Advances in Health Sciences Education (2008) 13:203–211
DOI 10.1007/s10459-006-9035-1

Is an Angoff Standard an Indication of Minimal Competence of Examinees or of Judges?

M. M. VERHEGGEN[1,*], A. M. M. MUIJTJENS[1], J. VAN OS[2] and L. W. T. SCHUWIRTH[1]

[1]*Department of Educational Development and Research, University of Maastricht, P.O. Box 616, Maastricht, 6200 MD, The Netherlands;* [2]*Department of Psychiatry and Neuropsychology, University of Maastricht, Maastricht, The Netherlands (*author for correspondence, Phone: +31-43-3885768; Fax: +31-43-3885779; E-mail: M.Verheggen@educ.unimaas.nl)*

(Received: 9 March 2006; Accepted: 7 September 2006)

Abstract. *Background:* To establish credible, defensible and acceptable passing scores for written tests is a challenge for health profession educators. Angoff procedures are often used to establish pass/fail decisions for written and performance tests. In an Angoff procedure judges' expertise and professional skills are assumed to influence their ratings of the items during standard-setting. The purpose of this study was to investigate the impact of judges' item-related knowledge on their judgement of the difficulty of items, and second, to determine the stability of differences between judges. *Method:* Thirteen judges were presented with two sets of 60 items on different occasions. They were asked to not only judge the difficulty of the items but also to answer them, without the benefit of the answer key. For each of the 120 items an Angoff estimate and an item score were obtained. The relationship between the Angoff estimate and the item score was examined by applying a regression analysis to the 60 items (Angoff estimate, score) for each judge at each occasion. *Results and conclusions:* This study shows that in standard-setting the individual judgement of the individual item is not only a reflection of the difficulty of the item but also of the inherent stringency of the judge and his/her subject-related knowledge. Considerable variation between judges in their stringency was found, and Angoff estimates were significantly affected by a judge knowing or not knowing the answer to the item. These findings stress the importance of a careful selection process of the Angoff judges when making pass/fail decisions in health professions education. They imply that judges should be selected who are not only capable of conceptualising the 'minimally competent student', but who would also be capable of answering all the items.

Key words: Angoff procedure, selection of judges, standard-setting

Introduction

When test results are used for pass/fail decisions, robust standards are needed. Previous research has shown that different standard-setting methods can produce different passing scores and that a ''gold standard'' does not exist (Downing et al., 2006). In health professions education a well-known

method to determine an absolute standard for examinations is the method described by Angoff (1971). This procedure involves estimation of the performance of borderline examines by a panel of judges. In order for the method to be robust it is often considered prerequisite that the panel of judges is familiar with the performance level of the students who take the test and that they should be credible experts in the topic being tested (Norcini and Shea, 1997). Usually, these experts are expert teachers or experienced clinicians. Originally, each judge judges the items individually, but in some modifications of the procedure a consensus meeting may be held as well (Cusimano, 1996). Typically, judges are presented the items with the correct answers.

The Angoff method has a number of obvious advantages. First, it is highly intuitive, in that it is easy to explain to stakeholders that the cut-off score was determined using a panel of experts. A second advantage is that the pass–fail score is known beforehand. A third is that pass–fail scores determined by an Angoff procedure are reasonably stable across time and across panels of judges, i.e. they have sufficient intra-panel and inter-panel reproducibility (Norcini and Shea, 1992; Norcini et al., 1987, 1988).

An important disadvantage of the method, however, is its time-consuming nature. It requires a panel of experts to spend large amounts of their valuable time to the rather tedious work of judging items. Also, if items are re-used and changed in content and/or wording they will have to be submitted to an Angoff panel again.

Since the quality of the Angoff ratings is highly dependent on the quality of the panel members, there are serious possible threats to the validity and reproducibility of Angoff scores. The reproducibility of Angoff ratings has received a substantial amount of attention in the literature. Norcini et al. (1987) found that judges gave the same mean ratings for matched sets of items when ratings were collected before, during and after a meeting. Another study reported that when subsets of judges rate subsets of items, a reliable standard was still obtained (Norcini et al., 1988). Furthermore, Norcini and Shea (1992) described that different groups of experts set the same standard for the same test material.

Williams et al. (2003) summarised in a recent review the main factors that may lower the reproducibility of judgements. Although their paper was principally aimed at judgements in a clinical setting and direct observation, some of their findings seem also pertinent to Angoff judgements. They mainly suggest ensuring that a sufficiently high number of judges are included in the procedure, that the judges are well informed about their task and that the possibility for discussion is included. The latter aims to decrease any possible variation originating from individual misconceptions of some judges.

