Comparing small-sample equating with Angoff judgement for linking cut-scores on two tests

Tom Bramley Research Division

Introduction[1]

The educational measurement literature makes a clear distinction between the activities of standard setting on the one hand, and test equating and linking on the other. For example, these topics occupy different chapters in the standard reference work Educational Measurement (Brennan, 2006). Test equating is usually defined in a fairly narrow, technical way such as: "Equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably" (Kolen & Brennan, 2004, p.2). Standard setting, on the other hand, is usually defined more broadly such as "…the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states of performance" (Cizek, 1993, p.100). The main issues in test equating tend to be around the definition of the 'correct' equating transformation, and the data collection designs and statistical methods necessary to estimate it. In standard setting, however, the procedures are "… seldom, if ever, impartial psychometric activities, conducted in isolation. Social, political and economic forces impinge on the standard-setting process…" (Cizek & Earnest, 2015, p.213). In particular, standard setting processes involve human values and judgements, and differences in these are to be expected.

Conceptually, however, the processes of standard setting and test equating are clearly very closely related. The performance standard can be conceived of as a point on an abstract continuum, and the aim of the standard setting process as being to find the score on the raw scale of the particular test at hand that corresponds to this point. This seems very similar to the conceptualisation of equating in Item Response Theory (IRT) – the raw scores on two tests that correspond to the same level on the unobservable underlying trait are deemed equivalent.

If we are prepared to conceive of the abstract continuum on which the performance standard is located and the latent trait of the IRT model as one and the same, then we can see that carrying out separate standard setting exercises on Tests X and Y is in theory no different from attempting to equate them (at the point on the latent trait corresponding to the cut-score) by an IRT approach. Of course, the *results* of applying such dramatically different approaches to the same problem could be expected to differ.

Although it would seem most logically justifiable to carry out a standard setting exercise just once (to establish one definitive example of a realisation of the abstract performance standard on a concrete test) and then to use statistical equating to link all subsequent (or other) forms to that, in practice it may well be that a standard setting method is used (perhaps alongside other methods) to inform or set the cut-score on subsequent forms. Thus, the standard setting method is used in practice as a test equating (standard maintaining) method. There are several scenarios where this might arise, for example:

i) if the test is very high-stakes (e.g., a licence-to-practise test) where procedures require 'stakeholder' involvement in setting the cut-score on each test form;

ii) if sample sizes are so low on each test form that statistical equating methods are not trusted;

iii) if contextual factors (such as cost, need for test security, local culture and expectations) prevent some of the necessities for equating methods such as pre-testing, administration of an anchor test, or embedding of field-test items into live tests;

iv) if there is a need to determine a cut-score before any 'live' performance data has been collected.

The conceptual similarity between equating and standard setting raises questions of the relative accuracy[2] of the two methods. Our starting assumption was that in an ideal world a large-sample equating exercise would be the preferred way to map a cut-score from one test to another parallel one. However, since the standard error of equating in a test equating exercise depends upon the sample size, continually reducing the sample size presumably will reach a point at which the equating error becomes greater than the error that would arise from carrying out two separate standard setting exercises. The equating error from the latter will depend on the details of the method used, but for all methods that rely on the judgement of item difficulty by experts, a fundamental issue is the extent to which those judgements correspond to the actual empirical difficulty. One of the motivations for this research was the realisation (see Benton, 2020, this issue) that estimates of item difficulty based on extremely small samples of empirical data (N<10) can correlate better with the actual (full population) values than estimates based on expert judgement. The aim of this study was to compare, by simulation, the accuracy of mapping a cut-score from one test to another by expert judgement versus the accuracy with a small-sample equating method.

Method

Standard setting method

The standard setting method we simulated was the 'mean estimation' method – a variant of the more well-known Angoff Method (e.g., Loomis

[1] This is a shortened and simplified version of a paper presented at the AEA-Europe conference in 2017 (Bramley & Benton 2017).

[2] In this article we use 'accuracy' in the general sense of overall accuracy including both bias and random error.

