Comparison of a rational and an empirical standard setting procedure for an OSCE

Anneke Kramer,[1] *Arno Muijtjens,*[2] *Koos Jansen,*[1] *Herman Düsman*[1], *Lisa Tan*[1] *& Cees van der Vleuten*[2]

Purpose Earlier studies of absolute standard setting procedures for objective structured clinical examinations (OSCEs) show inconsistent results. This study compared a rational and an empirical standard setting procedure. Reliability and credibility were examined first. The impact of a reality check was then established.

Methods The OSCE included 16 stations and was taken by trainees in their final year of postgraduate training in general practice and experienced general practitioners. A modified Angoff (independent judgements, no group discussion) with and without a reality check was used as a rational procedure. A method related to the borderline group procedure, the borderline regression (BR) method, was used as an empirical procedure. Reliability was assessed using generalisability theory. Credibility was assessed by comparing pass rates and by relating the passing scores to test difficulty.

Results The passing scores were 73·4% for the Angoff procedure without reality check (Angoff I), 66·0% for the Angoff procedure with reality check (Angoff II) and 57·6% for the BR method. The reliabilities (expressed as root mean square errors) were 2·1% for Angoffs I and II, and 0·6% for the BR method. The pass rates of the trainees and GPs were 19% and 9% for Angoff I, 66% and 46% for Angoff II, and 95% and 80% for the BR method, respectively. The correlation between test difficulty and passing score was 0·69 for Angoff I, 0·88 for Angoff II and 0·86 for the BR method.

Conclusion The BR method provides a more credible and reliable standard for an OSCE than a modified Angoff procedure. A reality check improves the credibility of the Angoff procedure but does not improve its reliability.

Keywords education medical/*standards; educational measurement; *clinical competence; reproducibility of results; Netherlands.

Medical Education 2003;37:132–139

Introduction

In objective structured clinical examinations (OSCEs), performance is measured with checklists or rating scales. When the examination is used for decisive purposes, results must be compared with a standard of adequacy. Setting the standard is a judgmental process involving an arbitrary decision of what is considered to be 'good enough'.[1–3] To guarantee that this process is defensible and controllable, several standard setting procedures have been developed.[1,3] These procedures can be divided into two categories: relative (norm-referenced) and absolute (criterion-referenced) procedures.[4] A relative standard is based on the performance of the test takers as a group (e.g. by establishing that the top 80% of examinees will pass). The absolute approach defines the passing score in terms of how many items or tasks have to be performed correctly to pass. Although relative procedures are easy to use and to explain, they have serious disadvantages. Firstly, the standard for a test will vary from group to group, depending on the ability of the group taking the test. Secondly, some examinees will pass and others will fail regardless of how correctly they perform and, as a consequence, the minimum standard of adequate performance will vary. As specific, well-defined tasks are measured in OSCEs, absolute standard setting procedures are more suitable, particularly when the results are to be used for certification.[3–5] This study therefore focuses on absolute procedures.

Among absolute procedures, a fundamental distinction can be drawn between rational (test-centred) and empirical (examinee-centred) procedures.[3,6] In a

[1]National Centre for Evaluation of Postgraduate Training in General Practice (SVUH), Utrecht, the Netherlands
[2]Department of Educational Development & Research, University of Maastricht, the Netherlands

Correspondence: Anneke W M Kramer MD, Mauritsstraat 92, 3583 HV Utrecht, the Netherlands. Tel: 00 31 30 251 8032; E-mail: Carol_ann@tiro.nl

 © Blackwell Publishing Ltd

Key learning points

For certification and licensure purposes, absolute standard setting procedures are preferable to relative procedures.

Evidence suggests considerable variability in absolute standards when different methods and judges are used.

This study examines the reliability and credibility of a rational (a modified Angoff method) and an empirical (a newly developed method related to the borderline group method) standard setting procedure.