Some studies have addressed the validity of the Angoff scores mainly from an angle of judge inconsistencies. Potential causal factors of inconsistencies

both within and between judges have been classified into three categories: (a) the background of judges, (b) the items and their contexts, and (c) the standard-setting processes (Plake et al., 1991). Concerning the judge-related factors Chang et al. (1996) showed that when judges differ in their conceptualisation of minimal competency, judgemental inconsistencies may arise. They also stated that inconsistency within judges may occur when they are unable to maintain their conceptualisations of minimal competency across items on the test.

Also, the judges' fields of expertise and professional skills are suspected to influence their ratings of the items during standard-setting (Verhoeven et al., 2002). With some exceptions, research into the influence of judges' item-related expertise on standard-setting is quite limited (Chang et al., 1996; Saunders et al., 1981). Therefore, the purpose of the current study is 2-fold. First, to investigate the impact of judges' item-related knowledge on their judgement of the difficulty of items, and second, to determine the stability of differences between judges. To this effect, judges were presented with two sets of items on different occasions without the answer key. They were then asked to not only judge the difficulty of the items but also to answer them. The following research questions were investigated:

1. Does knowing the correct answer or not lead to different estimates of the difficulty of an item in an Angoff procedure?
2. Do judges have a stable level of stringency on separate occasions?

Materials and Methods

Materials

Key-feature items were used. These items are used in the regular undergraduate medical programme for the examination of a fifth-year psychiatry rotation. From a bank containing 740 items, two comparable sets of 60 items each were drawn. Before including items in the bank, they were studied by a review committee on content, wording and relevance in order to improve the validity and reliability of the items. The questionings representing the items were in the MCQ format. Items contained 1, 2 or 3 questions.

Judges

Thirteen registrars in psychiatry pertaining to the local psychiatric rotational scheme participated in the study. Each of these judges judged the difficulty of each of the items in both sets. Participants were given 4 weeks to judge one set of items. After 6 months the second set of items was provided for judgement. The procedure was similar to a standard Angoff procedure except for the fact that the answer keys were not provided. Instead, the judges were asked to provide the correct answer. Therefore, for each of the 120 items an

item score and an Angoff estimate was obtained from all 13 judges. Prior to the procedure, all participants were thoroughly briefed about both the Angoff procedure and the purpose of the psychiatry test. All judges were knowledgeable in the content area in which they were making decisions. Participants were aware of the make-up of a minimally competent examinee as it pertains to examination content as all judges were closely involved in the regular undergraduate medical programme for the examination of a fifth-year psychiatry rotation.

Analysis

For each judge the relationship between the Angoff estimate and the items score was examined by applying a regression analysis to the 60 items (Angoff estimate, score) at each of the two occasions. The structure of the regression equations was determined by a total of 2 (sets of 60 items) $\times$ 13 (judges) yielding 2 regression equations for each judge. In the analysis the Angoff estimate was the dependent variable and the item score the predictor. Item scores varied between 0 and 1 (the minimum and maximum score) as an item could contain 1, 2 or 3 questions.

The regression line is defined by two parameters, the slope and the intercept. The slope represents the steepness of the line, and the intercept the ordinate of the line for an arbitrary position at the X-axis. Usually, this position is taken to be $X = 0$, the intercept being the ordinate of the line at the intersection of the Y-axis. However, for interpretative purposes we have preferred to define the intercept as the ordinate of the line at $X = 1$. Thus the intercept represents the mean Angoff estimate for those items to which the judge knew the correct answer, and therefore it can be interpreted as a measure for the inherent judge stringency.

The intercepts are defined as Angoff estimates and therefore are expressed as percentages (of minimally competent students who would answer the questions correctly).

The slope indicates the (average) extent to which knowledge of the correct answer affects the Angoff estimate. It is expressed as the percentage of increase of the Angoff estimate between the scores 0 and 1 (the minimum and maximum score). Each individual Angoff estimate can therefore be predicted by: intercept + (slope $\times$ item score).

Mean values of the intercepts and slopes were calculated for each of the two item sets. An analysis of variance (SPSS 12) was used to disentangle the between-judge and within-judge variance components that influence both the intercepts and the slopes. In the analysis the intercept and the slope for judge j at occasion k were decomposed according to

$$\text{Intercept}_{jk} = \text{MI} + a_j + e_{jk}; \quad \text{Slope}_{jk} = \text{MS} + b_j + d_{jk}$$

with MI the general intercept mean, a_j the judge effect, and e_{jk} the judge-occasion interaction for the intercept, and MS, b_j, and d_{jk} similar components for the slope. Then the variances of the judge effect, Var(a), and Var(b), respectively, represent the between-judge variances in intercept and slope, respectively, and Var(e) and Var(d), the corresponding within-judge variances. The consistency of the judge effect across occasions is indicated by the corresponding reliability measures, e.g. for the intercept Var(a)/(Var(a) + Var(e)). For the slope it is more meaningful to consider the distance to zero slope in order to investigate the substantiality of the effect of knowing the correct answer. Then the variance of intercept concerns MS and b_j, and the relative contribution of these components to the total variance can be expressed as [MS^2 + Var(b)] / [MS^2 + Var(b) + Var(d)].

