DOCUMENT RESUME

ED 387 505 TM 023 665

AUTHOR Webb, Melvin W., II; Miller, Eva R.
TITLE A Comparison of the Paper Selection Method and the Contrasting Groups Method for Setting Standards on Constructed-Response Items.
PUB DATE 20 Apr 95
NOTE 22p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Constructed Response; Criteria; Educational Assessment; Grade 8; Interrater Reliability; Judges; Junior High Schools; Performance; *Reading Tests; *Scoring; *Standards; *Test Items
IDENTIFIERS *Contrasting Groups Method; Early Warning Test NJ; National Assessment of Educational Progress; *Paper Selection Method; Standard Setting

ABSTRACT

As constructed-response items become an integral part of educational assessments, setting student performance standards on constructed-response items has become an important issue. Two standard-setting methods, one used for setting standards on the National Assessment of Educational Progress (NAEP) in reading in grade 8 and the other used to set standards on the 1993 New Jersey Early Warning Test (EWT) in reading for grade 8, are compared. In the paper selection method, used for the NAEP, judges first conceptualize students who are just at the borderline between categories of performance and then select actual papers from a set of all levels of performance to represent what students at the borderline would have produced. For the EWT, the contrasting groups method was used. Judges internalize the concepts to be assessed and then select students above and below the criterion of success. Papers by these students are scored, and a point between the two score distributions is selected as the standard. Comparison suggests that the contrasting groups method may yield results that are more accurate than those of the paper selection method, although results are not definitive. Two figures and three tables illustrate the comparisons. (Contains 14 references.) (SLD)

A Comparison of the Paper Selection Method and the Contrasting Groups Method for Setting Standards on Constructed-Response Items

by

Melvin W. Webb II, Ed.D.
Philadelphia School District

Eva R. Miller, Ph.D.
New Jersey Department of Education

presented to

the Annual Meeting of
The National Council on Measurement in Education

San Francisco, CA

April 20, 1995

INTRODUCTION

According to Berk (1986), the process of setting standards on standardized tests is the most complicated technical issue in criterion-referenced measurement. It is "controversial to discuss, difficult to execute, and impossible to defend" (p.565). When Berk made this statement, setting standards on constructed-response items was not really a part of the larger standard-setting issue. Now, with constructed-response items becoming an integral component of the National Assessment of Educational Progress (NAEP), state assessments such as New Jersey's High School Proficiency Test for eleventh grade (HSPT11) and their Eighth Grade Early Warning Test (EWT), and even commercial publishers' mainstream achievement tests (e.g., the Stanford Achievement Test-9th edition (SAT-9)), setting student performance standards on constructed-response items has become an issue with which we must wrestle.

Reckase (1994) points out that "there is very little in the measurement literature that provides guidance for setting standards on performance assessment tasks" (p.2). Yet research conducted by one of the present authors and his former colleagues on the 1992 NAEP in Reading (ACT, 1992, March) indicates that constructed-response items may present even more complications for standard-setting than do multiple-choice items. In addition, proven standard-setting methods for multiple-choice items, such as the modified Angoff and the Nedelsky, simply do not exist for constructed-response items. As Hambleton and Plake (1994) said, in referring to the 1992 NAEP Achievement-Levels Setting project (ACT, 1992, January), "there may be only one previous large-scale

initiative to set standards on performance assessments that has been well documented" (p.2).

Given the paucity of research on standard-setting for constructed-response items, how do we go about judging the relative merits of the different methods being used for setting standards on these type items? According to Plake (1994, April), several important features need to be considered when judging the effectiveness of a particular standard-setting method. These include:

- the accuracy of the decisions resulting from the application of the standard; this is primary
- the ease of administration
- panelists' comfort with the final decision rule
- panelists' confidence in the results
- potential replicability of the decision rule resultant from the standard-setting procedure

In the absence of standards for standard-setting, Plake's list of features to consider when evaluating competing methods for setting standards on constructed-response items provides a convenient starting point. This paper will compare two such methods, one used for setting standards on the National Assessment of Educational Progress (NAEP) in Reading (Grade 8) , and the other used for setting standards on the 1993 New Jersey Early Warning Test (EWT) in Reading (Grade 8), using a modified version of Plake's criteria. The potential replicability of the decision rule resultant from the standard-setting procedures, however, will not be addressed other than through references to a simulation study by Reckase (1994, June).