& Bourque, 2001). It is applicable to tests containing polytomous as well as dichotomous items. If the test consists solely of dichotomous items it is the same as the Angoff Method. Experts estimate the difficulty of each of the items in a test in terms of the mean score likely to be obtained on each item by a group of minimally competent examinees (MCEs). If the test is pass-fail then the MCEs are those who are just competent enough to pass. If the test is graded into more than two categories, there are different groups of MCEs for each cut-score. The cut-score is derived by summing the estimated means and then averaging across judges, rounding the result to an integer if necessary (or averaging and then summing – it makes no difference).

Previous research (e.g., Impara & Plake, 1998) has suggested that although estimating the mean scores of MCEs can be difficult for experts in an absolute sense, they are more adept at discerning the correct rank order of the difficulty of items. Hence, judgements from experts can potentially be transformed onto the correct scale before being used to inform standard setting (Thorndike, 1982; Humphry, Heldsinger, & Andrich, 2014). Since judgements can be transformed to the correct scale, the correlation between estimated difficulties and actual difficulties (often measured by item facilities – mean mark divided by maximum possible mark) provides a reasonable idea of the value of the information from such methods, as discussed above. In our simulation (described in more detail later) we wanted to vary this level of correlation and assess the effect on the outcome.

Equating method

There are a variety of equating methods appropriate for use with small samples (for example, see Livingston & Kim, 2009; or Kim, von Davier, & Haberman, 2008). We wanted a method suitable for the 'non-equivalent groups anchor test' (NEAT) design. This is because for equating test forms which are only produced once or twice a year (such as GCSEs or A Levels) it is not usually possible to get one group of examinees to take both forms, or to obtain randomly equivalent groups of examinees. It is much more frequently possible to obtain two different groups and adjust statistically for differences in ability between them by means of an anchor test. We chose chained linear equating (e.g., Puhan, 2010) because it requires fewer parameters to be estimated than the theoretically preferable (with large samples) equipercentile equating. Puhan (2010) reports that, across a range of conditions, chained linear equating tends to perform well compared to other linear equating techniques for the NEAT design.

We were also interested in exploring the effect of clustering on the small-sample equating outcome. In practice, it might only be logistically feasible to obtain examinees from a single class in a small number of schools for an equating exercise, so it was of interest to see how a small clustered sample differed from a genuinely random sample of the same size.

In brief, the equating scenario consisted of a Test X (where we assumed the cut-scores were known) and a Test Y where we needed to set equivalent cut-scores. We simulated mean estimation judgements at two levels of correlation (0.6 and 0.9) between estimated and empirical values, and derived the cut-score on Test Y by adding up the simulated means for the items on Test Y. We compared this with a chained linear equating method in two conditions:

1. Random samples of 30 examinees from three schools in Group A took Test X, and from three different schools in Group B took Test Y, and all 180 examinees took an anchor Test V;
2. Simple random samples of 90 examinees in Group A took Test X and in Group B took Test Y, and all 180 examinees took an anchor Test V.

In both cases we considered two cut-scores, one at the lower end of the raw score scale and one at the higher end.

Data

The dataset forming the basis of all the analyses reported here was artificially constructed from a large real dataset containing the responses of 15,731 examinees to a test with a maximum possible raw score of 200. The questions were made up of sub-questions (henceforth items), and the items ranged in tariff (maximum score) from 1 (i.e., dichotomous) to 5 (i.e., polytomous with six score categories). The facility values of all items were calculated and two Tests X and Y, each with a maximum possible raw score of 60, were constructed by selecting two sets of items comprising fifteen 2-tariff items and ten 3-tariff items by systematically alternating selection from the items ordered by facility value. An anchor Test V was constructed from 20 dichotomous items (which was all the dichotomous items and hence no selection method was required).

