Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 14, Number 17, October 2009 ISSN 1531-7714

Item Banking with Embedded Standards

Robert G. MacCann and Gordon Stanley
Oxford University Centre for Educational Assessment
University of Oxford, UK

An item banking method that does not use Item Response Theory (IRT) is described. This method provides a comparable grading system across schools that would be suitable for low-stakes testing. It uses the Angoff standard-setting method to obtain item ratings that are stored with each item. An example of such a grading system is given, showing how a grade and a scaled score could be calculated for a particular student.

Item banking can be a useful way for educational systems to monitor educational achievement. With online testing now becoming commonplace, it is much easier to distribute tests, mark them, and report the results without the burden of excessive paper handling. As Rudner (1998) points out, item banking has major advantages in terms of test development. It is a very time-consuming endeavour for schools to be creating new tests each year. Even if this were done, the interpretation of the test scores would only have a local meaning as the mean difficulties of the tests would vary from school to school.

Rudner's paper is presented in the context of Item Response Theory (IRT) models to equate the different forms of the test that can be drawn from the bank. This paper puts forward a method of item banking that does not use IRT models but can still deliver test scores that are approximately comparable across a national education system and can be related to system norms. Such a method may be suitable for low-stakes testing where a school wishes to determine how it is performing in relation to the rest of the cohort. These features can be approximately achieved through the use of the Angoff standard setting method, combined with an online item banking operation.

Why use an item banking system that does not employ IRT? Some organisations may wish to test over a broad curriculum area within a subject, where items from different topics are covered. For example, within a subject area such as Mathematics, a summative test may be desired that encompasses all the course content taught in a semester. This may include quite distinct topics such as algebra, coordinate geometry and functions. With IRT, care should be taken in the writing of the items to ensure that they are measuring a unidimensional trait. The method outlined in this paper, however, makes no assumptions about item unidimensionality. Consequently, there are no concerns about item fit, in an IRT sense. However, this does not mean that item quality can be ignored. As in all tests, the quality of items is paramount, if the maximum information about each examinee is to be obtained. The allocated test can comprise a set of heterogeneous items that measure general achievement across a broad range of topics within a subject. Naturally if desired, the test could be restricted to a particular topic. A second advantage is that this method works in the metric of the test score, rather than an underlying ability trait – it should be easy to explain to teachers and to interpret results. A third advantage is that it does not require specialist statistical knowledge to program. Thus the complexities of joint maximum likelihood (or other) estimation procedures employed in IRT can be bypassed.

The Item Bank

In practice, the items stored in the bank would be objectively scored (0 or 1) multiple choice items. The reason for this is that such items can be automatically marked by the central computer. In theory, the method to be outlined in this paper could be made to work with constructed response items, but a mechanism for marking these would need to be found. In the future, the automatic marking of constructed response items by computer will become more commonplace (e.g. Burstein, 2003; Attali and Burstein, 2006). Educational Testing Service (2006) already has a web-based marking system for constructed response items in its TOEFL system. For the moment, however, assume that the items are multiple choice.

Regardless of the test equating mechanisms within the bank (whether IRT or otherwise), it is good practice to attempt to make the tests delivered as similar as possible in mean difficulty. If the tests are not too different in difficulty, then the equating mechanisms will work more efficiently. It is also important, from a face validity perspective, that the tests appear to be not too different in difficulty. Secondly, it is also desirable that the tests should have a similar spread of content. Suppose for example, that the bank contained items that tested basic arithmetical operations – addition, subtraction, multiplication and division. Then in a bank without constraints, it is possible for one student to draw a test that contains mostly addition items, whereas another student may draw a test with mostly division items. Not only would these tests be likely to differ strongly in difficulty (with addition being much easier), but they are actually testing different subject matter.

The notion of comparability of tests includes the comparability of the domains that they cover. A common way to facilitate comparability is to sample items according to blueprints that specify content, sometimes item difficulty and perhaps other item characteristics. To ensure that the tests are as similar as possible, the item bank could be stratified by content area and item difficulty. Then for a given type of test (e.g. general achievement in a subject), a certain proportion of items would be randomly drawn from each content area/difficulty stratum to ensure an appropriate balance of items was maintained.

The Angoff Method

Many standard setting systems around the world currently use the Angoff method to set cutscores that delineate levels of student performance. In this method, a panel of judges is employed to rate the items in a test. In attempting this task, the judges would have at their disposal a set of descriptors which articulate the types of knowledge and skills of students in each performance band. Some systems would have so-called Standards Packages, which give examples of the performance of past students in each performance band. For example, for past multiple choice items, the percentage correct may be given for borderline students at each performance band, along with the percentage choosing each option and the overall percentage correct. For constructed response items, the judges may be given sample answers and the score awarded to each answer.

