Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 9, Number 5, March, 2004 ISSN=1531-7714

Estimating the Standard Error of the Judging in a modified-Angoff Standards Setting Procedure

Robert G. MacCann and Gordon Stanley
Board of Studies, NSW Australia

For a modified Angoff standards setting procedure, two methods of calculating the standard error of the judging were compared. The Central Limit Theorem (CLT) method is easy to calculate and uses readily available data. It estimates the variance of mean cut scores as a function of the variance of cut scores within a judging group, based on the independent judgements at Stage 1 of the process. Its theoretical drawback is that it is unable to take account of the effects of collaboration among the judges at Stages 2 and 3. The second method, an application of equipercentile (EQP) equating, relies on the selection of very large stable candidatures and the standardisation of the raw score distributions to remove effects associated with test difficulty. The standard error estimates were then empirically obtained from the mean cut score variation observed over a five year period. For practical purposes, the two methods gave reasonable agreement, with the CLT method working well for the top band, the band that attracts most public attention. For some bands in English and Mathematics, the CLT standard error was smaller than the EQP estimate, suggesting the CLT method be used with caution as an approximate guide only.

Standard setting procedures are now being widely used throughout the world to determine levels of student competence in educational programs. Reviews of such methods have been presented by Hambleton and Eignor (1980), Shepard (1980), Berk (1984), Berk (1986) and Jaeger (1993). One of the most commonly used methods is the Angoff (1971) procedure or its variants, referred to as modified Angoff procedures. This method has been extensively researched and compared to rival standard setting methods (e.g. Harasym, 1981; Livingston and Zieky, 1983; Cross, Impara, Frary and Jaeger, 1984; Hambleton and Plake, 1995; Impara and Plake, 1997; Plake, 1998; Giraud, Impara and Buckendahl, 2000; Buckendahl, Smith, Impara and Plake, 2001).

To interpret the results of the Angoff method over time, a knowledge of the standard error of the judges' decisions is useful. Each year the cut score in raw marks that defines a certain standard of achievement may vary, along with the associated percentages of students above that cut score. An important question when comparing cut scores across years is whether any difference observed may simply be attributable to natural variation in the judging, or whether it is sufficiently large to suggest a difference in the standard reached. Thus, an estimate of the standard error of the judging process is needed. This poses an important practical problem for educational systems, particularly those that are just starting to implement the process and have little data from which to observe the variation.

One simple and convenient way to estimate this standard error is through the Central Limit Theorem (e.g. see Hoel, 1961), which estimates the variance of mean cut scores as a function of the variance of the cut scores within a judging group. However, this method requires independence of the judges' decisions, which may not occur at all stages of a modified Angoff procedure (see the comments of Linn, 2003). For example in the Australian state of New South Wales, the School Certificate (SC), on which the data in this paper is based, is a mandatory examination-based Year 10 certificate which uses a modified Angoff method. In the procedure used, the independence of judges decisions only occurs at Stage 1. Given that in Stages 2 and 3 extra information is considered, it is possible that an estimate based on Stage 1 would over-estimate the error involved. That is, the Central Limit Theorem (CLT) procedure may give a conservative estimate of the standard error. The aim of this paper is to evaluate how well the CLT procedure performs when applied to the modified Angoff method employed in the New South Wales (NSW) programs. This evaluation is obtained by finding a second empirical estimate of the standard error, which it is argued, is approximately correct for the special cases in which it is employed in this paper. If the CLT method gives reasonable agreement with this second method for the special cases here, then it may prove to be a suitable method in general as an approximate estimate of the standard error involved.

The Standards Setting Procedures in New South Wales

The New South Wales educational system holds statewide tests at Year 10 for the School Certificate award. These tests are set and administered by a government body, the Board of Studies, which also sets the curricula for these programs. The Year 10 School Certificate tests to be considered here in English and Mathematics are compulsory and are attempted by approximately 80 000 students. The format of the School Certificate tests is a mixture of multiple choice and extended response items. For English, the test comprises 45 multiple choice items (worth 45 marks), some short

answer items where a few lines of response are required (worth 15 marks) and two extended response writing tasks (worth 20 marks each). For Mathematics, the test comprises 25 short items, each requiring the student to supply the answer (worth 25 marks), 55 multiple choice items (worth 55 marks), and four 5-mark items, each requiring a number of steps to be completed for full marks (worth 20 marks). While an attempt is made in the test construction process to ensure that test difficulty does does not vary too greatly from year to year, this ideal is difficult to obtain in practice and some variation in difficulty inevitably occurs. The internal consistency reliabilities of the scores, as determined by Cronbach's Alpha, are high and uniform as shown in Table 1.