LIMITATIONS

The comparisons made in this paper are limited by certain features of the assessments and the standard-setting processes used. The NAEP and the EWT are different tests, developed for different purposes. The NAEP, for example, is a survey instrument which utilizes matrix sampling techniques to administer a very large item pool to a small number of students, nationwide. No individual completes more than a small percentage of the entire item pool, individual results are not reported, and there are no individual consequences for students, no matter where the standards are set on the score scale. The EWT, on the other hand, contains a limited number of items, examinees complete all items on the assessment, individual results are reported, and there are consequences for individuals, large numbers of whom will be affected by the placement of the standards on the score scale. How these different attributes of the assessments may have influenced the judges' decisions relative to standard-setting is impossible to determine. In addition, differences in the standard-setting designs developed for the NAEP and the EWT also may have influenced the judges' decisions. For example, different panels of judges were used for the two standard-settings, and the composition of the panels differed. The NAEP panel was composed of teachers (57 percent), non-teacher educators (16 percent), and members of the general public (27 percent), while the EWT panel was composed entirely of teachers. Because both standard-setting methods under consideration relied on judges' ability to make decisions about the quality of actual student work, differences in the composition of

the two panels may have affected the outcome (Jaeger, 1991; Reid, 1991).

Despite these and other differences in the assessments and the standard-setting designs, the authors felt the assessments and designs shared enough common characteristics to make comparisons of the standard-setting methods worthwhile.

THE TWO METHODS DESCRIBED

Paper selection method. The method developed by American College Testing (ACT) and the National Assessment Governing Board (NAGB) to set standards on the 1992 NAEP's constructed-response items has been generally referred to as the "paper selection" method (ACT, 1992, January). In this method, judges first conceptualize students who are just at the "borderline" between categories of performance. They then select actual student papers (responses to Reading prompts), from a set of papers representing all possible levels of performance, that students at the borderline would have produced.

For the 1992 Grade 8 NAEP in Reading, judges were presented with twenty-four papers for each of the constructed response items, with six papers at each of the four score points. Judges were instructed to read all twenty-four papers for each item carefully, and to select three papers from the set, one each to represent borderline student performance on that item at the three achievement levels (Basic, Proficient, and Advanced). All papers had been scored previously, but judges were not given the scores.

This process was repeated across three rounds, with feedback in the form of intrajudge and interjudge consistency data, presented to judges between rounds (ACT, 1993, August). In addition, during Round 3, judges were allowed to discuss paper selections they found problematic with other judges in their group.

Scores for the final paper selections from Round 3 were used to compute the numerical standards (Basic, Proficient, and Advanced) for the constructed-response items, and these standards were combined with the numerical standards derived from the multiple-choice and short-answer items to produce the final numerical standards for the 1992 NAEP in Reading (Luecht, 1993, August).

Contrasting groups method. According to the National Academy of Education, the contrasting groups method typically involves having a group of judges internalize the construct to be assessed, then select students that are above and below the criterion of success. Papers written by these students are then scored, and a point between the two score distributions is selected to be the standard (NAEP, 1993). The contrasting groups method developed by National Computer Systems (NCS) and the New Jersey Department of Education (NJDOE) differed from that described by NAE (NCS, 1995, February). For the EWT, the judges (all of whom were teachers familiar with the generic constructed-response rubric) read a selected sample of twenty actual student papers for each constructed-response item (N=4). As was the case for the NAEP process, papers were in a random order in relation to their actual scores. At the time of first reading, judges were blind to the

scores. Judges were instructed to work individually to sort the papers into three categories: 1) does not need instructional intervention, 2) may or may not need instructional intervention, and 3) needs instructional intervention. Judges were told that any categoric classification or combination of the papers was possible, including their classifying all papers into only one of the three categories. The sorting process was repeated across three rounds, with feedback provided to judges in the form of intrajudge and interjudge consistency data and rubric scores for each paper between rounds (NCS, 1995, February).

The resulting standards were computed according to a formula which simply averaged the scores of papers classified as minimally competent (categories 1 and 2 above) with the average score of those papers categorized as not competent (category 3). The constructed-response standard was then combined with the dichotomous standard to produce the final numerical standard for the EWT in Reading (NCS, 1995, February).

