DOCUMENT RESUME

ED 372 113 TM 021 940

AUTHOR Hambleton, Ronald K.; Plake, Barbara S.
TITLE Using an Extended Angoff Procedure To Set Standards on Complex Performance Assessments.
PUB DATE Apr 94
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Assessment; Evaluation Methods; *Interrater Reliability; Pilot Projects; *Scores; Scoring; *Secondary School Teachers; *Standards; Test Construction
IDENTIFIERS *Angoff Methods; Performance Based Evaluation; Polytomous Scoring; *Standard Setting

ABSTRACT

The number of performance-based assessments is increasing rapidly, but to date there is no established procedure for setting standards on these assessments. This paper describes several extensions to the Angoff procedure to accommodate the characteristics of a performance-based assessment and presents the results of research in applying this procedure. The extensions included a revised task for panelists that involved the specification of expected scores for just barely certifiable candidates on polytomously scored exercises and weights to reflect the relative importance of scoring dimensions at the exercise level and of the exercises themselves. Results obtained from using the new procedure with 12 panelists (teachers of adolescents with experience in pilot tests of the assessment) were mixed. On one hand, panelists seemed to find the new procedure straightforward to apply and were able to reach a high level of agreement among themselves about the standards. Confidence levels in the resulting standards were high. However, when given the choice, panelists definitely preferred a standard-setting procedure that was at least partly conjunctive as opposed to the extended Angoff procedure that was completely compensatory in nature. Ten tables and one figure present study findings. (Contains 7 references.) (SLD)

Using An Extended Angoff Procedure to Set Standards on Complex Performance Assessments

Ronald K. Hambleton
University of Massachusetts at Amherst

and

Barbara S. Plake
University of Nebraska-Lincoln

Abstract

The number of assessments that are performance-based is increasing rapidly but, to date, there is not an established procedure for setting standards on these assessments. The purposes of this paper are to describe several extensions to the Angoff procedure to accommodate the characteristics of a performance-based assessment and to present the results of our research to apply this new procedure. The extensions included a revised task for panelists which involved the specification of (1) expected scores for just barely certifiable candidates on polytomously-scored exercises, and (2) weights to reflect the relative importance of scoring dimensions at the exercise level and the exercises themselves.

The results obtained from the new procedure were mixed. On the one hand, the panelists seemed to find the new procedure straightforward to apply and were able to reach a high level of agreement among themselves about the standards. Confidence levels in the resulting standards were high. However, when given the choice, panelists definitely preferred a standard-setting procedure which was at least partly conjunctive as opposed to the extended Angoff procedure which was completely compensatory in its nature. Suggestions for follow-up research are offered in the paper.

5/16/94

Using An Extended Angoff Procedure to Set Standards on Complex Performance Assessments[1,2,3]

Ronald K. Hambleton
University of Massachusetts at Amherst

and

Barbara S. Plake
University of Nebraska-Lincoln

The Angoff standard-setting procedure and various modifications of it have been widely applied to multiple-choice credentialing exams (Jaeger, 1989; Shepard, 1984; Sireci & Biskin, 1993). In the absence of any established standard setting procedures for use with performance assessments, it seemed reasonable, therefore, to explore the utility of the Angoff standard-setting procedure, or at least a modification of it, as a first step in a program of research concerning standard setting on complex performance assessments such as the National Board for Professional Teaching Standards' (NBPTS) assessment packages for identifying highly accomplished teachers. There may be only one previous large-scale initiative to set standards on performance assessments which has been well-documented in the measurement literature. Readers are referred to work by the American College Testing Program and the National Assessment Governing Board in setting standards on the performance-oriented component of the 1992 NAEP in mathematics, reading, and writing (Mullis, Dossey, Owen, & Phillips, 1993; National Academy of Education, 1993).

With the typical performance assessment, candidate total scores are obtained by summing scores on a set of polytomously-scored exercises. The

---

[1]Laboratory of Psychometric and Evaluative Research Report No. 259. Amherst, MA: University of Massachusetts, School of Education.

[2]Paper presented at the meetings of AERA and NCME, New Orleans.

[3]The material contained herein is based on work supported by the National Board for Professional Teaching Standards. Any opinions, findings, conclusions and recommendations expressed herein are those of the authors and do not necessarily reflect the views of the National Board for Professional Teaching Standards.