Table 1: Cronbach Alpha reliabilities for School Certificate English and Mathematics

Year	English	Mathematics
1998	0.921	0.966
1999	0.920	0.963
2000	0.921	0.966
2001	0.925	0.954
2002	0.918	0.963

The standards setting procedure for the School Certicate uses a modified version of the Angoff procedure. All items in the School Certificate English and Mathematics tests were rated by the judges. For each test, the raw mark scale was obtained by simply summing the scores on each item to get a total score. The Angoff procedure gave cut scores on these total score scales. Rather than having just one cut score to indicate minimal competence, five cut scores are produced to separate the students into six performance bands, Band 6 being the highest level and Band 1 the lowest. These performance bands are defined by written descriptors which indicate the academic characteristics of a typical student in the band for that subject and which help the judges gain a mental image of the type of student they are considering when judging how well such students would score on each item. In addition to these descriptions, actual examples of students' responses at each band level from previous years have been published and promulgated. These exemplars, known as Standards Packages, have been placed on CD-ROMs and distributed to all schools in NSW. These Standards Packages are useful to teachers in formulating their internal school assessments and in training the teacher judges for the standards-setting operation.

For each course, five cut scores that separate the six bands are finally produced from the Angoff standards setting. Each cut score is then mapped to an arbitrary reporting scale scale that is the same for all School Certificate courses. The Band 6 cut score is mapped to 90, the Band 5 cut score is mapped to 80, Band 4 to 70, Band 3 to 60 and Band 2 to 50. Band 1 has mapped marks below 50. Two other mapping points are used: the maximum possible raw mark score is mapped to 100, and zero is mapped to zero. All scores falling between pairs of anchor points are mapped by linear interpolation.

Typically, six to nine experienced teacher judges are used in a subject. These are assisted by Field Officers, staff hired to observe the operation, to aid the judges in interpreting the output and to ensure that the judges have all the materials they need for the task. After being appointed, the judges and Field Officers undertake an intensive training exercise involving a practical simulation of the standards-setting procedure. During this training they gain additional familiarity with the six levels of performance, as defined by the exemplars in the Standards Packages.

The standards-setting procedures takes place during the supervised marking. In Angoff's original note on his method (Angoff, 1971, pp. 514 -5), he discusses judgements about the likely performance of a single hypothetical marginal student. Then as a footnote, he suggests a slight variation in which judges are asked to think about a group of such students and to estimate the proportion of such students who would answer each item correctly. The latter method is used in the NSW procedures. Each judge works independently. For a given band level, the total cut score for the whole paper is obtained by summing the item cut scores. In Stage 2, the judges are given the results of the Stage 1 judging and are able to freely discuss these. They are also given statistical feedback on how student groups, who gained particular marks on an item, performed on the total paper. They are then able to modify their cut scores if desired. In Stage 3, they are given the results of the Stage 2 judging and given the complete scripts of students who obtained a mark close to the total cut scores. After looking at the quality of this work, the judges are given a final opportunity to modify their cut scores again. When the cut scores are finalised, they are averaged across the judges to get a single cut score for each band. These cut score means are then scaled to the arbitrary reporting scale as described above (a Band 6 cut score scales to 90, a Band 5 cut score scales to 80 and so on). The resulting final marks are called aligned marks, and

are the ones reported to students.

The physical arrangements involve the judges working around a large table which has room to accommodate the various recording sheets, Band descriptors, examination papers, marking scales, scripts and other documents involved in the process. In Stage 1, where the judges give independent judgements, the judges are physically close enough for it to be possible for a judge to observe the cut score decisions of a neighbouring judge but the presence of the Field Officer, who sits at the table and observes the process, would inhibit any collaboration at Stage 1. In any case, the judges know that they are free to collaborate at Stages 2 and 3, so there is little incentive for them to collaborate at Stage 1. At all stages of the process, the judges are prevented from obtaining any information about the associated percentages in the bands as a result of their decisions. They only find out these percentages at the same time as the general public, when the results are published on the Board of Studies website.