The empirical method (the borderline regression method) provides a reliable, credible and feasible standard for an OSCE.

rational procedure, the standard is provided by expert judgement based on rational analysis of the test content. Examples of this approach are the Angoff and Ebel methods.[1] An advantage of the rational procedure is that it is founded on the content of the test; however, an accompanying risk is that standards may be set unrealistically high. Therefore, it is recommended that judges undergo a reality check by, for example, providing performance data.[1] In the empirical procedure, the standard is determined by judgement of the performance of individual candidates relative to a performance standard based on some external criterion or on overall test performance. Examples of this approach are the contrasting groups and borderline group methods.[1] The empirical procedure provides a more holistic approach to standard setting and seems therefore particularly appropriate for performance tests in which a few, relatively lengthy tasks are assessed.[6]

Both procedures have been subject to study and evidence suggests considerable variability in standards when different methods and judges are used.[3] There is also evidence that reliability varies among different methods.[3] Studies examining the impact of performance data on standards for written tests found contradictory results. Some results indicated that performance data reduced the range of judgements, while others did not find it to have any influence.[4,6] Several studies have compared rational and empirical standard setting procedures for OSCEs. Some studies demonstrate higher pass rates for the empirical procedure,[3,7,8] while others show higher pass rates for the rational procedure.[3,9,10] Kaufman *et al.*[11] found comparable passing scores for an Angoff and a borderline group method. They also investigated the reliability of the Angoff method. Their conclusion was that a defensible and feasible passing score could be established using both methods, but that a large number of judges or stations would be required to obtain acceptable reliability for the Angoff procedure.

The need for more research into rational and empirical standard setting procedures for OSCEs is obvious. In particular, issues concerning the impact of a reality check and reliability require further investigation. Therefore this study focused on the following research questions:

1 What levels of reliability are shown by rational and empirical standard setting procedures in an OSCE and what is the impact of a reality check on the reliability of the rational procedure?
2 What is the credibility of both procedures and what is the effect of a reality check on the credibility of the rational procedure?

We chose to use a modified Angoff method for the rational procedure. It was modified so that, unlike the typical Angoff procedure, no group discussion followed by a second judgement procedure was applied. The empirical procedure consisted of a method related to the borderline group method and will be referred to as the borderline regression (BR) method. In the BR method, OSCE examiners rate clinical performance on a global rating scale. Checklist scores are subsequently regressed on the global ratings. The resulting equation is then used to calculate the checklist passing score. Woehr *et al.* described a similar procedure using the relationship between test scores and criterion performance data to arrive at a passing score.[12] The advantage of the BR method is that it uses the complete rating scale rather than the classes of performance used by the borderline group or contrasting groups methods. Reliability of the Angoff method was assessed using generalisability theory. A new procedure was developed to examine the reliability of the BR method. The procedure was based on a division of the available data into a number of random subsamples that were treated as representations of a group facet in a generalisability study. Credibility was judged by comparing the pass/fail rates of the different methods for a group of trainees in general practice and a reference group of experienced general practitioners (GPs). We considered that if a passing score was to be credible, a majority of GPs should pass (e.g. 70%). Because test procedures, including OSCEs, differ in levels of difficulty, we also related station passing scores derived from both methods to station difficulty. In our opinion, a procedure that is sensitive to station difficulty is more credible.

Methods

The OSCE

The OSCE was taken by 86 trainees in their final (third) year of training, randomly selected from the eight Dutch postgraduate training institutes for general practice. In addition, 35 GPs who also serve as trainers volunteered to act as the reference group for the examination.

The OSCE included 16 stations, representing a cross-section of the domain of clinical skills in general practice. Eight stations involved manikins, while eight involved standardised patients. Seven-minute stations were scheduled for seeing a patient with impaired vision, for insertion of a naso-gastric tube, for resuscitation, for urinary catheterisation, for a patient with painful micturition, for intravenous cannulation, for injection into the shoulder and for laboratory investigation of vaginal discharge. Fifteen-minute stations were scheduled for seeing a patient with a painful knee, a patient with a fish bone in the throat, a patient with backache, for testing pulmonary reversibility for diagnosing asthma, for physical examination of the female breast, for wound suturing with anaesthesia, for applying a compression bandage for a venous leg ulcer and for annual control of a patient with diabetes mellitus type 2.

