Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges

B H Verhoeven,[1] *A F W Van der Steeg,*[2] *A J J A Scherpbier,*[1] *A M M Muijtjens,*[3] *G M Verwijnen*[1] *& C P M van der Vleuten*[4]

Introduction Progress testing is an assessment method that samples the complete domain of knowledge that is considered pertinent to undergraduate medical education. Because of the comprehensive nature of this test, it is very difficult to set a passing score. We obtained a progress test standard using an Angoff procedure with recent graduates as judges. This paper reports on the reliability and credibility of this approach.

Methods The Angoff procedure was applied to a sample of 146 progress test items. The items were judged by a panel of eight recently graduated students. Generalizability theory was used to investigate the reliability as a function of the number of items and judges. Credibility was judged by comparing the pass/fail rates resulting from the standard arrived at by the Angoff procedure with those obtained using a relative and a fixed standard.

Results The results indicate that an acceptable error score can be achieved, yielding a precision within one percentage on the scoring scale, by using 10 judges on a full-length progress test (i.e. 250 items). The pass/fail rates associated with the Angoff standard came closest to those of the relative standard, which takes variations in test difficulty into account. A high correlation was found between item-Angoff estimates and the item *P*-values.

Conclusion The results of this study suggest that the Angoff procedure, using recently graduated students as judges, is an appropriate standard setting method for a progress test.

Keywords Undergraduate medical education, *methods; educational measurement, *methods; *problem-based learning; sensitivity; physicians.

Medical Education 1999;33:832–837

Introduction

Tests and examinations drive student learning.[1,2] For the students, the examination programme is the real curriculum.[3,4] A close match between educational objectives and assessment programme can prevent students from following a 'hidden curriculum' of undesirable assessment-based objectives. To achieve this, longitudinal final objective assessment methods have been developed in various medical schools, such as the Quarterly Profile Examination (QPE), the Progress Test (PT) and the Personal Profile Index (PPI).[5–8] These comprehensive examinations reflect the final objectives of the curriculum and sample the complete domain of knowledge that is considered pertinent to undergraduate medical education. These tests are administered periodically (e.g. 3 or 4 times per year) to all medical students regardless of their year of training. This format is intended to reinforce desirable learning behaviour in that it precludes test-directed studying, discourages students from leaving their individual learning paths, encourages functional long-term knowledge and provides feedback to which learning activities can be tailored. Research has shown that these educational objectives are generally attained.[6,7,9,10] Making pass/fail decisions, however, i.e. setting standards, has not yet been addressed properly. Because of the comprehensive nature of progress testing and because each test administration requires a different passing score for each class, setting a passing score is quite complicated. A convenient and widely used standard setting method is a normative approach, with relative cut-off scores being defined by the overall performance of each class. The advantage is that vari-

Departments: [1]Skillslab, University of Maastricht; [2]Medical student, University of Maastricht; [3]Department of Medical Informatics, University of Maastricht; [4]Department of Educational Development & Research, University of Maastricht

Correspondence: G.M. Verwijnen, Skillslab, University of Maastricht, PO Box 616, 6200 MD Maastricht, The Netherlands

ations in test difficulty are automatically corrected for, but there are also several drawbacks. Firstly, a number of students will always fail, regardless of examinees' abilities. Secondly, the heterogeneity of the student population reduces the validity of the reference group. Thirdly, examinees can deliberately influence the passing score, and, finally, the standard is not known in advance.[11,12] An absolute standard does not have these shortcomings. Its use is appropriate when mastery of content is involved and the percentage of qualified examinees is unknown.[13] This study investigates the use of an Angoff procedure for setting standards for the PT.[14] The PT is one of the main instruments used in the Maastricht medical school problem-based curriculum to assess knowledge and reinforce students' self-directed learning. Four PTs are administered annually to all students (approximately 1000), regardless of their class. The PT consists of approximately 250 true/false items of different taxonomic levels. It samples knowledge across all disciplines and content areas relevant for the medical degree. The items may include facts and figures or they may contain clinical problem vignettes.[8]

Compared with other types of tests, a PT poses two additional problems in applying an Angoff procedure. Firstly, an Angoff procedure requires the use of expert judges familiar with students' level of performance.[11,13] Usually, the judges are teachers who are experts in the subject matter being tested.[15] With progress testing, it is difficult to find credible experts for all topics to be tested and experts' familiarity with test-takers' expected levels of performance is questionable, particularly in a problem-based curriculum. We would argue that it is the students who are the only real experts, since they are the 'consumers' of the curriculum. Students are also able to conceptualize the target candidates. Therefore, we decided to use a panel of recently graduated students in an Angoff procedure for setting standards.