The examinees came from 323 schools, each contributing between 1 and 238 examinees (mean 48.7, median 33). Each school had a 5-digit identification number, which was known to be non-randomly assigned. Two non-equivalent groups of examinees of roughly the same size were created by assigning those in schools with ID numbers below a certain value to Group A and the rest to Group B. Scores on the anchor test correlated around 0.8 with scores on Test X and Y in both groups.

Table 1 shows that Test Y was slightly easier than Test X (higher mean score) but the lower SD of scores on Test Y shows that the difference in difficulty was not uniform across the score range. It is also clear that Group A was of higher ability than Group B (its mean score was higher on all tests).

Table 1: Descriptive statistics for scores on Tests X, Y and V

Test	*All (N=15,731)*		*Group A (N=7,752)*		*Group B (N=7,979)*	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
X (max 60)	31.76	13.29	33.00	13.39	30.55	13.08
Y (max 60)	32.36	12.16	33.46	12.37	31.29	11.86
V (max 20)	10.01	3.44	10.30	3.50	9.72	3.34

As described in the introduction, because of the conceptual similarity between the 'abstract continuum' on which the performance standard is located and the 'latent trait' of IRT, we defined the correct equating function to be the one arising from IRT true score equating on the complete dataset (i.e., X, Y and V items calibrated concurrently for both groups in a single-group design with no missing data). We focused on two different cut-scores on Test X: 15 out of 60, and 45 out of 60. The 'definitive' equated cut-scores on Test Y arising from the IRT true score equating were 17.20 and 44.21.

Equating via simulating judgements in a standard setting method

We simulated expert judgement of item difficulty by adding random error to the 'correct' (empirical) values. We simulated two levels of correlation: 0.6 (a value representative of published Angoff studies, see for example Brandon, 2004), and 0.9 (a much higher value than usually found, in order to represent a very optimistic view of what might be achievable in ideal conditions).

The technical details of the simulation are described in Bramley and Benton (2017). The process was repeated 1,000 times for each of two different values of the correlation *r* (0.9 and 0.6) and for the two different Test X cut-scores (15 and 45).

The simulated judgements were used to produce equated cut-scores, using the standard setting method previously described. The distributions of equated cut-scores were then compared with the definitive (correct) cut-score. Specifically, bias B was defined as the mean difference (across replicates) between the equated score for each replicate and the correct cut-score; error variance E was defined as the variance of the equated cut-scores; and the root mean squared error RMSE (Root Mean Square Error) was calculated as sqrt(B^2+E).

Equating via a traditional small-sample equating method

For condition 1, all schools with 30 or more examinees were selected and then a two-stage sampling process first selected at random three schools from each group, and then a random sample of 30 examinees from each school. This process was replicated 1,000 times. For condition 2, we selected 1,000 simple random samples (with replacement) of 90 examinees from Group A and 90 from Group B. An equated cut-score on Test Y for each of the Test X cut-scores (15 and 45) was derived by chained linear equating in each replicate in each condition (see Bramley & Benton, 2017 for the equations). The distribution of equated scores across the 1,000 replicates was then compared with the definitive cut-score in the same way as for the simulated judgements.

Results

Table 2 shows that in all cases except small-sample equating with the clustered sample (condition 1) the bias made a negligible contribution to the overall RMSE. The more realistic value for the correlation (0.6) had RMSE values nearly twice as high as that for the optimistic value (0.9) at both cut-scores. The % distributions in Table 2 refer to equated cut-scores on Test Y rounded to the nearest integer. This is on the assumption that in practice, if an integer cut-score were required to be set on Test Y, the correct values would be 17 and 44. This causes a slight asymmetry because an equated score of 44.6 (say) would be rounded to 45 and be 1 too high, whereas a less accurate equated score of 43.6 would be rounded to the correct value of 44. For simulated correlations of 0.9, the equated cut-score was within ±1 of the correct score around 75% of the time (cut-score of 15) or 80% of the time (cut-score of 45), but for simulated correlations of 0.6 only around 50% were in this range, and around 25% were three or more score points away.