Eighty-four examiners were involved, all of whom were experienced GPs as well as teachers at the postgraduate training institutes. Most of them were familiar with the OSCE format. Each examiner was trained in the rating of two different stations. The same examiners rated the same two stations. The purpose of the training was to reach a consensus in scoring between raters. A task-specific checklist, developed from national guidelines for general practice, was used to score clinical performance.[13,14] In addition to the checklist, performance was assessed using a 10-point global rating scale.[8,10] The examiners were made aware that a global rating below 5·5 represented inadequate performance. The checklist score of a station was defined by the percentage of correctly performed items on the checklist. The total test score was calculated by averaging the 16 station scores. The global rating was only used for standard setting purposes.

Using the checklist score, the norm oriented reliability of the test was 0·66 (generalisability coefficient) and the domain oriented reliability was 0·59 (dependability coefficient).

The judges

The OSCE examiners served as judges for the standard setting procedures. They participated on a voluntary basis.

The rational standard setting procedure

A modified Angoff procedure was applied to estimate the passing score on item content.[1] In order to include a reality check, the procedure was performed twice: once before the rating training (Angoff I) and once directly after the examination (Angoff II). Judgements were made individually using detailed written instructions. For purposes of efficiency and feasibility, no opportunity for discussion and no adjusted second judgements were included in the procedure. The judges were asked to estimate for each item on the checklist which proportion of the borderline candidates would correctly perform this item. The borderline candidate was defined as a candidate who performs at a level between pass and fail. The Angoff passing score was calculated for each station by averaging the estimates across judges and items. The Angoff passing score of the total test was defined by averaging the 16 station passing scores.

The empirical standard setting procedure

The BR method was applied to establish a passing score based on the empirical approach. By using the global ratings of the overall performance (on the 10-point scale), the pass/fail borderline was defined at 5·5 on the scale (5·5 is the traditional borderline for the Dutch marking system). The corresponding passing score per station on the checklist score scale was obtained by regressing the checklist scores on the global ratings, and then calculating the checklist score on the regression line for the global rating set at 5·5. The corresponding BR passing score for the total test was defined by averaging the 16 station passing scores.

Statistical analysis

Descriptive analysis

Mean OSCE checklist and global rating scores and standard deviations of trainees and GPs were calculated.

Mean Angoff I, Angoff II and BR passing scores were calculated. The results were tested for statistically significant differences using a paired *t*-test.

Reliability of the rational standard setting procedure

The reliability of the Angoff procedure was assessed using generalisability theory.[15] The test, consisting of 16 stations, was divided into eight pairs of stations and each pair was judged by a different subset of raters (judges nested within station pairs). All raters in the subset judged each of the two stations (judges were crossed with stations within a pair of stations) A

repeated generalisability analysis was performed to allow for a crossed design analysis. A crossed station-by-judge ANOVA design ($S \times \mathcal{J}$) was used for each pair of stations and variance components were estimated using the GENOVA package.[16] An overall estimate for each variance component was obtained by averaging the corresponding variance estimations across the station pairs.

Because we wished to estimate the error of setting a standard for a given test, the variance of the station main effect was not included in the passing score error variance. The root mean squared error (RMSE) of the Angoff passing score was estimated accordingly by

$$RMSE^2_{ANG} = \sigma^2_{\mathcal{J}}/n_{\mathcal{J}} + \sigma^2_{S\mathcal{J}}/(2 \times n_{\mathcal{J}}) \qquad (1)$$

where $n_{\mathcal{J}}$ is the number of judges and σ is the estimated variance component for the associated effect ($\mathcal{J}$ = judge main effect; $S\mathcal{J}$ = station–judge interaction effect). The $RMSE_{ANG}$ is an estimate of the standard error of the average Angoff passing score across judges and stations, assuming equal numbers of judges per pair of stations. Because judges were nested within station pairs, the number of station-judge combinations in the design equals $2 \times n_{\mathcal{J}}$ instead of $n_S \times n_{\mathcal{J}}$ as with a crossed design. This implies that when the number of judges is not changed, an increase of the number of stations in the test does not result in a change in the predicted precision of the passing score (i.e. the number of stations has no direct influence on the precision of the passing score). Using the variance components $\sigma_{\mathcal{J}}$ and σ_S obtained in the generalisability analysis, the error in the test's Angoff passing score was calculated according to equation 1 for several numbers of judges ($n_{\mathcal{J}}$ = 16, 48, 80, 160). These numbers of judges correspond to nested designs with 1, 4, 10 or 20 raters per station pair, respectively, in a 16-station test.