Angoff standard-setting procedure could be extended to performance assessments by requesting standard-setting panelists to estimate the expected score of the just barely certifiable candidate on each polytomously-scored exercise. This is a simple extension of a panelist's task when applying the Angoff procedure to dichotomously-scored multiple-choice test items. That task is to assess the probabilities (i.e. expected scores) with which just barely certifiable candidates correctly answer test items. These expected exercise scores from each panelist can be added to obtain a standard of performance on the set of exercises and then the panelists' standards can be averaged to obtain the standard required for certification set by the full group. Other variations that might be included in the standard-setting process could be (1) group discussion between two sets of ratings, (2) the presentation of data to the panelists bearing on the consequences of the standards (i.e. the passing rates), and (3) asking panelists to provide the expected distribution for a group of just barely certifiable candidates across the score scale for each exercise.

In our current project with the National Board for Professional Teaching Standards to design and evaluate possible procedures for setting performance standards, possible extensions to the Angoff procedure were needed because of complexities in the performance assessment. First, five substantially different dimensions or characteristics for scoring each exercise were available. Each exercise was scored on three of these dimensions, though a different three (in general) were used with each exercise (Plake, 1994). Setting a performance standard on each exercise, using some procedure akin to Angoff's procedure, therefore, involved setting a standard of performance on each dimension for an exercise and then summing these standards to obtain a standard for the exercise. Second, the dimension scores on each exercise may not necessarily be equally important in the panelists' minds, nor would the exercises be necessarily equally important in compiling assessment package scores. There is always the possibility that panelists would want to differentially weight the dimensions and exercises in producing assessment package scores

and making credentialling decisions. Therefore, the Angoff procedure needed to be extended to allow panelists to weight (1) the dimensions used in scoring each exercise and (2) the exercises themselves, to reflect their views of the relative importance of the dimensions and exercises in obtaining assessment package scores. Third, the scoring of the exercises was quite complex and therefore panelists needed substantial training with the scoring protocols prior to setting standards. To minimize any confusion resulting from the complex scoring, it seemed best, therefore, to ask panelists to work through a process for setting standards on each exercise, prior to receiving training and setting standards on other exercises.

The purposes of this paper are to describe the extensions made to the Angoff procedure to accommodate the characteristics of the performance assessment package under study and to present the results of our research to field-test this extended Angoff procedure. The research was conducted with 12 panelists working for four full days to set standards on a portion (five of the nine exercises) of the NBPTS Early Adolescence/English-Language Arts Assessment Package. Not all nine exercises were needed because of the research nature of our work. The exercises and scoring dimensions were described by Plake(1994), hence that information will not be repeated here.

Method

Overview

The Angoff standard-setting procedure involved 12 panelists becoming familiar with a definition of the just barely certifiable candidate. Then, following three hours (approximately) of training on the scoring protocol for an exercise, panelists set standards on the three dimensions used in scoring an exercise and indicated their level of confidence in the standards they set. Following a group discussion of the panelists' standards, panelists provided a second set of standards along with new confidence ratings. Next, the panelists specified relative weightings of the dimensions to reflect their personal views about the best way to weight the dimensions for obtaining the standard on the exercise. Again, these

relative weightings were discussed and then a final set of relative weightings were provided by the panelists along with their confidence ratings. Finally, the panelists weighted the exercises in terms of their views of the relative importance of the exercises, discussed their ratings, and then provided final ratings of the relative weightings of the exercises. From the complete set of standards and relative weightings, a standard from each panelist for the assessment package was obtained, and then the standards were averaged to obtain the final standard.

In addition, panelists completed evaluation forms at the end of the rating process for exercises one, three, and five, participated in a post standard-setting focus group, and provided data on their preferences for various models of determining certifiability (e.g. compensatory, conjunctive, and combination).

Panelists

The 12 panelists who participated in the study were teachers of adolescents who lived in the Tampa Bay area and who had experience as assessors in earlier pilot tests of the Early Adolescence/English Language Arts assessment package prepared by the University of Pittsburgh. No panelists had exposure to more than two of the five exercises prior to the standard-setting experience.