The overall accuracy of small-sample equating, as measured by the RMSE, was better in condition 2 (simple random sample of 90 examinees from each test) than in condition 1. At both cut-scores, the condition 2 RMSE was roughly half-way between the RMSE values from simulated judgements with r=0.6 and r=0.9. The condition 1 RMSEs were about 0.7 score points higher than the corresponding condition 2 RMSEs, for both cut scores, showing the detrimental effect of clustering of examinees within schools on equating error. The condition 1 RMSEs were slightly higher than those from simulated judgements with a correlation of 0.6. In the best case for small-sample equating (condition 2) the cut-scores were within one score point of the correct value around 60% of the time for a cut-score of 15 and around 70% of the time for a

Table 2: Equated scores based on simulated judgements and small-sample equating (replications=1,000)

	Simulated judgement		*Equating condition…*		*Simulated judgement*		*Equating condition…*	
	r=0.6	*r=0.9*	*1*	*2*	*r=0.6*	*r=0.9*	*1*	*2*
Test X cut-score	15				45			
Correct Y cut-score	17.20				44.21			
Test Y mean equated cut-score	17.08	17.15	16.39	16.56	44.38	44.33	44.25	44.36
Test Y SD equated cut-score	2.30	1.25	2.41	1.70	2.06	1.12	2.18	1.45
Bias	–0.12	–0.05	–0.81	–0.64	0.18	0.12	0.04	0.15
RMSE	2.31	1.25	2.54	1.82	2.07	1.13	2.18	1.45
%<= –3	12.1	0.9	19.1	12.0	8.8	0.9	10.5	1.9
% –2	13.3	8.1	10.4	13.6	9.3	4.3	9.9	7.9
% –1	16.4	21.9	18.8	22.1	14.3	17.6	16.9	17.5
% 0	17.6	29.6	19.7	23.2	18.4	32.0	19.0	27.1
% +1	15.0	25.6	14.1	16.0	19.4	31.0	16.2	24.6
% +2	10.3	11.3	10.3	9.7	13.8	12.0	11.4	13.4
% >= +3	15.3	2.6	7.6	3.4	16.0	2.2	16.1	7.6

cut-score of 45. Bias made a small contribution to the RMSE at a cut-score of 15 and a negligible contribution at a cut-score of 45. The fact that sampling error was the main contributor to RMSE in all methods and conditions suggests that comparisons are not critically dependent on how the 'true' equating function is defined, because this would only affect the bias and not the sampling error.

Discussion

This study has compared, by simulation, the level of accuracy that might be obtained from a standard setting method (mean estimation) if applied as a test equating method to that which might be expected from a small-sample test equating method (chained linear equating). As expected, the standard setting method resulted in more accurate equating when we assumed a higher level of correlation between simulated expert judgements of item difficulty and empirical difficulty. For small-sample equating with 90 examinees per test, more accurate equating arose from using simple random sampling compared to cluster sampling at a given sample size. The actual values of RMSE depended on the cut-score, being generally larger for the cut-score where the correct equated cut-score on Test Y was further from the cut-score on Test X. The simulations based on the more realistic value for the correlation between judged and empirical difficulty (0.6) produced a similar RMSE to small-sample equating with cluster sampling. Simulations of standard setting based on the optimistic correlation of 0.9 had the lowest RMSEs of all.

As shown by Benton (2020, this issue), even very small samples of examinees can give a more accurate picture of the relative difficulty of items than estimates from experts. We may therefore be surprised that the small-sample approach trialled here did not perform even better. There are a number of reasons for this. One reason is that the equating approach adopted in the simulation study required calibration of examinee abilities across two groups using an anchor test. Small-sample equating with a single group design would be significantly more accurate. Even within the NEAT design, it may be that other approaches, such as Tucker linear equating or Rasch true score equating, may provide a more stable estimate of equivalent scores than chained linear equating.