Using this information, the judges are required to form an expectation of the type of work produced by students at a cutscore borderline. They are then asked to work through all test items and indicate how such borderline students would perform on each. For multiple choice items, or dichotomously scored items, their decisions would estimate the proportion correct. In practice, the judges may be asked to consider 100 such borderline students and to indicate how many of this group would be likely to get the item correct. For extended response items, the judges would estimate the mean or average score that the borderline students would obtain. In the first stage of this process, all decisions made by the judges are independent – the judges do not share information or observations with the other judges. To obtain the cutscore on the total test for a judge, all the item judgements are summed. For a given performance band, a single cutscore for the test is obtained by taking the mean or median of the judges' cutscores. This is the one-stage Angoff method as outlined in Angoff (1971).

Although the Angoff method was originally conceived as a one-stage test-centred process, it has now generally developed into a multi-stage procedure. In the first stage, the judges work independently. In later stages, they may receive data on their Stage 1 decisions and discuss the results. This group discussion process has been suggested by several researchers (Berk, 1996; Jaeger, 1982; Morrison, Busch and D'Arcy, 1994; Norcini, Lipner, Langdon and Strecker, 1987).

Other researchers have also suggested that the provision of data could inform the discussion (Cross, Impara, Frary and Jaeger, 1984; Linn, 1978; Norcini, Shea and Kanya, 1988; Popham, 1978). This sharing of data and group discussion could constitute a Stage 2. In some systems, a Stage 3 can occur, where work samples

for students near the cutscores can be provided. At each stage, the judges may modify their item ratings.

Regardless of how these Angoff ratings are obtained, when stored with other item data in the bank, they can be used to estimate whether a particular student receiving a randomly formed test has reached a given performance level.

The data stored with each item

In the type of item banking proposed here, the fundamental statistic of item difficulty is simply the proportion correct over the population of students (called its *p-value*). As has been frequently noted, this statistic is really an index of the easiness of the item, not its difficulty, so that it is sometimes called the item facility (Gower and Daniels, 1980). However, most workers in the field still refer to it as the item difficulty. This statistic is constantly being updated as the item is administered to new examinees. As more and more students attempt the item, the proportion correct becomes a better estimate of the proportion correct that would have been obtained had it been administered to the whole population. Apart from the *p-value*, each item would have a content identifier to enable the appropriate spread of items across content areas to be obtained.

In the item banking model outlined in this paper, each item would also have an Angoff rating for each performance band cutscore stored in the bank. For example, suppose that an education system was operating a standard setting based on six performance bands, denoted A, B, C, D, E and F. This requires five cutscores, X_A, X_B, X_C, X_D and X_E.

These five cutscores, the item proportion correct, *p*, and a content identifier would be stored for each item.

A uniform scale for reporting

In a standards-based system, a uniform reporting scale is required so that comparisons are meaningful. If a student is a borderline A on one test, and another student is a borderline A on a different test, then the students are regarded as equivalent in performance and should receive the same score. Therefore the cutscores on different tests, for the same performance band, should be scaled to the same value. An example of a suitable scale, based on a maximum possible score of 100 marks, is given below:

$$X_A \to 90, \quad X_B \to 80, \quad X_C \to 70, \quad X_D \to 60, \quad X_E \to 50.$$

Thus a borderline 'A' student would be scaled to a score of 90 regardless of the particular test attempted, and so on for the other performance bands.

Group or 'on-demand' testing

The test allocation system may be set up to provide different options. One option may be to enter a school code which provides the same randomly created test for a school class. That is, all students in the class would do the same test. This would allow useful feedback to be supplied at the group level, showing the strengths and weaknesses in different content areas of the school group in relation to system norms. A second option could provide for an 'on-demand' type of testing, where individual students would log on when they were ready to test their competence. This would be similar to a student taking a computer-based test of theory for gaining a driver's licence. The testing could occur at different times during the school year and the student would be given a unique test formed by random assignment. This on-demand testing would have the potential to allow the student to accelerate in certain educational modules.

How a student result would be calculated

Suppose Bill wants to test himself against the system norms for Mathematics. Using his school computer laboratory, he enters his username and PIN number and logs on to be given a web-based test. The items in this test are randomly drawn from an item bank as described above. A 100-item test, Test *X*, is administered. Table 1 below provides information about the items in this test and Bill's responses.