The CLT Method

One way of estimating the standard error is to use the Central Limit Theorem. For a given cut score, the variance in the independent judgements is noted and divided by the number of judges. The result gives an estimate of the variance of the *means* that would occur if the judging process was repeated a large number of times with fresh judges being randomly drawn from an infinite population of judges. For a given judging group, if the decisions of the judges are independent, then an estimate of the standard error is given by

$$\hat{\sigma}_E = \frac{S}{\sqrt{n}}, \qquad (1)$$

where S is the unbiased standard deviation of the cut scores across the group and n is the number of judges (usually six to nine).

As n is small, the distribution of errors will not be normally distributed, but will be distributed as Student's t. This distribution is symmetric, like the normal distribution, but is leptokurtic (thicker at the extremes). A conventional 95% confidence interval could be constructed using the Student's t distribution. This procedure is quite appropriate for Stage 1, where the judgements are made independently. However in Stages 2 and 3 the judges work together and usually some modification of their judgements is made. Thus the above estimate of the standard error (based on Stage 1) would not be able to take into account the effects of the collaboration in Stages 2 and 3.

In other examination systems, the median of the judges' decisions may be used instead of the mean. The standard error of the median is more subject to sampling fluctuations and is therefore less efficient. It is about 25% larger than the standard error of the mean and is estimated by

$$\hat{\sigma}_{Median} = 1.253\frac{S}{\sqrt{n}}. \qquad (2)$$

This formula may be used to give an approximate conversion. Thus the comparative results obtained in this paper can be applied to other systems which use the median.

The EQP Method

The method described here attempts to eliminate all sources of variation except for that attributed to the judging. To remove variation associated with changes in candidature selection and with changes in teaching/learning practices, extremely large and stable candidature courses were selected - the external tests in English and Mathematics at the School Certificate. About 80,000 students take each of these courses. Given that the Year 10 School Certificate population have a consistent retention of approximately 97% from Year 7 (the first year of high school), it is extremely unlikely that changes in these aligned distributions in recent years could be attributable to changes in selection. Secondly, while the School Certificate method of *reporting* had changed in 1998 from a norm-referenced system to reporting in performance bands, the curricula did not and have been unchanged for many years. In 2003, the curricula for these courses was revised and major changes are planned to be implemented for 2005. Thus, the Year 10 English and Mathematics courses give a unique opportunity to observe variation in cut scores in a system containing large stable candidatures and a stable curriculum. It is unlikely that cut score changes over the years could be attributed to curriculum factors as the Year 10 curriculum has remained constant. Also, given the candidature sizes, it is also unlikely that these differences are attributable to an improvement or decline in teaching and learning over the whole state. In the absence of a systematic statewide program designed to improve performance at these tests in the period under consideration, there is no reason to postulate a change in the statewide aligned distributions – in such a large population, an improvement in some schools would probably be balanced by a decline in others.

Therefore we are arguing that most of the variation in the aligned distributions would come from two sources. The first is variation in the intrinsic difficulty of the examination papers and their associated marking patterns, as reflected in the raw mark distributions. The second is variation in the judging – the variation we are interested in. To remove the first type of variation, a base year was chosen (1998) and all raw mark distributions in a course were converted to have the same mean, standard deviation and distribution shape as the 1998 raw mark distribution. This was effected by

applying an equipercentile transformation to each of the 1999-2002 raw distributions to change the shape to that of the 1998 distribution. For a definition of equipercentile equating, see Angoff (1971, p. 563). In practice, anchor points are obtained for score pairs corresponding to the same percentile rank and scores lying between these anchor points are converted by interpolation (see, for example, the curved line of relationship equating two tests on p. 573 of Angoff, 1971). The equipercentile transformations were then applied to the raw mark total paper cut score means to give a set of cut score means that is comparable over time. These cut score means are to be interpreted as being on the same scale as the 1998 raw marks for each course. The standard deviation of these transformed cut score means across the years from 1998 to 2002 is then the estimate of the standard error of the judges' means.

Results for the EQP method

Table 2 below shows the comparable cut scores in English for each band level for the EQP method and gives their unbiased standard deviation (SD), which is the standard error estimate. When expressed as a fraction of the standard deviation of the total test, this is the Standardised SD, which is useful for making rough comparisons between the English and Mathematics results. Table 3 shows the comparable statistics for Mathematics.

The standard errors have certain similarities in English and Mathematics. Firstly, they are small for the very top band, indicating that the judges have similar views as to what constitutes a Band 6 marginal candidate. Secondly, the standard errors are relatively large at 3.56 and 4.55 respectively for a marginal student at Band 3. This implies that the judges have a much less clear view of the type of student at the Band 3 cutoff than they have of the very best students.