Implementation of the Extended Angoff Procedure

The specific steps in applying the extended Angoff procedure for setting a standard on five exercises from the Early Adolescence/English Language Arts Assessment Package were as follows:

1. Panelists were provided with a three page generic description of the highly accomplished teacher prepared by the National Board for Professional Teaching Standards. Panelists discussed the definition of the "just barely certifiable" (JBC) candidate until they felt the definition was clear enough for them to consider the performance of the JBC candidate on each exercise. Panelists were encouraged to apply this definition in setting their standards. Basically, they needed first to adapt this definition to the highly accomplished Early Adolescence/English-Language Arts Teacher and second, they

needed to conceptualize the teacher who just barely met this definition. After about 30 minutes of discussion, panelists indicated that they felt comfortable with this step and indicated their readiness to move to the next step in the process.

2. Following extensive training on the first exercise and how it was scored, and most importantly, the meaning of the four score points on each dimension, panelists provided standards (i.e. expected scores) on the three dimensions used in scoring the exercise. See Figure 1 for a listing of the dimensions used in scoring each of the exercises. The score scale for each dimension was 1 to 4, with 4 being outstanding performance. Also, panelists provided ratings of their confidence level in the standards they set.
3. Panelists were then led through a discussion of their standards, which included a presentation of the average panelist standard on each dimension, and the high and low standards on each dimension among the panelists. Panelists who rated high or low in relation to other panelists were encouraged to discuss their standards, and then other panelists joined in the discussion. When panelists felt the discussion was complete with all views being heard and discussed, the process was moved to the next step.
4. Panelists provided a second set of standards for the JBC candidate and indicated their confidence level in the standards they set.
5. Panelists were asked to provide relative weightings of the dimensions which would be used in obtaining a standard at the exercise level, as well as their confidence level in the relative weights.
6. Following the specification of relative weights of the dimensions for use in compiling an exercise standard, the panelists discussed their relative weightings. Again, at the beginning of the discussion, panelists were provided with the mean relative weight assigned to each dimension and the high and low relative weights assigned to each dimension by panelists. Panelists who rated relatively high or low in relation to other panelists were encouraged to give reasons for their weights, and then all of the panelists joined in the

discussion. The discussion continued until panelists indicated that they were ready to move to the next step in the process.

7. Panelists were asked to provide a second set of relative weights for combining their standard on each dimension to obtain an exercise standard. Again, panelists indicated their confidence level in the relative weights.

Steps 2 to 7 were repeated for each exercise. The ratings were collected on easy-to-use rating forms. When panelists provided their second set of standards, confidence levels, and relative weights, they were aware of their earlier ratings. Hence, their previous ratings were available to them when they were making changes or holding to their original set of ratings.

Once the ratings were completed on the set of five exercises, panelists were asked to provide relative weights for the exercises in compiling total assessment package scores and to provide their confidence level in the relative weightings. Following a group discussion of these weights (again, the average ratings and high and low ratings were shared with the panelists to begin the discussion), panelists provided a second set of relative weights.

Following the completion of ratings on exercises 1, 3, and 5, panelists completed evaluation forms which were intended to monitor their impressions of the process and to provide a basis for making mid-course corrections in the process, if necessary. Also, panelists participated in a focus group discussion following the completion of the standard-setting process. This final evaluation was intended to provide in-depth information about the standard setting process.

Results

Analysis of the Panelists' Ratings

Table 1 summarizes the panelists' standards on the three scoring dimensions with each exercise on the two rating occasions (before and after group discussion of the standards) and associated confidence levels. Several observations can be made: (1) variability among the panelists

dropped considerably on the second set of ratings (apparently the group discussion was influential in bringing the panel closer in agreement in their perceptions of the standards and where they should be set), (2) panelists were considerably more confident about their second set of standards, and (3) nearly all standards following the group discussion were "3," regardless of dimension or exercise.

Table 2 summarizes the relative weights attached by the panelists to the scoring dimensions for each exercise for the purpose of producing standards at the exercise level. By and large panelists attached equal weights to the dimensions, and the group discussion had the effect of bring the panelists even closer together in their assignment of relative weights.

Table 3 summarizes the confidence levels of the panelists with respect to the relative weights of the dimensions at the exercise level. Panelists varied quite a lot in their confidence levels even for the second set of ratings. Clearly though, panelists' expressed more confidence in their second set of ratings, especially for exercises 1, 2, and 5.