Each test item has a *p-value*, which when summed over all items in the test, gives a population mean estimate of 57 (/100). That is, if the system population of students attempted this particular randomly drawn test, it would be expected that they would average 57. The standard-setting system awards six performance bands, A, B, C, D, E and F. The item Angoff ratings for these band cutscores are stored for each item. Only the item ratings for Bands A and B are shown here. Summing the ratings for Band A, a borderline 'A' student would be expected to score 86 on this particular test. Similarly, a borderline 'B' student would be expected to score 73. Summing Bill's item scores, his total score is 77. Therefore he has achieved Band B standard but not Band A. After completing his online

Table 1: Statistics for Bill's attempt at a randomly generated test

Item	p-value	Angoff ratings for each performance band			Bill's scores
		Band A	Band B	…	
1	p_{x1}	a_{x1}	b_{x1}	.	$x_1=1$
2	p_{x2}	a_{x2}	b_{x2}	.	$x_2=0$
3	p_{x3}	a_{x3}	b_{x3}	.	$x_3=1$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
100	p_{x100}	a_{x100}	b_{x100}	.	$x_{100}=0$
	$\bar{X}=\sum p_{xi}=57$	$X_A=\sum a_{xi}=86$	$X_B=\sum b_{xi}=73$		$X=\sum x_i=77$

test, the computer screen presents him with a testamur stating that on this Mathematics test, he has achieved a system-wide award of Band B and displays the descriptors for a Band B performance. He is able to print out this testamur and it becomes part of his portfolio of achievement.

Calculating a scaled score for each student

An educational body may wish to take a further step and award an actual score to Bill which reflects his system-wide standing. Recall that in our hypothetical reporting system, $X_A \to 90$, and $X_B \to 80$. Then using linear interpolation, a score *X*, lying between the cutscores, would be converted to the statewide scale by

$$X' = 80 + (90-80)\frac{(X - X_B)}{(X_A - X_B)}. \quad (1)$$

In Bill's case, his raw score of 77 is converted to a scaled score as follows:

$$X' = 80 + (90-80)\frac{(77-73)}{(86-73)} = 83.$$

Note that for raw scores that fall in the top performance band, the additional point $100 \to 100$ may be used as an anchor point, while for raw scores that fall in the bottom performance band, $0 \to 0$ may be used.

Setting up the Bank

There are many ways to set up an item bank. Some items may be obtained from past system-wide testing programs, with their *p-value*s and Angoff ratings being known from these administrations. Another method, which is relatively cost efficient, is to phase in the items and gather statistics through practice tests, where the feedback students are given would not relate to system-wide norms. An educational organisation may offer practice tests on its website for students to assess their knowledge and prepare for other more formal tests. Schools could be directed to attempt these tests as part of their assessment program and the feedback they receive on their students' performance would comprise raw scores and a breakdown of scores across topic areas. This limited feedback may still be considered useful by schools and teachers would welcome the provision of tests that they do not have to set. The items would be kept secure.

After stable *p-value*s have been gathered, the items could be reviewed by subject experts to provide Angoff ratings for each performance band. It is a moot point as to how much data (if any) is provided to the judges before they give their Angoff ratings. See for example, Busch and Jaeger, 1990. This policy would vary between educational systems. As the Angoff ratings are the basis for equating different randomly formed tests under this system, it is suggested here that the judging panel could have access to the empirical item *p-value*s, before they make their judgements. If auxiliary information is available – for example, statewide scores on some other measure (Test *Y*), then this could also be incorporated. For each particular item to be rated, the observed item *p-value*s, for given subgroups of students (at particular percentiles on *Y*), could be useful information for the judges.

Regardless of how it is done, each item would receive an Angoff rating for each performance band cutscore. When enough such items have been obtained, they would be moved to the operational program that gives approximate normative feedback for each student.

Adding items to the Bank

After the item bank has been established, new items may be written that are to be added to the bank. These items need to be administered to students so that their difficulty (*p-value*) can be determined. This may be accomplished with a scheme where the candidate attempts the usual number of items but their scaled score is based on a slightly smaller subset of these items (for example, 98 items instead of 100). The two extra items are seeded into this test, and in many other randomly formed tests, so that data on item *p-value*s may be accumulated.

When a new item has been administered to sufficient students (say 1000 times), its *p-value* is then regarded as sufficiently reliable to be used as part of the official score and its status is changed from that of a new item to that of an operational item.