The second problem concerns the definition of the 'borderline' student, because borderline performance in a PT depends on how far a student has progressed through the six-year curriculum. The PT is administered to all classes four times per year, i.e. in the course of the curriculum students are assessed on 24 occasions, requiring 24 different standards. The PT was introduced at Maastricht in 1976, and since that time a vast amount of data has been collected. The data shows that the score follows a specific and stable growth pattern across the 24 measurement points. This implies that when a standard is set for one point in the curriculum, the standards for the other 24 points can be derived mathematically.[8] Since the progress test items should reflect the final objectives of the curriculum, it was decided to estimate the standard for graduation level and define borderline performance in relation to this point in the curriculum (i.e. the time of graduation).

The purpose of this study is to assess the reliability and credibility of the Angoff procedure applied to a sample of items from one PT, using recently graduated students as judges of borderline performance. Generalizability theory was used to assess reliability. To judge credibility, we examined whether the standard resulting from the Angoff procedure yielded different pass/fail rates compared with pass/fail rates resulting from both a relative and a fixed standard. In addition, we investigated the association of the Angoff item estimates with corresponding item difficulties.

Methods

Materials

Prior to administration a sample of 150 items was drawn from the total of 256 items of the PT of May 1997. Between administration and calculation of student results, 7 items were excluded from the test due to the routine quality assurance procedure.[8] Of the study sample, 146 items remained.

Judges

The eight participating experts were medical doctors who had graduated from Maastricht University 5 months before the study was conducted (range 1–10 months). They were selected on the basis of graduation date, willingness to participate and availability. Two of them were seeking employment, three were residents and three were working on their PhD dissertation. When in medical school, the selected experts had average PT scores, indicating that they did not differ significantly from the rest of their class.

Angoff procedure

To estimate the PT's passing score based on item content and difficulty, a (modified) Angoff procedure was used.[14,16] The judges were instructed not to apply a correction for guessing. They were asked to estimate for each item the probability of an imaginary borderline test-taker, at the time of graduation, knowing the correct answer. The judges were given the correct answers. However, they were not given the percentage of examinees that answered each item correctly (*P*-values) to avoid bias that might affect the assessment of the credibility of the Angoff procedure. Prior to the standard setting procedure, the judges received a letter

describing the purpose of the study, the Angoff method and instructions for the day on which the study would take place. The panel of judges was scheduled to meet from 9·00 to 17·00 h with two breaks, one in the morning and one in the afternoon. The meeting started with a plenary discussion moderated by one of the researchers to establish a working definition of 'the borderline student'. The following definition was formulated and used during the experiment: 'A borderline student is a student who spends an average amount of time studying, whose knowledge is just sufficient to pass at graduate level, but who frequently has difficulty in scoring above the cut-off score of the PTs.' Subsequently, the judges received a booklet containing the selected PT items, each item with the answer key and two blank spaces where the experts could enter their estimates. They were asked to read one item and estimate the percentage of the borderline group that would know the correct answer at the time of graduation. All estimates were entered in the booklets and written on a whiteboard. The judges with the highest and lowest estimates explained their positions, usually followed by a short debate within the panel. All judges were free to enter an adjusted estimate in the second blank space. In this way, the first 10 items were judged. The judges then made preliminary estimates for the next 10 items and the leader polled the group for their estimates and wrote these on the whiteboard. Whenever there was a discrepancy of 20 or more percentage points between any two judges, a discussion followed. Before the next item was dealt with, the judges were given the opportunity to change their estimate, whether or not a debate had taken place. This procedure was repeated for all the remaining items.

Statistical analysis

Descriptive statistics

To judge the representativeness of the item selection, scores on sampled and non-sampled items were compared. Each judge's Angoff estimates were averaged across all items to establish the passing score per judge. Mean and standard deviations were calculated. All judges' estimates (converted to a percentage scale) on all individual items (8 × 146) were averaged to establish the passing score for the test.