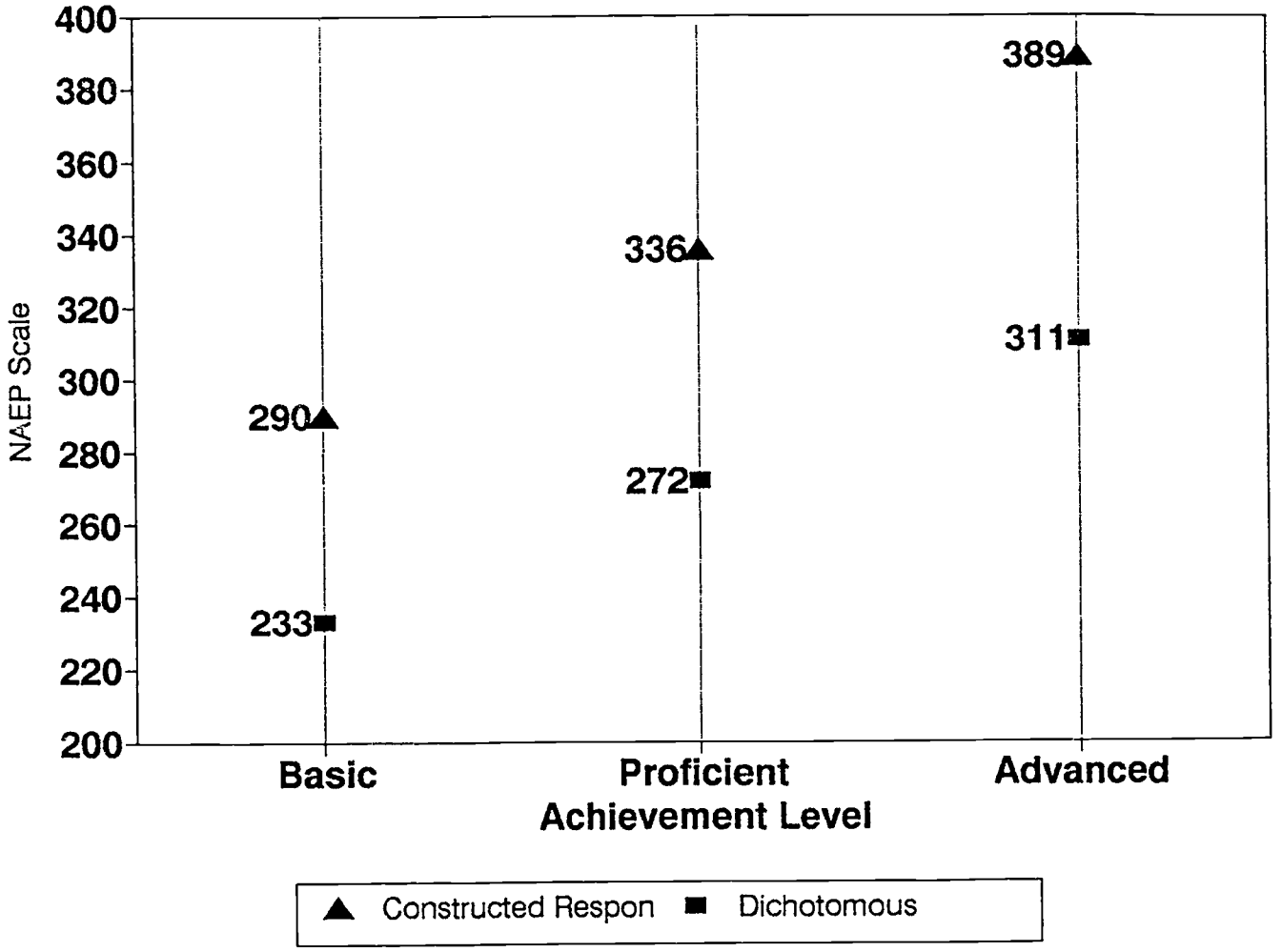
THE TWO METHODS COMPARED

The accuracy of the decisions. Reckase, in a simulation study, found that both the paper selection and contrasting groups methods over-estimated the percent of students passing the test, with the contrasting groups method over-estimating at a substantially higher rate than the paper selection method (Reckase, 1994). In terms of setting a standard, this would result in a standard that was artificially high, as more students would be expected to meet the standard than actually met the standard. To

determine if this result was observed for the NAEP and EWT standard-setting, the standard that would have resulted from using only the constructed-response items was compared to the standard that would have resulted from using only the dichotomously-scored (multiple-choice and short-answer) items. If Reckase's findings were replicated with real data, one would expect to observe a constructed-response standard that was considerably higher on the scale than that observed for the dichotomously-scored items. Because the modified Angoff method used for the dichotomously-scored items has been found to consistently over-estimate the percentage of examinees who would be successful (NAE, 1993, July), an even higher standard based on the constructed-response items may provide evidence that the method(s) in question yield inaccurate results.