The patterns, however, differ in their clarity in defining a bare "pass" candidate at the bottom of Band 2. For English, this gave the widest standard error of all the bands, whereas for Mathematics it gave the second smallest standard error. The Mathematics judging groups apparently have similar views as to what constitutes very good performances and what constitutes barely adequate performances. It is in the middle bands that they disagree the most. The English groups, on the other hand, were fairly consistent for Bands 1 – 4 but disagreed more for the marginal students in Bands 2 and 3.

Although the raw standard errors were lower for English than for Mathematics, the test standard deviation was also lower for English. When the standard errors are expressed as a proportion of this standard deviation, it can be seen that English has lower standard errors for bands 4, 5 and 6 and higher standard errors for bands 2 and 3.

Table 2: English cut score variation and EQP standard errors

Band	Year	Cut Score	Mean	SD (standard error estimate)	Standardised SD
6	1998	89.0			
	1999	86.0			
	2000	88.0			
	2001	87.7			
	2002	87.5	87.64	1.08	0.07
5	1998	78.0			
	1999	75.3			
	2000	77.0			

	2001	76.5			
	2002	76.5	76.66	0.98	0.07
4	1998	66.0			
	1999	64.1			
	2000	64.0			
	2001	66.3			
	2002	62.4	64.55	1.60	0.11
3	1998	51.0			
	1999	52.1			
	2000	48.5			
	2001	53.1			
	2002	44.2	49.78	3.56	0.24
2	1998	36.0			
	1999	37.0			
	2000	30.3			
	2001	36.2			
	2002	28.5	33.61	3.90	0.26

Table 3: Mathematics cut score variation and EQP standard errors

Band	Year	Cut score Variation	Mean	SD (standard error estimate)	Standardised SD
6	1998	85.0			
	1999	80.2			
	2000	80.7			

	2001	83.0			
	2002	83.3	82.44	1.98	0.09
5	1998	70.0			
	1999	61.2			
	2000	65.8			
	2001	63.0			
	2002	64.5	64.90	3.33	0.15
4	1998	50.0			
	1999	42.7			
	2000	47.0			
	2001	40.1			
	2002	42.0	44.36	4.04	0.18
3	1998	30.0			
	1999	24.0			
	2000	27.0			
	2001	18.8			
	2002	20.7	24.10	4.55	0.21
2	1998	14.0			
	1999	11.4			
	2000	12.9			
	2001	11.7			
	2002	8.6	11.72	2.03	0.09

Comparison of the Two Standard Error Methods

The second method uses the Central Limit Theorem to estimate the variation between cut score means as a function of the variation between the individual cut scores of the judges. As this variation is taken from the Stage 1 judging, it is possible that it will give larger standard error estimates than the EQP method. For each band level, the standard error estimates were averaged over the five years and are compared with the EQP estimates in Table 4.

Table 4: Comparison of the Standard Error Methods

Band	English Standard Errors		Mathematics Standard Errors	
	EQP observed	CLT	EQP observed	CLT
6	1.08	1.55	1.98	1.99
5	0.98	2.23	3.33	2.21
4	1.60	2.06	4.04	2.52
3	3.56	2.64	4.55	2.71
2	3.90	2.56	2.03	2.27

For English, Table 4 shows that the CLT method adequately approximates the EQP method for Bands 4 – 6, for which it gives a conservative estimate. But for the lowest two bands, EQP gives more variation in the standard error than would be predicted by the individual judges' variation through the CLT. In Mathematics, the CLT method gives good estimates of the EQP method for the highest and lowest bands, but under-estimates the latter for the middle three bands.

Discussion

It is not easy to account for the differential agreement between the methods across different bands. In the absence of data, one may have assumed that the effect of employing Stages 2 and 3 in the standards-setting process would have improved the accuracy over Stage 1, so that the CLT method (based solely on Stage 1) would be expected to *over-estimate* the standard error. However, for the lower two bands in English and for the middle three bands in Mathematics, the opposite is the case. Clearly, the selection of judges each year is not strictly a random sampling process, so that the group selected in 2002 may be more (or less) similar to the 2001 group than would be expected on the basis of random sampling. In addition, the judging over different years may not be strictly independent, even if the groups were entirely different random samples. This may occur because the percentages awarded in the bands are eventually made public knowledge and individual judges may be aware of the previous years' percentages and have some opinion about their appropriateness. Whether consciously or unconsciously, this may influence the degree of leniency/severity that they apply to the cut scores in the following year.