Results

The individual values for the intercepts and the slopes of all judges on both sets are presented in Table I.

The means of the slope values for both sets were 11.35% and 7.32%. The overall mean was 9.34. The slope is an indication of the influence of judge knowledge of the correct answer on his/her Angoff estimate. This influence appears to be considerable.

The means of the intercept values for both sets were 55.74% and 53.69 %. The overall mean was 54.71. The intercept indicates the stringency of the

Table I. Intercept and slope values for all judges on both sets of items

Judges	Intercept values for item set 1	Intercept values for item set 2	Slope values for item set 1	Slope values for item set 2
1	48.02	35.12	14.43	4.02
2	48.81	56.47	10.65	12.49
3	62.46	69.45	−4.82	11.50
4	56.22	43.20	27.86	5.13
5	57.67	56.94	6.84	−1.48
6	26.71	21.53	9.42	4.61
7	69.32	73.78	11.16	5.28
8	56.25	69.28	13.89	18.03
9	55.67	58.05	6.88	1.27
10	66.97	52.08	12.09	12.43
11	69.94	68.93	23.40	10.30
12	57.92	39.27	5.78	1.46
13	48.62	53.85	10.02	10.19

judge for the items to which s/he knew the correct answer, so the higher the intercept the higher the stringency.

Although the sets of items were different in content, mean intercept and mean slope values for the two different occasions were found to be quite comparable.

More striking though is the considerable variability of intercepts (stringency) between individual judges within each set. To evaluate this variability, it was compared to the variance within judges between sets (Table II).

For judge stringency the variance between judges was rather consistent and higher than the variance within judges: the consistency of the between-judge variation amounted to 0.74 (Var(a)/[Var(a) + Var(e)]). For the effect on the Angoff estimate of knowing the correct answer the between-judge variance (1.83) was negligible compared to the within-judge variance (48.50). However, the general mean, amounting to 9.34%, was statistically significant, so for all judges on average there was a substantial effect of knowing the correct answer, but the effect did not vary between judges. The fraction of the total variance explained by the general mean (and the between-judge effect) is considerable: it amounts to 0.65.

Hence, although the within – judge variation across occasions was quite large, there was a substantial effect due to knowing the correct answer. Not surprising, also for judge stringency the general mean was found to be significantly different from 0.

Judge number 6 has a considerable influence on the results. This judge is very mild in his/her judgements in comparison to the 12 other judges. However, there is no reason to believe that this judge did not seriously complete the task or did not understand the briefing about the Angoff procedure. Judge number 6 is an example of the diversity in judges that can exist in a panel of judges. Therefore it was decided not to exclude this judge from the results. Still though, even if judge number 6 would have been excluded from the results then for stringency the general mean and the variance between judges were still significant, amounting to 47.7% ($p < 0.001$) and 56.0 ($p < 0.03$), while the error variance slightly increased to 51.6. For the effect of knowing the correct answer, after exclusion of judge 6, the general mean still was significant,

Table II. Variance of intercepts and slopes within and between judges

	General mean (%)	Variation between judges	Variation within judges across occasions
Intercept	54.71***	135.14*** (SD = 11.62)	48.63 (SD = 6.97)
Slope	9.34***	1.83 (SD = 1.35)	48.50 (SD = 6.96)

***Statistically significant ($p < 0.001$).

amounting to 9.53 ($p < 0.001$), the between-judge variance was still small, 2.14 (NS), and the error variance slightly increased to 51.6.

Discussion

A careful selection process of judges is necessary when using an Angoff procedure for standard-setting in health professions education. Previous research showed the importance of the composition of the Angoff panel (Verhoeven et al., 2002), while few studies have addressed the effect of the knowledge of the Angoff panel members. This study demonstrated the impact of judges' item-related knowledge on their judgement of the difficulty of items in an Angoff procedure.

The results showed that the individual judgement of the individual item is not only a reflection of the difficulty of the item but also of the inherent stringency of the judge and his/her subject-related knowledge. The latter findings are in accordance with the results described by Chang et al. (1996). These results may have some implications for current concepts of the Angoff procedure.