For each exercise, panelists set a standard of performance on each dimension and assigned weights to the dimensions to reflect the relative importance of the dimensions in obtaining an exercise standard. (For example, suppose a panelist assigned standards of 3, 4, and 2 to the dimensions and assigned .25, .35, and .40 as the relative weights, respectively, then the exercise standard would be 3x.25 + 4x.35 + 2x.40 = 2.95.). Table 4 contains a summary of the exercise standards for the two sets of ratings. Again, movement of the exercise standards toward a score of 3.0 can be seen along with increased agreement among the panelists with respect to their exercise standards. In fact, the panelists' exercise standards are all within .07 of 3.0 for the second set of ratings and the standard deviation of the standards at the exercise level across panelists was never greater than .20.

Panelists were also asked to provide relative weights of the exercises for setting a standard on the assessment package. Table 5 summarizes the findings. As with the dimensions at the exercise level, the panelists consistently assigned near equal weights to the exercises. Also,

the first and second sets of mean relative weights of the exercises were near identical. Variability among the panelists was slightly lower on the second set of ratings. Confidence ratings on the exercise weights are summarized in Table 6. Panelists expressed a very high level of confidence in their second set of ratings.

Using the relative weights assigned by panelists to the exercises, it was possible then to compute the assessment package standards for the panelists. This was done by attaching the relative weights to the exercise standards and summing. Table 7 summarizes the findings. The panelists finished the process by establishing the standard as 3.0 with a high level of agreement among panelists (the standard deviation of the standards on the second set of ratings was 0.08). In terms of a total assessment package score, a score of 45 or higher out of a possible score of 60 would be needed for certification. The standard deviation of standards for the second set of ratings among the panelists was 0.96 points on the total assessment package score scale (which ranges from 15 to 60).

Evaluation Data

At the end of the standard-setting exercise, panelists were asked to complete a questionnaire which addressed their perceptions of the process itself, and their confidence in the resulting standards. Tables 8 and contain the panelists' evaluative data of the extended Angoff procedure. These results seemed very encouraging:

1. 67% of the panelists expressed confidence in the extended Angoff procedure producing a suitable standard.
2. 83% of the panelists expressed confidence in the actual standard for the assessment package. (Though they hadn't been informed during the process about what that standard would be, it was surely obvious to panelists because of the feedback between sets of ratings that the standard would be close to 3.0)
3. 67% of the panelists expressed confidence in their relative weightings of the dimensions and exercises.
4. In no instance did a panelist indicate low confidence in any aspect of the process.

And, with respect to the actual process itself, in every aspect, (1) the presentation and discussion of the definition of the highly accomplished teacher, (2) instructional materials about the exercises and scoring, (3) training, and (4) providing standards and relative weights, panelists rated the process as successful or very successful. There were no ratings in the categories of "somewhat successful" or "not successful".

In addition to evaluative questions, the panelists were asked a couple of questions relating to their views about standards for credentialing candidates. Table 10 contains the data. Some of these data are surprising as well as troublesome because they are inconsistent with the standard set using the extended Angoff procedure. To question 1, 10 of the 12 panelists (or 83.3%) indicated that they felt that there were certain exercises a candidate should pass to be certified. Unfortunately, the extended Angoff procedure does not produce a standard with that feature. A standard set with the extended Angoff procedure is essentially "compensatory." Candidates are certified as long as their total assessment package scores are equal to or above the standard set by the panelists. Thus, with a standard of 45, there are many ways of achieving that score including several where a candidate may do relatively well on one or two exercises and relatively poorly on the others.

From question 2, reported in Table 10, the panelists' preferences are clear. All panelists who answered this question (10 of 12) would require that exercises 1 and 5 (Student Learning, and Planning and Teaching) be passed for Board certification. In fact, most of the panelists (83.3%) would require that all of the exercises be passed for Board certification. Unfortunately, however, this standard-setting policy is not consistent with the policy that is obtained from the extended Angoff procedure. In fact, none of the panelists supported the use of a overall performance score as the standard (see question 4a) which is exactly the type of standard produced by the extended Angoff procedure. What the panelists wanted, and they differed widely in what type of credentialing policy they wanted (see the results to question 4 in Table 10), were various combinations of disjunctive, conjunctive and compensatory standard-setting policies.