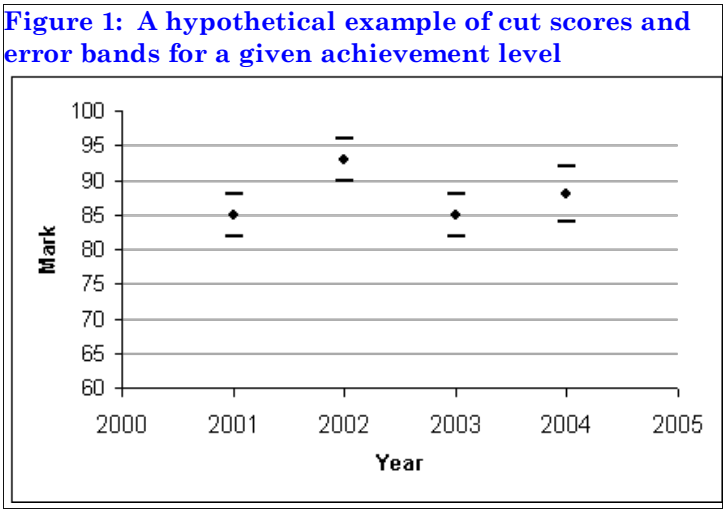
As expected, by Stage 3 of the modified Angoff procedure the effect of the judges' collaboration was to make their judgements more alike and hence the standard deviation of their cut scores was substantially reduced in both English and Mathematics. This information, however cannot be validly used to construct standard error estimates as the lack of independence of the decisions makes any estimate based on them virtually meaningless. For example in our Year 12 award, the Higher School Certificate, we have observed in some courses that the collaborating judges at Stage 3 have all agreed on the same values for their cut scores, giving zero variance, and hence the (invalid) use of the CLT formula would give a standard error estimate of zero. The only way to validly take account of the effects of Stages 2 and 3 would be to repeat the entire process from Stage 1 to Stage 3 with additional independent teams of judges. Each team would collaborate within themselves at Stages 2 and 3 but this would be entirely independent of the other teams. Then the observed variation between the team means would give an estimate of the standard error. This could be experimentally established as a "one off" trial but the time and expense involved would preclude it from being a regular part of the standard-setting procedures. It is the impracticality in conducting this type of experiment on a regular basis that provides the motivation to investigate the effectiveness of the CLT estimate based on Stage 1.

This analysis has removed the effects of differential distribution shapes caused by varying examination paper difficulties by converting all distributions to the 1998 raw mark scale. A further question of interest is whether differential measurement error in the test scores could effect the standard error estimates. In theory this could occur if the reliabilities of the scores varied markedly from year to year. The reliabilities affect the intercorrelations between

items and this affects the amount of regression that would occur relative to a total cut score based on the summing of cut scores across individual items. For example, imagine a test giving unreliable scores where the intercorrelations amongst the items are very low. For simplicity, consider a test where the items are extended response and parallel, and where the judges, being perfectly consistent, identify cut scores on each item that mark off the top 5% of students as being in the top ability band. Then the summing of these cut scores may give a total cut score that marks off only 1% of students achieving the top band, due to the low intercorrelations between items. In contrast, consider a similar situation where all factors are identical except that the item intercorrelations are high. Again the judges choose item cut scores that mark off the top 5% of students, but this time the total cut score identifies 4% of students in the top band, due to the higher item intercorrelations. Any equipercentile mapping that converts the distribution of the unreliable test to the same distribution as the reliable test (or vice versa) will not overcome this disparity in percentages. In fact, if the total cut scores on the two tests are exactly the same mark, then the equipercentile mapping, in equating all moments of the two distributions, will then transform these marks to be different. This effect obviously would be greatest for cut scores near the extremities of the distributions. Thus the total cut scores (and hence the standard error estimates) could vary solely as a a function of differing reliabilities from year to year. The empirical investigation of this effect is beyond the scope of this paper. However, in practice the high level of reliabilities for SC English and Mathematics and their uniformity across time (as shown in Table 1) imply that the differential effect from year to year would be minimal for these data.