Reliability

Generalizability theory was used to investigate the reliability.[17–19] As all items were rated by all judges, a crossed item-by-judge design ANOVA (i × j) was used, followed by variance component estimation using the GENOVA package.[20] Since we wish to estimate the error of setting a standard for a given test, the variance of the item main effect was not included in the error variance. The Root Mean Squared Error (RMSE) was estimated and expressed accordingly:

$$RMSE = \sqrt{\frac{\hat{\sigma}_j^2}{n_j} + \frac{\hat{\sigma}_{ij}^2}{n_i n_j}}$$

Where n_i is the number of items, n_j is the number of judges and $\hat{\sigma}^2$ is the estimated variance component for the associated effect. The RMSE is an estimate of the standard error of the mean of Angoff estimates across items and judges. It indicates the error involved in the test's passing score.[18]

Credibility

Because the judges were asked not to apply a correction for guessing, the Angoff estimate can best be compared to the percentage correct score. For the item sample a percentage correct score was calculated for each of the judges and used to determine a pass or a fail for the sixth year students only. The pass/fail rate obtained by using the Angoff standard was compared with the pass/fail rates obtained with a relative standard (mean test score of the sixth year students minus one standard deviation) and a fixed standard, respectively. The fixed standard was derived from the average pass/fail rates obtained with a relative standard for past test performances across 8 years.[12] Furthermore, the Pearson correlation was calculated for the mean of the judges estimates per item and the *P*-values (percentage correct).

Results

Table 1 presents the mean test score and standard deviation of the 69 sixth year medical students that took the PT of May 1997. Descriptive statistics are presented for the total PT and the two sub-tests (sampled and non-sampled part). In addition, the reliabilities of the three tests are given.

The average student results on the total PT and the two sub-tests are comparable. These scores are also comparable to those of previous PTs on the same occasion. There seems to be a common, relatively stable score that students achieve at the end of the undergraduate medical curriculum. For a correct interpretation of the scores, one should take into account that students can use a question mark option (zero points) and that students are instructed that incorrect answers will be penalized with minus 1 (–1) point. The true correlation (corrected for attenuation) between the

Table 1 Mean progress test scores of sixth year medical students (n = 69) and reliabilities of the total progress test, set of selected and non-selected items for the Angoff procedure

	Number of questions	Mean correct score (%)	SD	Reliability*	Standardized reliability†
Total PT	249	53·0	8·7	0·90	0·90
Sampled	146	52·3	8·4	0·83	0·89
Not sampled	103	53·8	10·2	0·83	0·92

*Cronbach's alpha
†Reliability corrected for the reduced number of items by using the Spearman Brown prophecy formula; standardization is towards 249 items.[22]

sampled items and the total PT is 1·00 (observed correlation: 0·96).[21] The sample used for this study appears to be representative. The mean Angoff estimate (i.e. the average estimation of all items and judges) is 41·4% with a standard deviation of 1·7. Table 2 presents the descriptive statistics of the Angoff estimates for each judge separately.

The mean total ratings of the eight judges show little variation. However, the standard deviations per judge are quite large, implying that the Angoff estimates differ considerably across items. Table 3 presents the results of the analysis of variance and the estimated variance components.

Approximately 82% of all variance can be attributed to variation between items. Apparently, a wide range of item difficulties is found in the PT. In line with results from Table 2, the percentage of variance associated with consistent variability of judges across items is very small (0·4% of the total estimated variance). Even the overall error term is relatively small with approximately 18% of the total variance.

Table 4 reports the Root Mean Squared Errors of the test's passing score as a function of the number of judges and the number of items in the test.

The RMSE is the error of the test's passing score expressed on the original scoring scale (i.e. the percentage correct scale). With eight judges and 150 items, an RMSE of 0·59 was achieved. An approximately double (i.e. 1·96) RMSE yields a 95% confidence interval for the test's passing score (41·4% ± 1·2%). This interval is relatively small compared to the standard deviation of the test scores (8·7%). However, with normally distributed test scores, a 1% shift in the passing score changes the failure rate by approximately 2·5%. This implies that we should aim at a precision of at least 1% on the scoring scale, which corresponds to an RMSE of 0·51. This would be achieved with 10 judges, each rating 200 items or more.