Conclusions

In several respects the results from applying the extended Angoff procedure to complex performance assessments were encouraging: (1) The procedure itself required a modest amount of training (though, of course, an extensive amount of time was needed to familiarize panelists with the exercises and scoring protocols), (2) a two step judgmental process (with discussion in between) resulted in high agreement among the panelists-clearly the two step process was preferable to a one-step process only, and (3) panelists indicated a high level of confidence in the procedure and the results. With many performance assessments, such findings would be highly supportive of the resulting standards.

But there was at least one troublesome finding in this study which calls into question the validity of the resulting standard for the (five-exercise portion) NBPTS Early Adolescence/English Language Arts assessment package. On a post standard-setting questionnaire, many panelists indicated that, given the choice, they wanted candidates to pass several of the exercises, but such a standard-setting policy is inconsistent with the type of standard set with the extended Angoff procedure. With the extended Angoff procedure, a compensatory policy is in place which means that candidates who achieve a total weighted score which equals or exceeds the standard (45 of 60 points in this study) are credentialed regardless of the actual number of exercises that they may pass. In practice, there would not even be a need to produce exercise scores, let along use them to identify "passing" and "failing" candidates. A candidate under the right circumstances could do very well on one or two of the five exercises and not so well (i.e., perform below the exercise standards) on the others and still be credentialed. There is nothing necessarily wrong with that result, except that, in this study, such a policy is inconsistent with the preferences of the panelists. The clear implication is that the panelists did not fully understand the implications of the extended Angoff procedure that they had implemented. Had they fully understood the procedure, they most likely would not have so strongly endorsed the procedure or the standard they set.

There was a second problem which surfaced during the focus group discussion following the standard-setting process. Many panelists felt uncomfortable setting minimum performance scores (i.e. JBC scores) on each dimension for each exercise, essentially, independently. They expressed a desire to look at the total set of dimension scores and exercise scores in a holistic fashion and establish multiple acceptable patterns of performance scores across the dimensions and exercises for being certified. For example, though a panelist may have set a standard of 3 on each exercise, that same panelist may have wanted to adopt complex rules for certifying candidates such as "certify a candidate who achieves a 3 on each exercise but allow the candidate to slip (below the standard) on one of the exercises." Some panelists too wanted to indicate where slippage would be allowed as part of their credentialing policy (e.g., a panelist might say "candidates can fail exercises 1, 3, or 4, but they must pass exercises 2 and 5 because those two are the most important."). Clearly, the extended Angoff procedure was not capable of providing the flexibility the panelists wanted in setting standards on the assessment package.

In sum, the extended Angoff procedure as used in this study probably produced standards that were not completely in line with the panelists' preferences. Standard-setting procedures which would allow panelists to rate full or partial score profiles as certifiable or not certifiable may hold more promise for the Early Adolescence/English Language Arts assessment package and other complex performance assessments. Such a strategy was field tested by others working on the project with encouraging results (Putnam, Pence, & Jaeger, 1994).

It should be noted that the extended Angoff procedure as implemented in this study could have been modified to some extent to make it more acceptable to the panelists. For example, the extended Angoff procedure could have been used to set standards on the exercises only. Then, another procedure could be implemented by the panel to assist them in determining which exercises would need to be passed, or which combinations of exercises would need to be passed, or even which combination of exercises would need to be passed combined with an overall assessment package score at or above

some specified level. Such a procedure might be considered in the future to produce a policy for certifying candidates which was "compensatory" at the exercise level but which might be quite complex across exercises (e.g., pass exercises 1 and 3, pass at least one of the remaining three exercises, and receive an overall assessment package score of 45).

Finally, perhaps the point should be made that panelists working with other performance assessments may be quite comfortable with a compensatory policy of standard setting. If so, the results of the current study show that it is possible to carry out the extended Angoff procedure and achieve a high level of agreement among the panelists, even after only two sets of ratings. Still, as with any standard-setting procedure, both internal and external evidence should be compiled to support the validity of the resulting standards before they are used to certify candidates.

References

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

Mullis, I. V. S., Dosser J. A., Owen, E. H., & Phillips, G. W. (1993). NAEP 1992 Mathematics Report Card for the Nation and the States (Report No. 23-ST02). Washington, DC: U.S. Department of Education.

National Academy of Education. (1993). Setting performance standards for student achievement. Washington, DC: Author.

Plake, B. S. (1994). The performance domain and the structure of the decision space. Paper presented at the meetings of AERA and NCME, New Orleans.