Reliability of the empirical standard setting procedure

The reliability of the BR method was also investigated. A newly developed procedure was used in order to obtain information about the accuracy of the BR passing score estimate. In this procedure, the sample of 86 trainees was split into random subsamples of, for instance, eight groups of 10 trainees per group. The BR passing score per station for each subsample was calculated as described before. The resulting 16 × 8 (station × subsample) matrix of BR passing scores was used as input for a station-by-subsample ANOVA analysis in the GENOVA package.[16] The analysis resulted in estimates of the variance components associated with stations (σ^2_s), subsamples (σ^2_G) and their interaction (σ^2_{sG}). The passing score of the test was obtained by averaging the passing scores of the stations. Accordingly, the RMSE of the passing score ($RMSE_{BR}$) was defined by

$$RMSE^2_{BR} = \sigma^2_G + \sigma^2_{SG}/n_S \qquad (2)$$

with n_S representing the number of stations in the test. Note that $RMSE_{BR}$ represents the estimation error of the BR passing score when the estimation is based on the data of a single subsample.

The random split of the original sample of 86 trainees was repeated for varying subsample sizes (seven subsamples of 12 examinees, six of 13, five of 17, four of 20 and three of 29). For each subsample size the variance components σ^2_G and σ^2_{sG} were estimated and the corresponding $RMSE_{BR}$ was calculated according to equation 2 for varying numbers of stations (n_s = 1, 4, 8, 12, 16, 20).

Thus, for each number of stations a series of six $RMSE_{BR}$ was obtained corresponding to the six subsample sizes used (10, 12, 13, 17, 20 and 29). Generally, the error variance of a parameter estimate is inversely proportional to the sample size. Using this relation, the $RMSE_{BR}$ for the original sample size (86 trainees) was calculated by extrapolating the series of six $RMSE_{BR}$ values for each number of stations.[17] The same relationship was used to calculate the $RMSE_{BR}$ for several other (hypothetical) numbers of trainees (40, 160 and 320).

Credibility

Pass/fail rates of the Angoff I, Angoff II and BR passing scores were calculated for trainees and GPs.

Pearson correlations of the station passing scores and *P*-values (mean percentage correctly performed items per station) were calculated for the Angoff I, Angoff II and BR methods.

Results

The OSCE scores of the trainees were higher than those of the GPs for both the checklist and global rating scores (68·6% versus 63·3% and 7·0 versus 6·8, respectively) (see Table 1).

Table 1 Mean OSCE checklist and global rating scores and standard deviations of trainees and general practitioners

Examinees	*n*	OSCE Checklist score Mean (SD)	OSCE Global rating score Mean (SD)
Trainees	86	68·6 (6·1)	7·0 (0·49)
General practitioners	35	63·3 (7·0)	6·8 (0·50)

Table 2 Estimated variance components of the Angoff I and Angoff II passing scores

	Angoff I			Angoff II		
Source	Estimated variance component	Standard error of estimated component	Percentage of total variance	Estimated variance component	Standard error of estimated component	Percentage of total variance
Judges (J)	0·37	14·6	0·03	0·67	15·2	0·06
Stations (S)	815·2	303·4	66·4	775·9	290·8	64·4
SJ,e	411·7	66·4	33·5	427·2	68·9	35·5

The mean Angoff I passing score of the total test was 73·4%. The mean Angoff II passing score was found to be significantly lower: 66·3% ($n = 16$, $P < 0·001$). The mean BR passing score of the total test was 57·6%. This passing score was significantly lower than both Angoff passing scores ($n = 16$, $P < 0·001$).

Table 2 shows the mean estimated variance components for Angoff I and II.

Due to practical reasons, only 51 of the 84 examiners served as judges in both Angoff procedures, with a minimum of four and a maximum of 11 judges per pair of stations. The generalisability analysis for the Angoff procedure was based on the 51 examiners that were involved in Angoff I as well as Angoff II.