Most important, however, is the fact that our simulations assumed that judged and empirical values for the mean scores of MCEs would differ only in their rank order, and that the mean and SD would (apart from sampling error) be the same. In fact, evidence both old (Lorge & Kruglov, 1953) and new (Humphry et al., 2014) suggests that expert judges tend to think that easy items are harder than they are, and that hard items are easier than they are. That is, the implied scale unit of estimated difficulty tends to be larger (i.e., less discriminating) than the scale unit of empirical difficulty: the judges' estimates are less spread out than the empirical values. Humphry et al. (ibid.) suggested applying a linear transformation to align the scale units, on the assumption that judges are unbiased when estimating passing proportions/probabilities of 50%. Although this assumption seems reasonably plausible, it nevertheless needs empirical support. In any event, we were not confident that we could choose realistic values for scale shrinkage effects to include in our simulation because they may depend on a number of contextual factors. This is an area for further research.

In our simulations, sampling error was the dominant contributor to RMSE, which suggests that attempting to reduce sampling error at the risk of increasing bias may also be worth considering. One way of achieving this would be to apply the 'synthetic linking' approach of Kim et al. (2008) where the final equated cut-score on Test Y is a weighted average of the Test X cut-score and the cut-score derived from the equating. This approach is clearly most suitable when there is some reason to believe that the two tests should have similar cut-scores – perhaps if they have been constructed to the same detailed specification.

The main issue is whether the aggregate of judges' estimates of item difficulty provides useful information about relative test difficulty. The article by Benton (2020, this issue) gives some cause for pessimism here, at least as far as the kind of data we see at GCSE and A Level is concerned. The degree of correlation between judged and empirical item difficulty is clearly an important factor in the usefulness of Angoff-related standard setting methods. Using a small-sample equating method may be preferable to using a standard setting method if typical levels of correlation are to be expected, and indeed this was the conclusion of Dwyer (2016), although it should be noted that the (actual, not simulated) correlations of the judge estimates in his study were in the range 0.39 to 0.49 – lower than observed in many other studies. If it were possible to increase the correlation beyond 0.6 by increasing the number of judges in a judging panel and/or training them to make the mean estimation judgements, then substantial improvements in the accuracy of the standard setting method could be obtained – in the simulation here a correlation of 0.9 was more accurate than the best small-sample equating scenario (a simple random sample of 90 examinees). However, Benton (2020, this issue) argues that rather than focusing on the absolute size of the correlation coefficient, the critical issue is the proportional reduction in error in predicting empirical difficulty from judged difficulty. This takes account of any overall biases and scale differences in judgements as well as disagreements in rank order.

In conclusion, it can be observed that in some contexts standard setting methods are used to achieve the same goal as test equating methods, namely determining cut-scores on test forms that relate to the same performance standard. IRT true-score equating provides a conceptual link between the two, if it is reasonable to conceive of the IRT latent trait as being the same as the abstract continuum containing the performance standard. The simulations reported here have suggested that the overall accuracy of Angoff-based standard setting methods could in some circumstances be similar to what might be expected from test equating with a NEAT design using small samples (N~100) of examinees. Of course, these findings all derive from simulations based on just one dataset, so we are not in a position to make general recommendations about what to do in particular applied contexts. We made choices about how to define the 'true' equating function and which particular standard setting method and small-sample equating method to use, all of which could be varied. The effect of using polytomous items rather than dichotomous anchor items could be explored, as could the effect of varying test length. Furthermore, our method of artificially constructing Tests X and Y ensured that they would be reasonably similar in difficulty. However, these findings point to a way in which practitioners could set up experiments or simulations that more closely match their own particular contexts, in order to discover whether using a standard setting method based on expert judgement might be more accurate than using a small-sample test equating method (or vice versa); or whether focusing effort on constructing parallel (equally difficult) tests would be a better use of available resource.

References

Benton, T. (2020). How useful is comparative judgement of item difficulty for standard maintaining? *Research Matters: A Cambridge Assessment Publication, 29*, 27–35.