Putnam, S., Pence, P., & Jaeger, R. M. (1994, April). A multi-stage dominant profile method for setting standards on complex performance assessments. Paper presented at the meetings of AERA and NCME, New Orleans.

Shepard, L. (1984). Setting performance standards. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 169-198). Baltimore, MD: The Johns Hopkins University Press.

Sireci, S. G., & Biskin, B. H. (1993). Measurement practices in national licensing examination programs: a survey. CLEAR Exam Review, 21-25.

Table 1

Summary of Panelists' Extended-Angoff Ratings
(N=12)

Exercise	Dimension	First Rating				Second Rating			
		Standard[a]		Confidence[b]		Standard		Confidence	
		$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
1	A	3.17	0.58	1.50	0.52	3.00	0.43	1.17	0.39
	C	2.75	0.45	1.67	0.65	3.00	0.00	1.33	0.78
	E	3.08	0.29	1.67	0.65	3.00	0.00	1.17	0.39
2	A	3.00	0.00	1.25	0.45	3.00	0.00	1.25	0.45
	C	3.08	0.29	1.25	0.45	3.08	0.29	1.25	0.45
	E	3.00	0.43	1.42	0.67	2.92	0.29	1.25	0.45
3	A	3.25	0.45	1.42	0.51	3.08	0.39	1.25	0.62
	C	3.00	0.00	1.42	0.51	3.00	0.00	1.17	0.39
	E	3.08	0.29	1.25	0.45	3.08	0.29	1.17	0.58
4	A	3.00	0.00	1.25	0.45	3.00	0.00	1.25	0.45
	B	2.75	0.45	1.92	0.51	2.92	0.29	1.58	0.67
	E	3.00	0.00	1.33	0.65	3.00	0.00	1.25	0.62
5	B	2.92	0.29	1.67	0.65	2.92	0.29	1.42	0.67
	D	3.00	0.00	1.42	0.51	3.00	0.00	1.25	0.45
	E	2.92	0.29	1.42	0.67	2.92	0.29	1.42	0.67

[a]The score scale ranges from 1 to 4 with 4 representing exemplary practice.

[b]Confidence ratings: 1=High, 2=Medium, 3=Low

Table 2

Summary of Panelists' Relative Weightings of Dimensions to Form Exercise Scores (N=12)

Exercise	Dimension	First Rating Percentages		Second Rating Percentages	
		$\bar{X}$	SD	$\bar{X}$	SD
1	A	0.354	0.027	0.352	0.020
	C	0.294	0.038	0.300	0.022
	E	0.352	0.040	0.348	0.020
2	A	0.332	0.024	0.332	0.022
	C	0.344	0.028	0.350	0.030
	E	0.324	0.030	0.318	0.022
3	A	0.374	0.028	0.363	0.024
	C	0.310	0.019	0.320	0.014
	E	0.316	0.025	0.318	0.018
4	A	0.340	0.027	0.338	0.020
	B	0.306	0.044	0.318	0.021
	E	0.354	0.024	0.344	0.010
5	B	0.315	0.016	0.310	0.016
	D	0.336	0.026	0.365	0.024
	E	0.349	0.028	0.324	0.018

Table 3

Summary of Panelists' Confidence* in the Relative Weightings of Dimensions

Exercise	First Rating $\bar{X}$	First Rating SD	Second Rating $\bar{X}$	Second Rating SD
1	1.75	0.87	1.25	0.45
2	1.67	0.49	1.25	0.45
3	1.58	0.51	1.50	0.67
4	1.92	0.90	1.67	0.89
5	1.92	0.67	1.33	0.49

*Confidence Ratings: 1=High, 2=Medium, 3=High

Table 4

Summary of Panelists' Standards at the Exercise Level[a]
(N=12)

Exercise	First Rating $\bar{X}$	First Rating SD	Second Rating $\bar{X}$	Second Rating SD
1	3.02	0.29	3.00	0.16
2	3.03	0.17	3.01	0.14
3	3.12	0.24	3.06	0.20
4	2.92	0.14	2.98	0.09
5	2.94	0.20	2.93	0.20

[a]Sum of expected dimension scores for just barely certifiable candidates weighted by the relative weights assigned to the dimensions.