For the Grade 8 NAEP in Reading, the paper selection method resulted in standards that were substantially higher (over 1 SD on the NAEP scale) at all three achievement levels (Basic, Proficient, and Advanced) than those resulting from the modified Angoff method (ACT, 1993, August; see Figure 1). The Basic level constructed-response standard was at a higher scale point than the Proficient dichotomous standard (290 and 272, respectively, on the NAEP scale), while the Proficient level constructed-response standard was at a higher scale point than the Advanced dichotomous standard (336 and 311, respectively). The Advanced level constructed-response standard was 389, over seventy scale points higher than the dichotomous standard (311). These results would appear to confirm Reckase's findings.

Figure 1
Grade 8 NAEP Reading Standards



For the Grade 8 EWT in Reading, the contrasting groups method resulted in minimally competent and clearly competent standards for the constructed-response items that were somewhat lower on a percent correct metric than the modified Angoff method used for the multiple-choice items (NCS, 1995, February; see Figure 2). For minimally competent, the constructed-response standard was 5.16 out of 12 possible points (43 percent correct) while the multiple-choice standard was 26.29 out of 44 possible points (60 percent correct).

Table 1

Comparison of Standards*

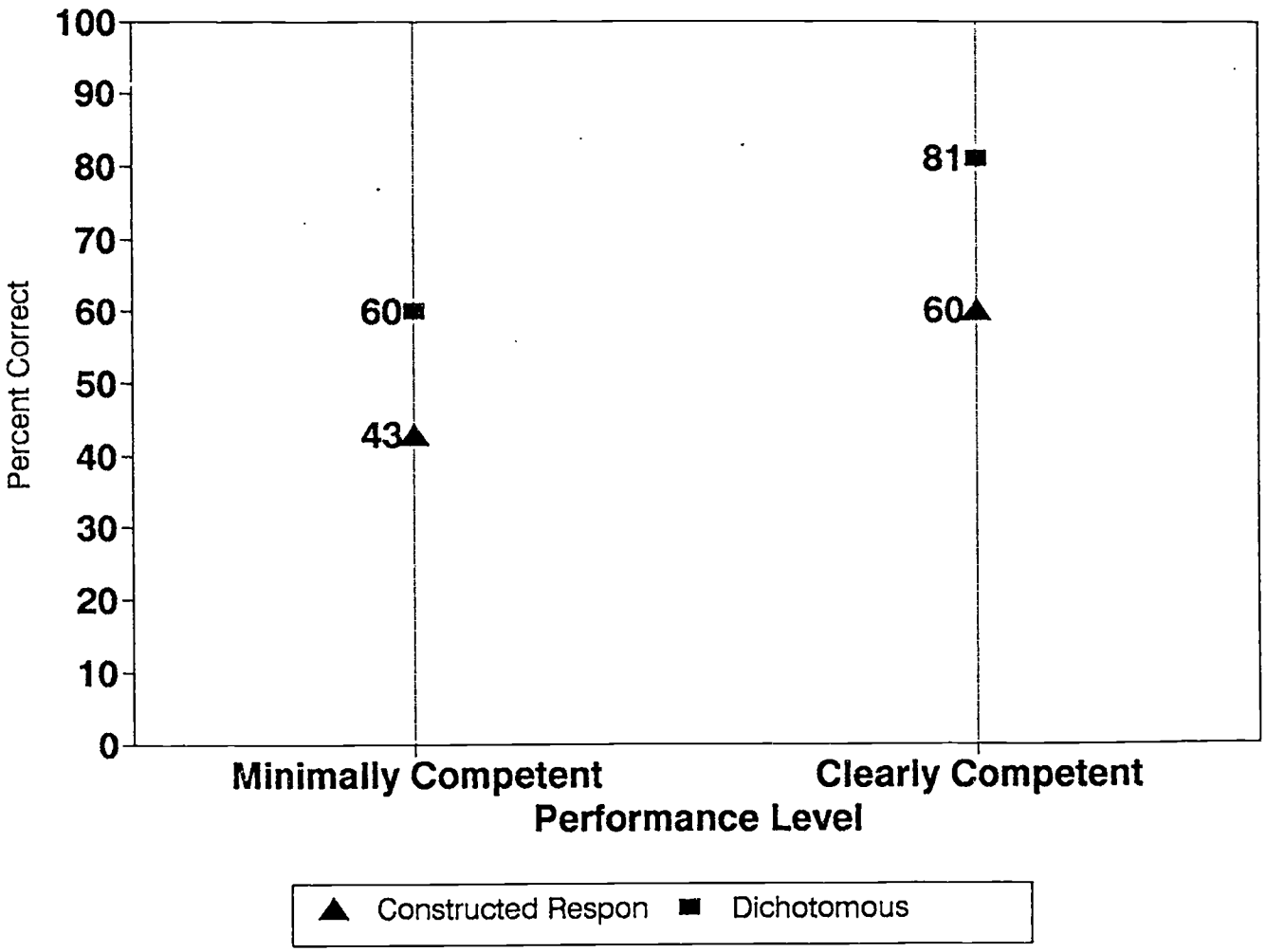
Method	Achievement Level		
	Basic	Proficient	Advanced
Round 3 Paper Selection	290	336	389
Round 4 Study	232	269	302
Angoff/dichotomous	232	272	311