Comparability Issues

The use of an item bank based on classical test statistics and Angoff item ratings may not give the level of comparability obtained under an item bank based on IRT but would be suitable for lower stakes testing. If an IRT bank is properly implemented and administered, then the *b* values (item difficulty) would be population independent. The classical item p-values are not. The classical bank relies on the random sampling from items stratified by content and difficulty. If this is well implemented, then the amount of variation in the difficulty of any test administered will be limited by the uniformity of item difficulties within each stratum. This is shown in Equation (2) below which gives the variance of the means of the randomly formed tests in terms of the variance of the p-values in each stratum, where a stratum is denoted by *j* and there are *m* strata. (Equation (2) is derived in the Appendix).

$$\sigma^2_{\bar{X}} = \sum_{j=1}^{m} n_j \sigma^2_{p_j} \ . \qquad (2)$$

In the extreme case, where the items in each stratum are of equal difficulty (zero variance of the p-values), (2) shows how the randomly formed tests will be of equal difficulty (zero variance of the test means).

A key issue in the effectiveness of the bank is the accuracy of the Angoff item ratings and how well they function in adapting to tests of varying difficulty. Apart from the above control through stratification, the Angoff item ratings for a particular achievement band and the item p-values can be compared to see whether they are consistent. For example, a particular item, *i*, might be somewhat easier than the other items in a stratum. Suppose the average stratum p-value was 0.5 and the particular item had a p-value of 0.6. One would expect that the corresponding item Angoff rating would also reflect this. If the Angoff ratings for Band C averaged 0.5 over the stratum, then one would expect that the Angoff Band C rating for the particular item to be in the vicinity of 0.6, or (at the least) higher than the stratum average.

Probably the best way to compare the p-values and the Angoff item ratings for a particular band is to plot them with the p-values on the X-axis and Angoff ratings on the Y-axis. This scatterplot would then reveal Angoff ratings that were inconsistent with the p-values. One could use 95% confidence intervals to determine the outliers, and refer these back to the judges for a re-assessment of their Angoff ratings.

In the operation of the bank, a similar check can be used to determine whether a particular test drawn has a difficulty that is consistent with the Angoff cutscore. Suppose a test designated for Bill was more difficult (in mean value) than the average test drawn from the bank. In addition, suppose that this particular item combination gave an Angoff cutscore that was inappropriately higher than that for an average test. Then in this extreme scenario, Bill would be doubly disadvantaged by receiving a harder test and a more stringent cutscore. A statistical check can be implemented to look for cases like this and, if detected, a redraw of items would be performed. The examinee of course would be unaware of these behind-the-scenes checks to ensure test fairness and the redrawing of the test items.

Conclusion

As the item banking program proceeds, the item difficulty of each item would be monitored at periodic intervals. Some items may need to be retired from the bank due to changes in the curriculum as they become outdated. Other items may become over-exposed and become too easy as a consequence. New items would be written and incorporated into the bank. As these processes occur, the bank needs to be monitored to

examine their effects. If it is desired that the original weighting of content areas and item difficulties be approximately maintained, then the ratios of items across the content areas would be periodically examined. In addition, the mean difficulty of each stratum would also be monitored. The newly written items would then be targeted to maintain the balance of the original bank. However, this can only be done very approximately and the bank may change slightly in mean difficulty over time. This can be easily monitored.

In theory any change in mean difficulty should not matter for obtaining comparable scores over time. The Angoff item ratings for a given performance band should reflect when the items get harder. If the items become more difficult, then the judges should be able to allow for this and lower their item cutscores. In practice, however, one would not want the bank to change too much in difficulty in case this does affect the item ratings (for example, see Bejar, 1983; Goodwin 1999) – hence the importance of the checks outlined in the previous section.

Such a system could be implemented for low-stakes testing and would allocate students to performance bands, delivering instant feedback on the band attained and the descriptors giving the performance characteristics of typical students in the band. In systems where several levels of performance are described, a system of linear interpolation may be used to give an approximate score on the statewide scale.

References

Attali, Y. & Burstein, J. (2006). Automated Essay Scoring with e-rater V.2. *Journal of Technology, Learning, and ssessment, 4*(3). Retrieved 16 June, 2009 fromhttp://escholarship.bc.edu/cgi/viewcontent.cgi?article=1049&context=jtla

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement*. (2nd ed., pp. 508-600). Washington, D.C.: American Council on Education.

Bejar, I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, 7, 303-310.

Berk, R. (1996). Standard setting: the next generation. *Applied Measurement in Education*, *9*, 215-235.

Burstein, J. (2003). The e-rater scoring engine: Automated Essay Scoring with natural language processing. In M. D. Shermis and J. C. Burstein (Eds.), *Automated Essay Scoring: Across disciplinary approach* (pp. 113–121). Mahwah, NJ: Lawrence Erlbaum Associates.

Busch, J. C., and Jaeger, R. M. (1990). Influence of type of judge, normative information, and iscussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement*, *27*, 145-163.