For both the Angoff I and II passing scores, approximately two thirds of the total variance was attributable to variation among stations, indicating a wide range of difficulty levels across stations. This implies that the universal Angoff passing score (i.e. the true Angoff passing score) may vary considerably for tests comprising small numbers of stations. The percentage of variance associated with judges was very small, while the variance component corresponding to station–judge interaction accounted for approximately one third of the total variance. This indicates that the judge main effect in the Angoff passing scores is minor.

Table 3 presents the predicted root mean squared errors (RMSE) of the test's passing score of Angoff I and II as a function of the number of judges in the test.

The RMSE is the error of the test's passing score expressed on the scoring scale. With 48 judges (approximately the number of judges who participated in both Angoff procedures), an RMSE of 2·1% on the scoring scale was achieved for Angoff I and II. Multiplying the RMSE by 1·96 yields a 95% confidence interval for the test's passing score. For the passing score found in our OSCE, this corresponds to an interval of 73·4% ± 4·1% for Angoff I and an interval of 66·0% ± 4·1% for Angoff II. This results in pass rates for trainees ranging from 9·3% to 45·3% for Angoff I and from 41·9% to 83·7% for Angoff II.

Table 4 shows the RMSEs of the test's passing score of the BR as a function of the number of trainees and stations in the test.

With 16 stations and 86 trainees, an RMSE of 0·6 was achieved. This yields a confidence interval of 57·6% ± 1·2%, resulting in pass rates for trainees ranging from 91·9% to 97·7%.

Table 5 presents the percentages of trainees and GPs passing the examination when each of the three passing scores is applied.

Table 3 Root mean squared error of the Angoff I and Angoff II passing scores as a function of the number of judges

Number of judges	Angoff I	Angoff II
16	3·6	3·7
48	2·1	2·1
80	1·6	1·6
160	1·1	1·2

Table 4 Root mean squared error of the BR passing scores as a function of the number of trainees and stations

	Number of stations					
Number of trainees	1	4	8	12	16	20
40	3·3	1·7	1·2	1·0	0·9	0·8
86	2·3	1·2	0·8	0·7	0·6	0·6
160	1·7	0·9	0·6	0·5	0·5	0·4
320	1·2	0·6	0·4	0·4	0·3	0·3

Table 5 Pass rates of trainees and general practitioners (GPs) for the Angoff I, Angoff II and BR passing scores

Procedure	Passing score %	Pass rate	
		Trainees ($n = 86$) %	GPs ($n = 35$) %
Angoff I	73·4	18·6	8·6
Angoff II	66·3	66·3	45·7
BR	57·6	95·3	80·0

The highest pass rate was obtained by the BR method. The lowest pass rate was obtained by the Angoff I method, where only 9% of the GPs and 19% of the trainees would have passed.

The correlation of station difficulties (*P*-values) and station passing scores was 0·69 for Angoff I, 0·88 for Angoff II and 0·86 for the BR method.

Discussion

This study concerns issues of reliability and credibility for a modified Angoff procedure and an empirical procedure related to the borderline group method, the borderline regression (BR) method.

Angoff procedure

This procedure, which is based on the content of the test and the concept of the hypothetical borderline candidate, has a tendency to set the standard too high.[1, 4,18] We therefore applied the procedure with and without a reality check (respectively, Angoff II and Angoff I). To obtain a feasible procedure, we had to deviate from the typical Angoff procedure by omitting the group discussion and adjusted the second judgement part of the procedure. The reality check proved to have hardly any influence on reliability. For both Angoff I and Angoff II, two thirds of the estimated variance was attributable to station variance while the estimated judge variance was very low (< 0·1%). The estimated overall variance (including the interaction between judges and stations) was one third. The station variance component was not considered part of the estimation of the error involved in the test's passing score because the procedure is aimed at estimating the passing score associated with a certain set of stations. With 80 judges (approximately the number of examiners involved in the test) the error in the passing score would still amount to 1·6% on the scoring scale. This implies a 95% confidence interval of a width equal to $2 \times 3{\cdot}2\%$. Given that with normally distributed test scores a 1% shift in the passing score changes the failure rate by approximately 2·5%, this error is too high. Using more judges would reduce the error but from a resource perspective this hardly seems feasible in regular practice.