In this study, considerable variance between and within judges was observed. This is not new. Many studies have demonstrated that judges differ in their leniency and that there is a judge by item interaction. These studies also show that when enough judges and items are used, the overall pass–fail score is sufficiently reproducible (Norcini and Shea, 1992). One other non-random factor influencing the judge $\times$ item interaction seems to be whether the judge would have known the answer to the question or not. This raises the question of whether the answer key should be given or not when judges judge the items?

This decision depends on the validity of the conclusions drawn from the standards derived from this method and the ability of the method to provide consistent and precise estimates.

There are some arguments in favour for giving the answer key and some against. If whether or not the judges know the answer influences their item judgement, an Angoff procedure in which the answers are not given will lead to more judge $\times$ item interaction. Therefore, such judgements seem less generalisable and less stable. From this it seems obvious that providing the correct answers would be preferable, and would produce more generalisable estimates.

On the other hand however, students also sit the test items without the correct answers. Providing the answer key to judges who would not have known the correct answer themselves would most probably result in them underestimating the item difficulty. It will lead them to backward reason from the provided answer, and this will produce a lower estimate of item difficulty compared to reality.

Therefore, although analyses would indicate that the Angoff ratings are more reliable, they may be less valid. In other words, the judgements would be *consistently* off the mark.

An alternative solution to the problem would be to provide the items without the correct answers during the Angoff procedure. The judgements pertaining to items the judge did not answer correctly could then be adjusted using the regression methodology described in this paper. Before this could be routinely used however, replications of this study in different settings would be needed to obtain a more accurate estimate of the effect.

The findings in this study put even more stress on a careful selection process of judges when setting standards in health professions education. Health professions educators setting standards should consider selecting judges who are not only capable of conceptualising the 'minimally competent student', but who would also be capable of answering all the items. That a careful balance is needed is further underpinned by a study by Verhoeven et al. (2002), which showed that panel expertise by itself, is not sufficient for obtaining a reliable passing score. Another study by Cusimano and Rothman (2003) demonstrated that providing judges with normative data describing the performance of students is useful in obtaining a more precise standard. Recent research by Downing et al. (2006), describing how 5 different standard-setting methods can be used to establish acceptable passing scores, showed that the key to defensible standards lies not only in the choice of credible judges but also in the use of a systematic approach to collecting their judgements. Ultimately, establishing valid and reliable passing scores in health professions education remains a challenge.

References

Angoff, W.H. (1971). Scales, norms and equivalent scores. In: R.L. Thorndike (ed.), *Educational Measurement*, pp. 508–600. Washington DC: American Council on Education.

Chang, L., Dziuban, C.D. & Hynes, M.C. (1996). Does a standard reflect minimal competency of examines or judge competency?. *Applied Measurement in Education* **9**(2): 161–173.

Cusimano, M.D. (1996). Standard-setting in medical education. *Academic Medicine* **71**(10 Suppl): 112–120.

Cusimano, M.D. & Rothman, A.I. (2003). The effect of incorporating normative data into a criterion-referenced standard setting in medical education. *Academic Medicine* **78**(10 Suppl): 88–90.

Downing, S.M., Tekian, A. & Yudkowsky, R. (2006). Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teaching and Learning in Medicine* **18**(1): 50–57.

Norcini, J.J., Lipner, R.S., Langdon, L.O. & Strecker, C.A. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement* **24**(1): 56–64.

Norcini, J.J., Shea, J.A. & Ping, J.C. (1988). A note on the application of multiple matrix sampling to standard-setting. *Journal of Educational Measurement* **25**(2): 159–164.

Norcini, J.J. & Shea, J.A. (1992). The reproducibility of standards over groups and occasions. *Applied Measurement in Education* **5**: 63–72.

Norcini, J.J. & Shea, J.A. (1997). The credibility and comparibility of standards. *Applied Measurement in Education* **10**: 39–59.

Plake, B.S., Melican, G.J. & Mills, C.N. (1991). Factors influencing intrajudge consistency during standard-setting. *Educational Measurement Issues and practice* **10**(2): 15–16, 22, 25–26.

Saunders, J.C., Ryan, J.P. & Huynh, H. (1981). A comparison of two approaches to setting passing scores based on the Nedelsky procedure. *Applied Psychological Measurement* **5**: 209–217.

Verhoeven, B.H., Verwijnen, G.M., Muijtjens, A.M.M., Scherpbier, A.J.J.A. & Van der Vleuten, C.P.M. (2002). Panel expertise for an Angoff standard-setting procedure in progress testing. Item writers compared to recently graduated students. *Medical Education* **36**(9): 860–867.

Williams, R.G., Klamen, D.A. & McGaghie, W.C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine* **15**(4): 270–292.