Table 5

Summary of Panelists' Relative Weightings of Exercises to Make Certification Decisions

Exercise	First Rating Percentages $\bar{X}$	First Rating Percentages SD	Second Rating Percentages $\bar{X}$	Second Rating Percentages SD
1	0.214	0.016	0.215	0.015
2	0.184	0.021	0.183	0.019
3	0.207	0.017	0.208	0.011
4	0.170	0.028	0.172	0.024
5	0.225	0.029	0.223	0.023

Table 6

Summary of Panelists' Level of Confidence
in the Relative Weightings of Exercises
(N=12)

Rating	Confidence Level			$\bar{X}$	SD
	High (1)	Medium (2)	Low (3)		
First	50%	25%	25%	1.75	0.83
Second	75%	25%	0%	1.25	0.44

Table 7

Summary of Panelists' Standards to Make Certification Decisions[1]
(N=12)

Rating	Descriptive Statistics			
	$\bar{X}^1$	SD	$\bar{X}^2$	SD
First	3.02	0.12	45.3	1.80
Second	3.00	0.08	45.0	0.96

[1]Standards reported on the 1 to 4 dimension scoring scale.

[2]Standards reported on the total assessment package score scale (15 to 60).

Table 8

Evaluation Data of the Extended-Angoff Procedure
(N=12)

Question	Panelist Responses (Percentage)
3. How confident are you that the Extended-Angoff Procedure will produce a suitable NBPTS standard?	
a. Very confident	8%
b. Confident	58%
c. Somewhat confident	33%
d. Not confident	0%
7. What level of confidence do you have in your final expected score ratings with the Extended-Angoff procedure?	
a. Very high	25%
b. High	58%
c. Medium	17%
d. Low	0%
8. What level of confidence do you have in your final weightings of dimensions and exercises with the Extended-Angoff procedure?	
a. Very high	25%
b. High	42%
c. Medium	33%
d. Low	0%

Table 9

Summary of Panelists' Evaluation of the Extended-Angoff Procedure
(N=12)

Question	- Rating -			
	Very Successful	Successful	Somewhat Successful	Not Successful
1. Rate the level of success of various aspects of the study:				
a. Presentation and discussion of the definition of the Highly Accomplished Teacher.	58	42	0	0
b. Instructional materials.	83	17	0	0
c. Training on the procedures.	83	17	0	0
d. Providing expected scores on the dimensions associated with each exercise.	75	25	0	0
e. Providing dimension weights at the exercise level.	67	33	0	0
f. Providing exercise weights to form an Assessment Battery score.	58	42	0	0

Table 10

Summary of Panelists' Opinions About Aggregation Models for Making Credentialing Decisions (N=12)

Question	Percentage
1. Do you feel that there are certain exercises that a candidate must pass at the Highly Accomplished level to be certified?	
a. Yes	83.3
b. No	0.0
c. Unsure	16.7
2. Identify any exercises that you feel a candidate must pass at the Highly Accomplished level to be certified, regardless of overall performance.[1]	
a. Student Learning	100.0
b. Assessment of Student Writing	90.0
c. Post Reading Interpretive Discussion	90.0
d. Instructional Analysis	80.0
e. Planning and Teaching	100.0
f. There are none.	0.0
3. Do you feel that there is a certain number of exercises that a candidate must pass at the Highly Accomplished level to be certified?	
a. any one of the assessment package exercises	0.0
b. any two of the assessment package exercises	0.0
c. any three of the assessment package exercises	0.0
d. any four of the assessment package exercises	16.7
e. any five of the assessment package exercises	83.3
4. Which combination of decision rules for certification would you favor?	
a. only achieving some overall performance score	0.0
b. only passing certain exercises	16.7
c. achieving some overall performance score plus passing certain specified exercises	16.7
d. only passing some specified number of exercises	8.3
e. achieving some overall performance score plus passing some specified number of exercises	33.3
f. other	25.0

[1]Two of the panelists omitted question 2.

Figure 1. Dimensions Scored in the Five Early Adolescence English Language Arts Exercises Examined in the Standard-setting Pilot Test

Exercise / Dimension	Dimension A Learner-Centeredness	Dimension B Cultural Awareness	Dimension C Content Knowledge	Dimension D Integrative Curriculum	Dimension E Coherent Pedagogy
Student Learning Exercise	X		X		X
Analysis of Student Writing	X		X		X
PRIDE	X		X		X
Instructional Analysis	X	X			X
Planning and Teaching		X		X	X