Credibility was assessed by comparing pass rates. Only 19% of the trainees and 9% of the experienced GPs would have passed the test with Angoff I. We do not consider such high failure rates to be credible for a decisive examination and for experienced GPs. The reality check improved credibility: 66% of the trainees and 46% of the GPs would have passed. However, the failure rates are still rather high. When credibility was examined by investigating to what extent the procedure takes test difficulty into account, the station passing scores of Angoff II had a higher correlation (0·88) with the station checklist scores than did those of Angoff I (0·69). This indicates that Angoff II is more sensitive to test difficulty then Angoff I. This should not be surprising as the reality check passes station difficulty information to the judges. Given the high correlation of passing scores and station difficulty for both Angoff I and Angoff II, it is also not surprising to find a large station component in the variance component estimation (Table 1). This seems to be a natural consequence of variability in station difficulty, as is usually found in OSCEs. Apparently, this is appropriately reflected in the passing scores. From this perspective, the large station variance component in the Angoff passing score seems a desirable outcome.

Borderline regression procedure

This procedure, based on the overall judgement of the actual performance by the examiners, provides a more lenient and holistic concept of competence than the Angoff procedure.[6,8] A reality check is inherent to the method. The current study confirms this leniency because pass rates would have been 95% for trainees and 80% for GPs. This finding supports the credibility of the procedure as a vast majority of passes is expected for GPs as well as for trainees at the end of their postgraduate training. Credibility was also supported by the high correlation of the station passing scores with the station checklist scores (0·86). The error in the test's passing score has an acceptable precision level (95% confidence interval of a width equal to $2 \times 1{\cdot}2\%$) with the number of stations and trainees used in the test, implying a pass rate ranging from 92% to 98%.

Methodological considerations

Our application of the Angoff procedure, without group discussion and adjusted judgement, may lead to less precise and more variable estimates, resulting in a lower level of reliability than the typical Angoff method. However, important disadvantages of the typical application are that it is organisationally complex, time-consuming and expensive, particularly when the procedure has to be applied twice for one test, as would have been required in our study.

Different designs were applied when estimating the errors involved in the Angoff and BR passing scores. This difference is a consequence of the nature of both procedures. In the Angoff procedure the passing score is obtained by judging the content of the stations, while in the BR procedure the passing score is estimated by judging the performance level of the trainees. Consequently, the size of the passing score error in the Angoff method is determined by the number of judges, whereas in the BR method, it is determined by the number of stations and trainees. Nevertheless, both RMSEs indicate the size of the passing score error when the two standard setting methods are applied in a practically feasible manner. Hence, the results help to decide which of the two procedures is preferable, given the restricted availability of resources in regular practice.

Generalisation of the results obtained in the current study is limited because the study was based on a single test and the numbers of trainees and judges were rather small.

Comparison of the Angoff and borderline regression procedures

On the basis of the results in this study, the BR procedure seems a more credible and reliable procedure with which to set a standard for an OSCE than does the modified Angoff procedure, even when a reality check is applied to the latter.

Our findings support discussion in the literature that the Angoff procedure may be less convenient for performance-based skills testing than for written knowledge testing. In the Angoff method the proportion of correct borderline examinees has to be estimated for each item, assuming that each item is content-independent. However, questions remain as to what extent this can be true for the OSCE.[6,19,20] The BR method focuses on overall performance per station rather than performance per item.

From a cost perspective, the BR method is preferable, although our application of the Angoff method (written instruction only, individual estimation without discussion) represents a good alternative.

Conclusion

We conclude that the BR method provides a reasonable and defensible standard for an OSCE and that the method is practically feasible.

Contributors

All authors contributed to the conception and design of the study and to the analysis and interpretation of the data.

AK, AM, JJ, LT and CvdV contributed to the drafting and critical revision of the manuscript.

Acknowledgements

We would like to thank all trainees and GPs who took part in the study. We are grateful for the support of staff from the eight Dutch training institutes for postgraduate training in general practice and to the members of the national test committee for the OSCE for their contribution to the development of the test. Thanks are also due to Ijsbrand Kramer for polishing our English.