Table 5 shows the percentages of sixth year students failing the PT using the standard arrived at by the Angoff estimate and the failure rates when applying a fixed and a relative standard, respectively.

Table 2 Mean and standard deviation of the Angoff estimates (146 items)

Judge	1	2	3	4	5	6	7	8
Mean (%)	41·1	41·3	42·7	40·1	42·3	38·3	41·6	43·7
SD	22·8	20·5	22·3	24·3	20·6	22·6	22·6	22·3

Table 3 Analysis of variance and estimated variance components

Source of variability (effect)	d.f. *	Sum of squares for score effects (SS)	Mean squares (MS)	Estimated variance component	SE †	Percentage of total variance
Items	145	486013·26	3351·82	407·89	48·87	81·8
Judges	7	2868·27	409·75	2·20	1·32	0·4
IJ, e	1015	90065·73	88·73	88·73	3·94	17·8

*degrees of freedom
†standard error of estimated variance component.

Table 4 Root Mean Squared Errors as a function of the number of items and the number of judges

	Number of judges									
Number of items	3	4	5	6	7	8	9	10	11	12
50	1·15	1·00	0·89	0·81	0·75	0·70	0·66	0·63	0·60	0·58
100	1·01	0·88	0·79	0·72	0·66	0·62	0·59	0·56	0·53	0·51
150	0·96	0·84	0·75	0·68	0·63	0·59	0·56	0·53	0·50	0·48
200	0·94	0·81	0·73	0·66	0·61	0·57	0·54	0·51	0·49	0·47
250	0·92	0·80	0·71	0·65	0·60	0·56	0·53	0·51	0·48	0·46
300	0·91	0·79	0·71	0·64	0·60	0·56	0·53	0·50	0·48	0·46

Table 5 Failure rates of sixth year medical students (n = 69) in the Progress Test for different standards

Standard used	Passing score (%)	Failure rate (%)
Angoff standard	41·4	7·2
Fixed standard	52·4	55·1
Relative standard	43·9	10·1

The failure rate obtained by the Angoff standard is lowest and comes closest to the failure rate of the relative standard. Finally, the correlation between item-difficulties (*P*-values) and item Angoff estimates was 0·81, indicating that the Angoff estimate does include item difficulty variation.

Discussion

The large variation in item difficulties is typical for PTs. Although the test is targeted to the final objectives of the curriculum and thus should yield high scores on all items at the final administration, this is routinely not the case and a considerable spread of item scores is found. Repeated and intensive feedback to item authors about scoring profiles of individual items and total tests, and specific instructions and workshops about item construction and test design have failed to bring about a reduction in the spread of item scores. Apparently, this is an inherent phenomenon of PTs. More than 80% of the variance could be attributed to item variance, leaving only a small portion of judge and other error variance. The small size of the judge variance is probably also partly a result of the procedure followed. The judges were allowed to revise their estimates, i.e. the individual estimates are not fully independent.

In the estimation of the error involved in the Angoff estimate across judges and items (the test's passing score), the item difficulty variance is not considered as part of the error because the students will be tested on this set of items and generalization is not towards other items. Only the (small) judge variance and overall error term (including the interaction effect between judges and items) are considered as error variance. Adding judges would considerably improve the reproducibility of the passing score. With 10 judges in the panel judging 200 items or more, an acceptable precision is reached, i.e. 1% on the scoring scale. In other words, with the normal sample size of items (approximately 250) an Angoff panel should consist of 10 judges.

Credibility was assessed by comparing the student failure rate associated with the Angoff method with the failure rates obtained with two conventional standards and by investigating the relationship between the Angoff score and item difficulty. Although different standards yield different outcomes, some credibility of the (modified) Angoff procedure can be inferred from the comparison in this study. The Angoff standard came closest to the relative standard, based on the mean and the standard deviation of this PT. Unlike the fixed standard, the relative standard takes account of variations in test difficulty. Compared to the fixed standard (52·4%), the relative standard of the test under study (43·9%) is lower, indicating that the difficulty of this test is above average. This is also supported by the same finding in the other five year groups that took the test on the same occasion. The scores obtained on previous tests by the group of sixth year students give no reason to expect this group to differ from other groups on this occasion in the curriculum. The necessity to include test difficulty in the judgement of a passing score is evident given the magnitude of item variance involved in progress testing. Credibility was also supported by the high correlation of the item Angoff estimates with the actual item scores. It shows that the Angoff standard is sensitive to item and thus to test difficulty.