* from Luecht, 1993 (March).

As the data in Table 1 show, judges' standards based on the Round 4 , modified Angoff-type method of estimating the percentage of students who would answer the item correctly (obtain a "passing" score) were very similar to their Round 3 modified Angoff standards, and substantially different from their Round 3 paper selection standards. This would also appear to support Reckase's finding that the paper selection method overestimates students performance.

For the EWT Special Study, judges were asked to choose an

Figure 2
New Jersey EWT Reading Test



"ideal" cut-score on the four constructed response items to represent minimally and clearly competent response items, a total of 12 points (4 items x 3 points) was possible. Table 2 presents the results from the Day 3 Special Study, from the Day 3 modified Angoff dichotomous ratings, and from the Day 3 contrasting groups method. As shown in Table 2, both the Day 3 Special Study standard and the Day 3 contrasting groups standard (on the present correct metric) were lower than the modified Angoff standard. The contrasting group method, in fact produced the lowest standard of the three methods, contradicting Reckase's findings and lending support to the contrasting groups methodology. Considering that students appear to do less well on constructed-response items than on multiple-choice items (NCES, 1993 July), the contrasting groups method used for the EWT would appear to produce more defensible standards.

Table 2

Comparison of Data From EWT Day 3 Modified Angoff, Contrasting Groups, and Special Study Methods*

Method	% Correct score Minimally Competent	% Correct score Clearly Competent
Day 3 Contrasting Groups	43%	60%
Day 3 Special Study	50%	73%
Day 3 Modified Angoff	60%	81%

* From NCS, 1995 (February).

Ease of Administration. For purposes of this paper, ease of administration will be defined as a) cognitive complexity of the task for judges, and b) physical/time burden for judges.

For the 1992 NAEP in Reading, panelists were asked to conceptualize students who were who are just at the "borderline"

between categories of performance, then to select student papers that students at the borderline of performance would have produced. That is, they were to select one paper that represented borderline basic performance, one that represented borderline proficient performance, and one that represented borderline Advanced performance. Cognitively, panelists found this to be a relatively easy task to perform, with 90% of the panelists indicating, that by Round 3, they had more than an adequate level of understanding of the task they were to perform and 95% indicating that their conception of borderline performance was more than moderately well-formed (ACT, 1993 August). The main difficulty panelists expressed was finding true "borderline" papers, which was probably due to the types of papers chosen for the project. In general, papers at each score point were selected as "solid" examples of those score points rather than as "borderline" examples (author's recollection).

