-14
No/Yes C			
Average (S.D.)	31.9 (9.5)	37.5 (15.3)	-5.6
Range	35 (12 to 47)	53 (7 to 60)	-18
No/Yes D			
Average (S.D.)	8.1 (6.2)	9.3 (5.2)	-1.2
Range	18 (0 to 18)	17 (0 to 17)	1

*Statistically significant ($p < .05$)
Wilks' Lambda = .90 ($F = .49$, $p < .8107$)

Correlation with Observed Data

Percent correct for each question was calculated from data gathered on CLEP examinees. The percents correct were correlated, using a Spearman rank-order correlation, with the average rating for each question.

United States History study. The percents correct for the United States History study questions were calculated from a sample of 3,210 examinees. The results of the correlations with judges' ratings are presented in Table 11.

Table 11
Correlations between Judge Ratings and Observed Data For the United States History Study

	Online Sample (N = 15)	Onsite Sample (N = 19)
Angoff B	.55	.56
Angoff C	.54	.51
No/Yes B	.52	.57
No/Yes C	.47	.58

Natural Science study. The percents correct for the Natural Science study questions were calculated from a sample of 2,320 examinees. The results of the correlations with judges' ratings are presented in Table 12.

Table 12
Correlations between Judge Ratings and Observed Data For the Natural Science Study

	Online Sample (N = 15)	Onsite Sample (N = 18)
Angoff B	.70	.64
Angoff C	.71	.68
No/Yes B	.66	.67
No/Yes C	.64	.64

Discussion

The goal of improving the discussion interface of the WBSS system was modestly realized. Unlike the pilot study (Harvey and Way, 1999), no significant differences were found in the tendency for judges to agree that the discussions were helpful in rating the exam questions. This was true for both the United States History study and the Natural Science study.

Several aspects of the group process continue to differ substantially for the online and onsite groups. In both studies, for example, the online group was less likely than the onsite group to view themselves as an active participant in the discussions. It is perhaps the nature of an online process that participants will feel less involved than when they are face-to-face. That should not stop the study facilitator from encouraging greater participation from online discussants. In the same way that many of us have a 'toolbag' of tricks to draw out shyer or less interested participants in a teaching situation, techniques for doing the same for an online audience are worth exploration.

The facilitator is still an important aspect of the equation in a successful online study, as evidenced by the significant differences in facilitator related statements for the Natural Science study questionnaire. Although not significantly different, the United States History study showed a very similar pattern.

There were substantial differences in Angoff rating results for the United States History study. There were no significant differences in Angoff ratings for either the pilot study or for the Natural Science study, so it does not appear to be inherent to the online mode of study. Nevertheless, the differences are non-trivial, 15 to 17 points on a 120-question test. Such differences would mean 12% fewer passing at the B level and 34% fewer passing at the C level. It will be important to explore reasons why the differences might have occurred.

One possible explanation for the differences in Angoff ratings for the United States History study, but not for the Natural Science study, is differences in the discussion of the typical student. It may be that the consensus definition at each level was qualitatively different for the two panels participating in the United States History study, but not for the Natural Science study panels. An analysis of transcripts of the discussions and the resulting summary definitions might bear this out. If such a difference were found, of course, it would beg the question of why this would be true in one case, but not the other. A short list of possibilities could include aspects of the facilitation, the test content, and the background characteristics of the judges. Each is worthy of attention in future research.

Another possibility for the difference could reside in the use of the no/yes method for the initial training of the United States History panels on factors affecting question difficulty. This session is the sole aspect of the study providing examinee data. While both panels practiced the no/yes method during their

question difficulty training, it may be that the onsite facilitator could introduce subtleties in the training that assisted the onsite panel in making their Angoff judgments, that were not available to the online panel. It would seem, however, that the correlations of question difficulty and judge ratings would be substantially different for the two modes of presentation if this were the explanation.

There seems little doubt that the working conditions for the online group are less pleasant, focused, or positive than those experienced by the onsite study. This finding is persistent across studies and echoes the results of the pilot study. Although this does not appear to affect rating results, it cannot be dismissed when considering an online study over an onsite study. On the other hand, the overall experience was viewed similarly by both groups. With the often substantial cost savings, potential for substantive review of training materials, and the potential for greater participation rates, an online standard setting study appears to be a reasonable alternative to an onsite study.

References

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

Faggen, J. (1994, November). *Setting Standards for Constructed-Response Tests: An Overview.* Research Memorandum RM-94-19. Princeton, NJ: Educational Testing Service.

Harvey, A. L. & Way, W. D. (1999, April). *A comparison of web-based standard setting and monitored standard setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

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

Author's Notes

Anne L. Harvey is a Measurement Statistician II in the Analysis Division at Educational Testing Service.

The author would like to thank Neil Dorans and Krishna Tateneni for their contribution to the design of the study. Jane Faggen, Nancy Olds, and Uma Venkateswaran did an excellent job as the facilitators for three of the panels. Richard Carvalho and Laura Scheffler took very good care of the logistics for the onsite studies. Reviews of earlier drafts of this paper by Diane Bailey, Robert Smith, and Walter D. Way resulted in some very helpful changes.

Requests for additional copies of this report may be directed to Anne L. Harvey, Mailstop 15-L, Educational Testing Service, Princeton, NJ 08541-0001

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

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: COMPARING ONSITE AND ONLINE STANDARD SETTING METHODS FOR MULTIPLE LEVEL OF STANDARDS	
Author(s): ANNE L. HARVEY	
Corporate Source: EDUCATIONAL TESTING SERVICE	Publication Date: APRIL 2000

II. REPRODUCTION RELEASE:

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

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

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

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

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

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

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

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

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

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

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

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

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

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

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

Sign here, → please

Signature: Anne L Harvey	Printed Name/Position/Title: ANNE L. HARVEY, DIR. OF PRODUCT DEV.
Organization/Address: THE COLLEGE BOARD, 45 COLUMBUS AVENUE, NY, NY 10023	Telephone: 212-713-8070 FAX: 212-649-8427
	E-Mail Address: aharvey@collegeboard.org Date: 3/13/02

(over)

ERIC®

Clearinghouse on Assessment and Evaluation

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

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

May 8, 2000

Dear AERA Presenter,

Hopefully, the convention was a productive and rewarding event. As stated in the AERA program, presenters have a responsibility to make their papers readily available. If you haven't done so already, please submit copies of your papers for consideration for inclusion in the ERIC database. We are interested in papers from this year's AERA conference and last year's conference. If you have submitted your paper, you can track its progress at http://ericae.net.

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

We are gathering all the papers from the **2000 and 1999 AERA Conference**. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

Please sign the Reproduction Release Form enclosed with this letter and send **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can mail your paper to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to: AERA 2000/ERIC Acquisitions
University of Maryland
1129 Shriver Laboratory
College Park, MD 20742

Sincerely,

Lawrence M. Rudner

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