Given that the CLT method does not quite account for the variation observed in some of the bands, one should be very cautious in making inferences about possible changes in achievement. One course of action would be to use a stringent alpha level in applying constructing confidence intervals around cut scores, using a probability of 0.01 say, instead of 0.05. Secondly, it would be advisable to observe the variation over a few years before drawing any inferences about change, to see if any trends are emerging. For example, consider the hypothetical set of standard error bands shown in Figure 1 below.

Figure 1: A hypothetical example of cut scores and error bands for a given achievement level



In Figure 1, the 2002 error band does not overlap with the 2001 band, but it would be unwise to assume that a significant change in the standard of achievement had been attained. In 2003 the aligned mark drops again and then rises somewhat in 2004. In this case, the error bands (which are certainly fallible estimates themselves) have not been sufficient to account for the observed variation. This type of result should indicate the need to wait for a few years of consistency in the pattern before assuming that there had been a real change in demonstrated achievement.

The above method of reporting may help in enabling changes in the standard of achievement to be identified. However, the interpretation of why these changes have occurred may be more difficult. In courses which are not compulsory, a change in demonstrated achievement may simply be caused by *selection factors* rather than improvements in teaching and learning. For example, if sufficient able students drop out of a course (to switch to other courses) and are replaced with lower ability students, then the standard of achievement in the course may drop, but this may have nothing to do with changes in teaching and learning. This effect would be more common in small candidatures.

References

Angoff, W.H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement*. (2nd ed.). Washington, D.C.: American Council on Education.

Berk, R.A. (1984). *A guide to criterion-referenced test construction*. Baltimore, MD: John Hopkins University Press.

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

Buckendahl, C.W., Smith, R.W., Impara, J.C. and Plake, B.S. (2001). *A comparison of Angoff and Bookmark standard setting methods*. Paper presented at the annual meeting of the National Council on Measurement in Education in

Seattle, WA.

Cross, L.H., Impara, J.C., Frary, R.B. and Jaeger, R.M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. *Journal of Educational Measurement*, 21, 113-130.

Giraud, G., Impara, J.C. and Buckendahl, C. (2000). *Educational Assessment*, 6, 291-304.

Hambleton, R.K. and Eignor, D.R. (1980). Competency test development, validation and standard setting. In R.M. Jaeger and C.K. Tittle (Eds.), *Minimum competency achievement testing: Motives, models, measures and consequences* (pp. 367-396). Berkeley, CA: McCutchan.

Hambleton, R.K. and Plake, B.S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8, 41-56.

Harasym, P.H. (1981). A comparison of the Nedelsky and modified Angoff standard setting procedure on evaluation outcomes. *Educational and Psychological Measurement*, 41, 725-734.

Hoel, P. G. (1961). *Introduction to Mathematical Statistics* (2nd ed.). New York: John Wiley.

Impara, J.C. and Plake, B.S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, 34, 355-368.

Jaeger, R.M. (1993). Certification of student competence. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed.), pp. 485-514. New York: American Council on Education and Macmillan Publishing.

Linn, R.L. (2003). Performance standards: utility for different uses of assessments. *Education Policy Analysis Archives*, 11 (31).

Livingston, S.A. and Zieky, M.J. (1983). *A comparative study of standard-setting methods*. Research Report No. 83-38. Princeton, N.J.: Educational Testing Service.

Plake, B.S. (1998). Setting performance standards for professional licensure and certification. *Applied Measurement in Education*, 11, 65-80.

Shepard, L.A. (1980). Technical issues in minimum competency testing. In D.C. Berliner (Ed.), *Review of Research in Education*. (Vol. 8, pp. 30-82). Washington, D.C.: American Educational Research Association.

About the Authors

Dr Robert MacCann is Head, Measurement and Research Services, within the NSW Board of Studies, Sydney Australia.

Professor Gordon Stanley is President of the NSW Board of Studies and Adjunct Professor, School of Policy and Practice, in the Faculty of Education and Social Work, University of Sydney.

Contact:

Dr Robert MacCann, Head, Measurement & Research Services,
Board of Studies NSW
GPO Box 5300,
Sydney 2001 Australia

maccann@boardofstudies.nsw.edu.au

Descriptors: Standards Setting; Modified Angoff Procedures; Standard Error; Central Limit Theorem; Equipercentile Equating

Citation: MacCann, Robert G. & Gordon Stanley (2004). Estimating the standard error of the judging in a modified-angoff standards setting procedure. *Practical Assessment, Research & Evaluation*, 9(5). Available online: http://PAREonline.net/getvn.asp?v=9&n=5.