JOURNAL OF APPLIED MEASUREMENT, *10*(4), 438-454

Standard Setting with Dichotomous and Constructed Response Items: Some Rasch Model Approaches

Robert G. MacCann
Oxford University
Centre for Educational Assessment

Using real data comprising responses to both dichotomously scored and constructed response items, this paper shows how Rasch modeling may be used to facilitate standard-setting. The modeling uses Andrich's Extended Logistic Model, which is incorporated into the RUMM software package. After a review of the fundamental equations of the model, an application to Bookmark standard setting is given, showing how to calculate the bookmark difficulty location (BDL) for both dichomotous items and tests containing a mixture of item types. An example showing how the bookmark is set is also discussed. The Rasch model is then applied in various ways to the Angoff standard-setting methods. In the first Angoff approach, the judges' item ratings are compared to Rasch model expected scores, allowing the judges to find items where their ratings differ significantly from the Rasch model values. In the second Angoff approach, the distribution of item ratings are converted to a distribution of possible cutscores, from which a final cutscore may be selected. In the third Angoff approach, the Rasch model provides a comprehensive information set to the judges. For every total score on the test, the model provides a column of item ratings (expected scores) for the ability associated with the total score. The judges consider each column of item ratings as a whole and select the column that best fits the expected pattern of responses of a marginal candidate. The total score corresponding to the selected column is then the performance band cutscore.

Requests for reprints should be sent to Robert MacCann, Oxford University Centre for Educational Assessment, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY, United Kingdom, e-mail: victoria.hayman@education.ox.ac.uk.

The last two decades have seen the growth and evolution of standard-setting methods in an attempt to improve the quality of education. In the broadest sense, this involves recasting the curriculum in greater detail than was previously done so that the standards of achievement required of students are more explicitly mapped. It also involves ensuring that the testing reflects this curriculum, and requires the reporting of student achievement in a way that relates to the standards. This reporting usually makes use of performance descriptors, accompanied by work samples (where appropriate) describing what the typical student can do at each proficiency level.

In the New South Wales (NSW) Australia education system, an important part of capturing and clarifying standards is the production of Standards Packages. These comprise a comprehensive set of examples of student responses to past examination questions at each level of proficiency, and are disseminated widely to schools and educational organizations on compact disks. By studying these packages, students can gain a clearer idea of the characteristics of the typical answer at each proficiency level and are able to more efficiently address weaknesses in their preparation for future examination attempts. Teachers are also able to use them to acquire an understanding of the state standards at each proficiency level and can adjust their teaching accordingly.

An important aspect of standard setting is establishing cutscores which separate the levels of proficiency on the test score distribution. This process involves a team of judges, carefully chosen to satisfy the needs of the particular educational system. These teams require an appropriate mix of representativeness and expertise. Two widely used methods of standard setting are the Angoff method (Angoff, 1971) and its variations, and the more recent Bookmark method (Mitzel, Lewis, Patz and Green, 2001). The Angoff method was originally conceived as a one-stage process, focusing on the difficulty of the test items and requiring the judges to estimate how students at a particular proficiency level would perform on the items. However, several researchers have advocated group discussion as a means of the team members acquiring a more complete understanding of the criteria and reasoning used by others in allocating their item ratings (Berk, 1996; Jaeger, 1982; Morrison, Busch, and D'Arcy, 1994). In addition, other researchers have suggested supplying the judges with data and giving feedback on their judgments (for example, Linn, 1978; Norcini, Shea and Kanya, 1988; Popham, 1978). Thus the Angoff procedure has now typically developed into a multi-stage process where, for example, the judges independently make their judgments in Stage 1, discuss their decisions in Stage 2 (often with statistical data from Stage 1), and confirm their decisions in Stage 3 (often with work samples of borderline students).

In contrast with the Angoff method, which can be implemented without item response theory (IRT) procedures, the Bookmark method was designed to use IRT from the beginning. In this paper, Rasch modeling will be used to illustrate the Bookmark method for a test comprising a mix of item types. In addition, this example will be used to illustrate three methods of applying Rasch modeling to improve the Angoff standard-setting procedures.