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.

Educational Testing Service (2006). New technologies improve automatic scoring. *Innovations, Issue 2*, p. 3.

Glass, G. V. and Stanley, J. C. (1970). *Statistical Methods in Education and Psychology*. Prentice-Hall, New Jersey.

Goodwin, L. D. (1999). Relations between observed item difficulty levels and Angoff Minimum passing levels for a group of borderline examinees. *Applied Measurement in Education*, *12*, 13-28.

Gower, D. M. and Daniels, D. J. (1980). Some Factors which Influence the Facility Index of Objective Test Items in School Chemistry. *Studies in Educational Evaluation, 6,* 127-136.

Jaeger, R. (1982). An Iterative Structured Judgment Process for Establishing Standards on Competency Tests of Theory and Application. *Educational Evaluation and Policy Analysis*, *4*, 461-475.

Linn, R. (1978). Demands, cautions and suggestions for setting standards. *Journal of Educational Measurement, 15*, 301-308.

Morrison, H., Busch, J. and D'Arcy, J. (1994). Setting reliable national curriculum standards: a guide to the Angoff procedure. *Assessment in Education*, *1*, 181-199.

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

Norcini, J., Shea, J. and Kanya, D. (1988). The effect of various factors on standard setting. *Journal of Educational Measurement, 25*, 57-65.

Popham, W. (1978). As always provocative. *Journal of Educational Measurement, 15*, 297-300.

Rudner, Lawrence (1998). Item banking. *Practical Assessment, Research & Evaluation*, 6(4). Retrieved June 15, 2009 from http://PAREonline.net/getvn.asp?v=6&n=4.

APPENDIX
Variation in difficulty of tests drawn by stratified random sampling

Suppose a large bank comprised *m* strata, each stratum with items homogeneous in content and difficulty. Let σ_{p_j} be the standard deviation of the *p-values* of all items in Stratum *j*.

Suppose that n_j items were randomly sampled from Stratum *j* and the mean of their *p-values* calculated. If this sampling process were to be repeated many times, then by a well-known formula (for example, Glass and Stanley, 1970), the variance in the mean *p-value* across samples would be given by

$$\sigma^2_{\bar{p}_j} = \frac{\sigma^2_{p_j}}{n_j}. \tag{A1}$$

The items drawn in a particular sample from Stratum *j* may be regarded as a mini-test with mean, $\bar{X}_j$, the mean being derived as the sum of the *p-values*. Thus we may write

$$\bar{p}_j = \frac{\sum p_j}{n_j} = \frac{\bar{X}_j}{n_j}. \tag{A2}$$

From (A1) and (A2):

$$\sigma^2_{\bar{X}_j} = n_j \sigma^2_{p_j}. \tag{A3}$$

This gives the variance of the means of the mini-tests drawn from Stratum *j* by repeated sampling.

A total test is formed by repeating this sampling process in all other strata and pooling the items. For a particular sample, the total test mean is given by the sum of the means of each mini-test drawn from its stratum. That is

$$\bar{X} = \sum_{j=1}^{m} \bar{X}_j. \tag{A4}$$

As the sampling in one stratum is independent of that in all other strata, then the correlations between pairs of stratum means will be zero. Hence from (A4), the variance of the total test means obtained by repeated sampling is given by

$$\sigma^2_{\bar{X}} = \sum_{j=1}^{m} \sigma^2_{\bar{X}_j}. \tag{A5}$$

Substituting from (A3):

$$\sigma^2_{\bar{X}} = \sum_{j=1}^{m} n_j \sigma^2_{p_j}. \tag{A6}$$

This equation gives the variance in test means that would be obtained across randomly formed tests in terms of the variances of item *p-values* within each stratum. The latter is known. The item bank stores all the item *p-values* in a stratum, and their variance can be calculated. Equation (A6) shows the importance of stratification. If the strata can be made quite homogeneous for item difficulty, then the p-value variance will be small for each stratum, resulting in a small variance of the test means.

Note

This paper was written whilst the first author was a Visiting Research Fellow at Oxford University Centre for Educational Assessment, United Kingdom.

Citation

MacCann, Robert G. and Stanley, Gordon (2009). Item Banking with Embedded Standards. ***Practical Assessment, Research & Evaluation***, 14 (17). Available online: http://pareonline.net/getvn.asp?v=14&n=17.

Corresponding Author

Robert G. MacCann
Oxford University Centre for Educational Assessment
15 Norham Gardens
Oxford OX2 6PY
UK.

Email: victoria.hayman [at] ox.ac.uk or robert.maccann [at] optusnet.com.au