In terms of physical/time burden on judges, each judge has to read and evaluate between 144-168 papers, depending on whether they had 6 or 7 constructed response items in their half of the item pool. After reading a set of papers for a prompt, most judges sorted papers into three piles (Basic, Proficient, Advanced) then re-evaluated papers in each pile to select the single paper in each pile that represented borderline performance. Despite the number of papers to be evaluated, 85% of the judges indicated they had a sufficient has to read and evaluate between 144-168 papers, depending on whether they had 6 or 7 constructed response items in their half of the item pool. After reading a set of papers for a prompt, most judges sorted papers into three piles (Basic,

Proficient, Advanced) then re-evaluated papers in each pile to select the single paper in each pile that represented borderline performance. Despite the number of papers to be evaluated, 85% of the judges indicated they had a sufficient with some re-reading of papers for prompts/achievement levels about which they were unsure. In general, the paper selection method appears to be a reasonable standard-setting method in terms of ease of administration.

For the EWT, panelists were given a set of 20 papers per prompt. Using item-specific scoring rubrics and their content knowledge, judges were asked to sort the papers for each prompt into tow piles, with one pile designated "does not need instructional intervention" and the second designated "may or may not need instructional intervention." Judges were then to resort the latter pile into two piles designated "may need instructional intervention "and" needs instructional intervention". The result was three piles of papers separating papers from students who need instructional intervention, who may need instructional intervention, and who don't need instructional intervention. This represented a change in focus re student performance from the modified Angoff method, where the concepts of "minimally competent" and "clearly competent" performance found the basis of judges ratings. As much, it introduced the possibility of confusion into the process.

Judges however, indicated that they found the task easy to perform with 95% of the judge indicating their level of understanding of the task was more than acceptable (NCS, 1995). While 80% of the judges indicated their conception of minimally

competent performance was more than moderately well formed (91% for clearly competent performance), the same question was not asked relative to conceptions of needs instructional intervention, may or may not need instructional intervention, and does not need instructional intervention. Because judges were not asked to define these concepts, and to reach group consensus on their meaning, one must assume that judges tended to use their own conceptions of these levels of performance.

In terms of physical/time demands on judges, the contrasting groups methods was less burdensome than the NAEP paper selection method. Both methods basically required judges to sort papers into three piles, a similar tasks requiring similar amounts of exertion. The contrasting groups methods, however did not require the extra step of selecting a representative paper from each of the three piles. When asked about the amount of time allocated for the process, 86% of the EWT judges indicated they had sufficient time for the task. (NCS, 1995). Judges did not appear to be fatigued following this part of the standard-setting process, and there was not evidence of a fatigue effect in the data from judges sorting the papers (author's recollection).

In general, the contrasting groups method used for the EWT appears to have slight edge over the paper selection method in terms of ease of administration.

Judges' comfort with the final decision rule. Unfortunately, the panelists' evaluation questionnaires from the NAEP and EWT standard-setting project did not ask judges about the standards based solely on the constructed-response items, but both

questionnaires asked judges about the defensibility and the reasonableness of the overall standards. In lieu of data about panelists comfort with the constructed-response standards, panelists' comfort with the final standards, which combined the dichotomous items' and constructed response items' standards, will be used. It is not unreasonable to assume that if panelists were not comfortable in their opinions of the reasonableness and defensibility of the final standards.

Both evaluation questionnaires asked judges whether they felt the standard-setting "study" had produced recommended standards that were defensible, and that would be considered reasonable. Both questionnaires used a 5 point Likert-scale which ranged from "not at all" (coded 1) to "to a great extent" (coded 5). Table 3 displays the results.

Table 3

Are Standards From NAEP and EWT Defensible and Reasonable?

	NAEP RATING %						EWT RATING %					
	5	4	3	2	1	Mean	5	4	3	2	1	Mean
Defensible	40	45	5	10	0	4.15	62	36	01	0	0	4.61
Reasonable	55	40	5	0	0	4.50	66	34	0	0	0	4.66

* From Act, 1993 (August) and NCS, 1995 (February)

From the data in Table 3 it appears that judges who participated in both the standard-setting projects felt very comfortable with the final standards produced by the standard-setting studies.

Judges' confidence in the results. Again neither evaluation

questionnaire contained questions about judges' confidence in the results from the constructed-response items, but both questionnaires asked judges questions that can be used to gauge their confidence in the overall results.

In addition to asking judges directly about whether the standards were defensible and reasonable, the NAEP questionnaire asked judges to indicate their level of confidence in the achievement level ratings they provided. Judges indicated a high level of confidence with 90% saying they were more than somewhat confident. In addition, 100% of the judges indicated a willingness to sign a statement recommending use of the final standards (ACT,1993, August).