Methods

The Data

The applications of Rasch modeling to standard setting were conducted on data from the Year 12 testing program in NSW, Australia. This program is primarily attempted by school leaver students, generally of age 17-18 years, although it is open to, and attempted by, a range of mature-age candidates. Many of these students go on to college. The test items were taken from the Economics external test, a non-compulsory test containing a mixture of multiple-choice items (scored 0 if incorrect, 1 if correct) and constructed response items of varying mark values. The latter items required an open-ended response varying from a few lines, up to a page, depending on the mark value. The test totaled 50 marks over all items. The test items are summarized in Table 1 below. The analyses described below were based on a simple random sample of 5000 students from the Economics course candidature.

Table 1

Structure of the test on which the examples are based

Part	Items	Mark value
Part A (20 marks)	20 items (A1 to A20)	1 each
Part B (30 marks)	B1	2
	B2	2
	B3	2
	B4	2
	B5	2
	B6	2
	B7	4
	B8	4
	B9	5
	B10	5

Rasch Modeling

The Rasch model was developed independently of the other IRT models by Rasch (1960), with a clear philosophy of measurement in mind. In one sense, it may be regarded as the simplest member of the family of IRT models, a model having only one item parameter—its difficulty. In another sense, it is different in that the sought-for property of *specific objectivity* is obtained (if the data fit), permitting the separation of person ability estimation and item difficulty estimation (Rasch, 1960, 1966). The model was popularized by Wright and his colleagues (Wright and Panchapakesan, 1969; Wright, 1977; Wright and Stone, 1979). It was later extended to handle polychotomous items by Andrich (1978, 1982, 1988) for his Rating Scale and Extended Logistic models, and by Masters (1982) and Masters and Wright (1984) for the Partial Credit model.

In the NSW education system, Andrich's Extended Logistic Model (Andrich, 1988) is widely used for test analysis. This model has been incorporated into the RUMM software package—Rasch Unidimensional Measurement Models (Andrich, Sheridan, Lyne, and Luo, 2000). The RUMM software accepts raw data files with the students in rows and the items in columns. It produces a person ability estimate for each student on the *logit* scale (log odds units), a scale ranging from minus infinity to plus infinity. Although this scale stretches from minus to plus infinity, in practice most of the values range from about –5 to +5 logits.

For dichotomous items, RUMM produces an item difficulty estimate for each item, also on the logit scale. For polychotomous items, it again produces an item difficulty estimate on the logit scale, but also produces *item threshold estimates*, one for each item mark value that exceeds zero. The item thresholds, when combined with the item difficulty, indicate the ability required to just reach the next mark level. A study of the relative gaps between these thresholds can give considerable insight into how difficult it is for students to obtain each mark.

Model for Polychotomous Items. Consider a constructed response item (j) where m_j is the value of the largest score category on the item. Then for students (i) of ability θ_i, the probability that an observed score on the item will equal a particular value (x) is given by:

$$P(X_{ij} = x) = \frac{e^{x(\theta_i - \delta_j) - \tau_{j1} - \tau_{j2} - \ldots - \tau_{jx}}}{\sum_{k=0}^{m_j} e^{k(\theta_i - \delta_j) - \tau_{j1} - \tau_{j2} - \ldots - \tau_{jk}}}, \quad (1)$$

where δ_j is the item difficulty and the τ_{jk} are the centralized item thresholds. The latter are associated with the ability required to just reach a particular score level.

Using (1), for an item worth 3 marks (m_j = 3) the denominator (D) becomes:

$$1 + e^{1(\theta_i - \delta_j) - \tau_{j1}} + e^{2(\theta_i - \delta_j) - \tau_{j1} - \tau_{j2}} + e^{3(\theta_i - \delta_j) - \tau_{j1} - \tau_{j2} - \tau_{j3}}.$$

Then the probabilities of obtaining 0, 1, 2 and 3 respectively are given by:

$$P(X_{ij}=0)=\frac{1}{D},$$

$$P(X_{ij}=1)=\frac{e^{1(\theta_i-\delta_j)-\tau_{j1}}}{D},$$

$$P(X_{ij}=2)=\frac{e^{2(\theta_i-\delta_j)-\tau_{j1}-\tau_{j2}}}{D},$$

$$P(X_{ij}=3)=\frac{e^{3(\theta_i-\delta_j)-\tau_{j1}-\tau_{j2}-\tau_{j3}}}{D}.$$