Funding

This study was initiated by the Registration Committee of Postgraduate Training in General Practice (HVRC), financially supported by the Foundation of Postgraduate Training in General Practice (SBOH) and executed by the National Centre for Evaluation of Postgraduate Training in General Practice (SVUH) (all in the Netherlands).

References

1 Livingstone SA, Zieky MJ. *Passing Scores.* Princetown: Educational Testing Service 1982.
2 Norcini J. Approaches to standard-setting for performance-based examinations. In: Harden RM, Hart IR, Mulholland H, eds. *Approaches to the Assessment of Clinical Competence Part I.* [Proceedings of the 5th Ottawa International Conference on Medical Education and Assessment; September 1–3 1992; Dundee, Scotland.] 1992;32–7.
3 Cusimano M. Standard setting in medical education. *Acad Med* 1996;**71** (Suppl. 10):S112–20.
4 Norcini JJ. Research on standards for professional licensure and certifications examinations. *Eval Health Prof* 1994;**17**:160–77.

5 Newble D, Jolly B, Wakeford RE, eds. *The Certification and Recertification of Doctors. Issues in the Assessment of Clinical Competence*. Cambridge: Cambridge University Press 1994.
6 Kane MT, Crooks TJ, Cohen AS. Designing and evaluating standard setting procedures for licensure and certification tests. *Adv Health Sci Educ* 1994;**4**:195–207.
7 Van Luijk SJ, van der Vleuten CPM. A comparison of standard setting methods applied to a performance-based test. In: Harden RM, Hart IR, Mulholland H, eds. *Approaches to the Assessment of Clinical Competence Part I*. [Proceedings of the 5th Ottawa International Conference on Medical Education and Assessment; September 1–3 1992; Dundee, Scotland.] 1992;326–30.
8 Rothman AI, Cohen R. A comparison of empirically- and rationally-defined standards for clinical skills checklists. *Acad Med* 1996;**71** (Suppl. 10):S1–3.
9 Norcini JJ, Stillman PL, Sutnick AI *et al.* Scoring and standard setting with standardised patients. *Eval Health Prof* 1993;**16**:322–32.
10 Dauphinee WD, Blackmore DE, Smee S, Rothman AI, Reznick R. Using the judgements of physician examiners in setting the standard for a national multi-centre high stakes OSCE. *Adv Health Sci Educ* 1997;**2**:201–11.
11 Kaufman DM, Mann KV, Muijtjens AMM, van der Vleuten CPM. A comparison of standard-setting procedures for an OSCE in undergraduate medical education. *Acad Med* 2000;**75**:267–71.
12 Woehr DJ, Arthur W, Fehrmann ML. An empirical comparison of cut-off score methods for content-related and criterion-related validity setting. *Educ Psy Meas* 1991;**51**:1029–39.
13 Grol RPTM. National standard setting for quality of care in general practice: attitudes of general practitioners and response to a set of standards. *Br J General Pr* 1990;**40**:361–4.
14 Jansen JJM, Tan LHC, van der Vleuten CPM, Van Luyk SJ, Rethans JJ, Grol RPTM. Assessment of competence in technical clinical skills of general practitioners. *Med Educ* 1995;**29**:247–53.
15 Brennan RL. *Elements of Generalizability Theory*. Iowa City: ACT Publications 1983.
16 Crick JE, Brennan RL. A general purpose analysis of variance system. Version 2.2. [Computer program.] Iowa City: American College Testing Program 1984.
17 Muijtjens AMM, Kramer AWM, Kaufman DM, van der Vleuten CPM. Using resampling to estimate the precision of an empirical standard setting method. *Appl Meas Educ* (in press).
18 Norcini J. The credibility and comparability of standards. *Appl Meas Educ* 1997;**10**:3959.
19 Clauser BE, Clyman GC. A contrasting-groups approach to standard setting for performance assessments of clinical skills. *Acad Med* 1994;**69** (10):42–44.
20 Ross LP, Clauser BE, Margolis MJ, Orr NA, Klass DJ. Standard setting. An expert-judgement approach to setting standards for a standardised patient examination. *Acad Med* 1996;**71** (Suppl. 10):S4–6.

Received 13 *November* 2001; *editorial comments to authors* 12 *February* 2002, 14 *June* 2002; *accepted for publication* 30 *August* 2002