In conclusion, the results of this study suggest that the Angoff procedure is an appropriate standard setting

method for a PT. The use of recently graduated students as judges appears to be justifiable. Feasibility could be a problem, since considerable resources are required in order to reach a reproducible passing score estimation. Further research could focus on the effect of minimizing the resources and logistics needed for the Angoff procedure. For instance, one might look at the use of different groups of judges judging fewer items per test or explore the reduction of the panel size by providing initial estimates by item authors and/or reviewers.

References

1 Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Med Educ* 1983;**17**:165–71.

2 Frederiksen N. The real test bias. Influences of testing on teaching and learning. *Am Psychologist* 1984;**39**:193–202.

3 Van der Vleuten C, Newble D, Case S, Holsgrove G, McCann B, McRae C *et al.* Methods of assessment in certification. In: Newble D, Jolly B, Wakeford R. eds. *The Certification and Recertification of Doctors*. Cambridge: Cambridge University Press, 1994: p. 105–25.

4 Van der Vleuten CPM. Beyond Intuition [inaugural lecture]. Maastricht: Datawyse, 1996.

5 Arnold L, Willoughby TL. The quarterly profile examination. *Acad Med* 1990;**65**:515–6.

6 Blake JM, Norman GR, Smit EKM. Report card from McMaster: student evaluation at a problem-based medical school. *Lancet* 1995;**345**:899–902.

7 Blake JM, Norman GR, Keane DR, Mueller CB, Cunnington J, Didyk N. Introducing Progress Testing in McMaster University's Problem-based Medical Curriculum: Psychometric Properties and Effect on Learning. *Acad Med* 1996;**71**:1002–7.

8 Van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Med Teacher* 1996;**18**:103–9.

9 Van Berkel HJM, Nuy HJP, Geerlings T. The influence of progress test and block tests on study behaviour. *Instructional Sci* 1995;**22**:317–33.

10 Van Til CT. Voortgang in voortgangstoetsing: Studies naar de aansluiting van de voortgangstoets op probleemgestuurd onderwijs [PhD Dissertation Maastricht University]. Wageningen: Ponsen & Looijen, 1998.

11 Norcini JJ, Shea JA. The Credibility and Comparability of Standards. *Appl Measurement Educ* 1997;**10**:39–59.

12 Muijtjens AMM, Hoogenboom RJI, Verwijnen GM, Van der Vleuten CPM. Relative or absolute standards in assessing medical knowledge using progress tests. *Adv Health Sci Educ* 1998;**3**:81–7.

13 Norcini JJ. Research on standards for professional licensure and certification examinations. *Eval Health Professions* 1994;**17**:160–77.

14 Angoff WH. Scales, Norms, and Equivalent Scores. In: Thorndike RL. eds. *Educational Measurement*. Washington DC: American Council on Education, 1971: p. 508–600.

15 Impara JC, Plake BS. Teachers' ability to estimate item difficulty: a test of the assumptions in the Angoff standard setting method. *J Educational Measurement* 1998;**35**:69–81.

16 Livingston SA, Zieky MJ. *Passing Scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service, 1982.

17 Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: John Wiley & Sons Inc., 1972.

18 Brennan RL. *Elements of Generalizability Theory*. Iowa City: ACT Publications, 1983.

19 Shavelson RJ, Webb NM, Rowley GL. Generalizability Theory. *Am Psychologist* 1989;**44**:922–32.

20 Crick JE, Brennan RL. *A general purpose analysis of variance system [computer program], Version 2.2*. Iowa City: The American College Testing Program, 1984.

21 Nunnally JC. Theory of measurement error. In: *Psychometric Theory*. 2nd edn. New York: Mc Graw-Hill Book Company, 1978: p. 190–224.

22 Crocker LM, Algina J. Procedures for estimating reliability. In: *Introduction to Classical and Modern Test Theory*. Orlando: Harcourt Brace Jovanovich College Publishers, 1986: p. 131–56.

Received 26 *October* 1998; *editorial comments to authors* 1 *March* 1999; *accepted for publication* 1 *April* 1999