For the EWT, judges were asked to indicate to what extent the standard-setting study provided an opportunity to use their best judgement in recommending standards. 98% of the judges indicated that they were able to use their best judgment more than to "some extent". In addition, 100% of the judges indicated that the New Jersey Department of Education could include their name and organization in a listing of standard-setting judges that would be included in the final report. While not asking about confidence in the results directly, these responses would seem to indicate a high level of satisfaction, with and possibly confidence in the results.

Conclusions and Recommendations

It is always problematic to compare the results of non-parallel studies, yet this paper has attempted to do just than. While the limitations of the paper were stated up front, it is incumbent upon the authors to remind the reader at this point that

differences in the two assessments themselves, and in the design of the two projects, may render all the comparisons made above suspect.

Based upon the evidence cited earlier, it appears as though the contrasting groups method may yield results that are more accurate than those from the paper selection method, or at least results that 1) are closer to those produced by the modified Angoff method and 2) have better intrajudge consistency. To verify this tentative conclusion a study that used both methods on the same assessment, and with the same set of judges, would be useful.

Both methods appear to be relatively easy for judges to cope with from a cognitive and physical standpoint, although the shear volume of work required by the NAEP process (judges reviewing 144-168 papers compared to only 80 for the EWT) led to some fatigue and fatigue related selection problems for those judges. If the number of prompts and papers had been equal, there may have been little or no difference observed in the ease of administration of the two methods.

Panelists' comfort with, and confidence in, the results of the standard-setting processes used were both very high, with neither the NAEP or EWT process showing a clear advantage. Questions related to these factors, however, did address the overall process and not the constructed-response methods specifically, so results here may not reflect judges' opinions of the paper selection and contrasting groups methods directly.

Obviously, this paper does not provide a definitive answer to the question "which method is better? "Some of our findings

contradict some of the findings related to the contrasting groups method from Reckase's (1994) simulation study, but then Reckase's contrasting group method differed from that used by NCS and New Jersey Department of Education. Clearly, more research is needed on methods of setting standards on constructed response items.

REFERENCES

1. American College Testing, 1991 (November). Design document setting achievement levels on the 1992 NAEP in Reading, Writing, and Mathematics. Iowa City, IA: Author.

2. American College Testing, 1992 (March). Results of pilot study for achievement levels setting. Technical report for the National Assessment Governing Board. Iowa City, IA: Author.

3. American College Testing, 1993 (August). Setting achievement levels on the 1992 National Assessment of Educational Progress in Reading: A final report. Iowa City, IA: Author.

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

5. Hambleton, R. and Plake, B. (1994) April. Using an extended Angoff procedure to set standards on complex performance assessments. Pace at the American Educational Research Association annual meeting, New Orleans.
   - there may be only one previous large-scale initiative to set students on PAs which as been well-documented in the measurement lecture.

6. Jaeger, R. (1991). Selection of judges for standard-setting. Educational Measurement - Issues and Practices, 10,(2), 3-14.

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

8. Luecht, R. M., 1993 (March). NAGB round 4 reading results. Unpublished technical report. Iowa City, IA: Author.

9. Luecht, R. M., 1993 (April). Using IRT to improve the standard setting process for dichotomous and polytomous items. Paper at the Annual Meeting of the National Council on Measurement in Education. Atlanta, Ga.

10. National Academy of Education, 1993 (July). Setting performance standards for student achievement. A report of the NAE Panel on the Evaluation of the 1992 Achievement Levels. Boulder, CO.: Author.

11. National Computer Systems, 1995 (February). New Jersey standard-setting study: Report of Activities. Final report on the New Jersey Grade 8 Early Warning Test (EWT) in Reading, Writing, and Mathematics. Eden Prairie, MN. Author.

REFERENCES (Continued)

12. Plake, B. (1994) April. An integration and reprise: What we think we have learned. Paper at the American Educational Research Association annual meeting, New Orleans.

13. Reckase, M. (1994) June. Standard setting on performance assessments: A comparison between the paper selection method and the contrasting groups method. Paper at the National Conference on Large Scale Assessment sponsored by the Council of Chief State School Officers, Albuquerque, NM.

14. Reid, J. (1991). Training judges to generate standard-setting data. Educational Measurement:classes and Practices, 10(2), 11-14.