Bramley, T. & Benton, T. (2017). *Comparing small-sample equating with Angoff judgment for linking cut-scores on two tests*. Paper presented at the annual conference of the Association for Educational Assessment-Europe (AEA-Europe), Prague, Czech Republic, 9–11 November, 2017.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*(1), 59–88.

Brennan, R. L. (Ed.). (2006). *Educational Measurement* (4th ed.). Washington, DC: American Council on Education/Praeger.

Cizek, G. J. (1993). Reconsidering Standards and Criteria. *Journal of Educational Measurement, 30*(2), 93–106.

Cizek, G. J., & Earnest, D. S. (2015). Setting performance standards on tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp.212–237). New York: Routledge.

Dwyer, A. C. (2016). Maintaining equivalent cut scores for small sample test forms. *Journal of Educational Measurement, 53*(1), 3–22.

Humphry, S., Heldsinger, S., & Andrich, D. (2014). Requiring a consistent unit of scale between the responses of students and judges in standard setting. *Applied Measurement in Education, 27*(1), 1–18.

Kim, S., von Davier, A. A., & Haberman, S. (2008). Small-sample equating using a synthetic linking function. *Journal of Educational Measurement, 45*(4), 325–342.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices*. (2nd ed.). New York: Springer.

Impara, J.C., & Plake, B.S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions of the Angoff standard setting method. *Journal of Educational Measurement, 35*, 69–81.

Livingston, S.A., & Kim, S, (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46*, 330–343.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods and Perspectives* (pp.175–217). Mahwah, NJ: Lawrence Erlbaum Associates.

Lorge, I., & Kruglov, L. (1953). The improvement of estimates of test difficulty. *Educational and Psychological Measurement, 13*, 34–46.

Puhan, G. (2010). A comparison of chained linear and poststratification linear equating under different testing conditions. *Journal of Educational Measurement, 47*(1), 54–75.

Thorndike, R. L. (1982). Item and score conversion by pooled judgment. In P. W. Holland & D. B. Rubin (Eds.), *Test Equating* (pp. 309–317). New York: Academic Press.

How useful is comparative judgement of item difficulty for standard maintaining?

Tom Benton Research Division

Introduction

Developing a way to accurately estimate the relative difficulty of two tests before any students have taken them has long been a holy grail in test development. At one time or another, various organisations have explored how well we can discern the relative difficulties of assessments without actually trialling them with students. Recent research on this topic has been produced by Cito in the Netherlands (van Onna, Lampe, & Crompvoets, 2019), ETS in the United States (Attali, Saldivia, Jackson, Schuppan, & Wanamaker, 2014) and Cambridge Assessment in the UK (Curcin, Black, & Bramley, 2009). Item trialling is often undesirable as it places some of the burden of test development upon schools and students, and can lead to concerns over the security of items.

If accurate predictions of item difficulty were possible then, in the context of UK examinations, this would mean being able to accurately set grade boundaries for this year's GCSE exams before any students have attempted the paper. It would also provide an alternative to the current approach of "comparable outcomes" to awarding and its inherent implication that (broadly speaking) the percentage of pupils achieving high grades will not change from the previous year (Benton, 2016). Outside of the UK context, being able to accurately predict the difficulty of items might allow "lowering the sample sizes required for item pretesting, leading to lower costs and increased security of items" (Attali et al., 2014, p.7).

The previous article (Bramley, 2020) has considered the extent to which a particular form of expert judgement (the 'mean estimation' variant of the Angoff Method) might provide sufficiently accurate information on the relative difficulty of two tests. The present article explores the value of expert judgements of item difficulties derived in a different manner – by comparative judgement (CJ).

In this context, a CJ study requires expert judges to sort sets of items according to their perceived difficulty (PD). The rationale for using CJ is that previous research has indicated that judges tend to "be good at predicting the relative difficulties of items but not absolute levels" (Mislevy, Sheehan, & Wingersky, 1993, p.59). Placing items in a rank order of difficulty is conceivably a more intuitive task then estimating the proportion of minimally competent candidates who will answer them correctly, as must be done under the Angoff Method. As such,