Note that the denominator is equal to the sum of the numerators. Hence all four probabilities sum to 1, as they must. The above equations use the *centralized thresholds* from RUMM, which sum to zero. RUMM also produces uncentralized thresholds which are equal to the sum of the item difficulty and the centralized threshold:

$$\tau_{jk}(\textit{uncentralized}) = \delta_j + \tau_{jk}. \qquad (2)$$

If (2) is substituted into (1), so that the latter equation is written in terms of uncentralized thresholds, then δ_j is cancelled out and disappears from the equation. RUMM provides a toggle to switch between centralized and uncentralized thresholds. This paper presents the equations in terms of the centralized thresholds. Whenever the item difficulties (deltas) are explicitly shown in the equations, it implies that the centralized thresholds are being used. When the uncentralized thresholds are used, the deltas are not part of the equations. Naturally the same results are obtained, however the equations are represented.

Model for Dichotomous Items. From Equation (1), the model simplifies for items scored (1, 0). For these item types, only one threshold τ_{j1} is required, but *this equals zero if it is a centralized threshold*. Therefore (1) reduces to the following:

$$P(X_{ij}=x)=\frac{e^{x(\theta_i-\delta_j)}}{\sum_{k=0}^{1} e^{k(\theta_i-\delta_j)}}. \qquad (3)$$

Thus the probability of getting the item wrong (given θ) is:

$$P(X_{ij}=0)=\frac{1}{1+e^{(\theta_i-\delta_j)}}. \qquad (4)$$

The probability of getting the item right (given θ) is:

$$P(X_{ij}=1)=\frac{e^{(\theta_i-\delta_j)}}{1+e^{(\theta_i-\delta_j)}}. \qquad (5)$$

The Relationship between Ability and Total Score. The conversion of ability to total score for the Economics data is shown in Figure 1. An important property of Rasch models is that, if the items are compulsory (as in our example), then the total score is a sufficient statistic for estimating student ability (Andersen, 1973). For complete data, it follows that all students on the same total score will receive the same ability estimate, and vice versa. This property does not hold for non-Rasch models if item pattern scoring is used (which is the optimal scoring for such models). With item pattern scoring, the ability is determined by the specific pattern of the responses—for such models the graph below would have a similar shape but would resemble a scatterplot rather than a line.

The conversion from total score to ability estimate in logits is shown in Figure 2. This conversion is nearly linear for most of the mark range, but as one approaches the extremes, the line curves sharply. Thus, a small total mark difference near the extremes can result in a relatively large difference in ability estimates. For students gaining zero marks or full marks, it is difficult to justify giving an ability estimate.

Expected Score on an Item. The conversions given in Figures 1 and 2 are usually provided by reputable software packages, but can be calculated from Equation (1) as follows. First, (1) can be used to estimate the *expected score on an item* for students at ability level θ. The expected item score is obtained by multiplying the probability of gaining each score (from (1)) by the score value itself and summing over all possible score values. That is, for item *j*:

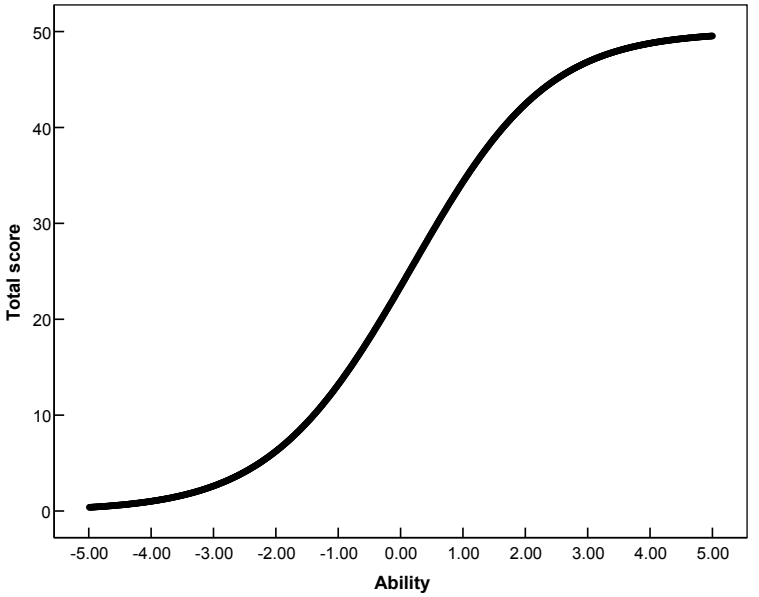

Figure 1. Converting from ability (logits) to total score.

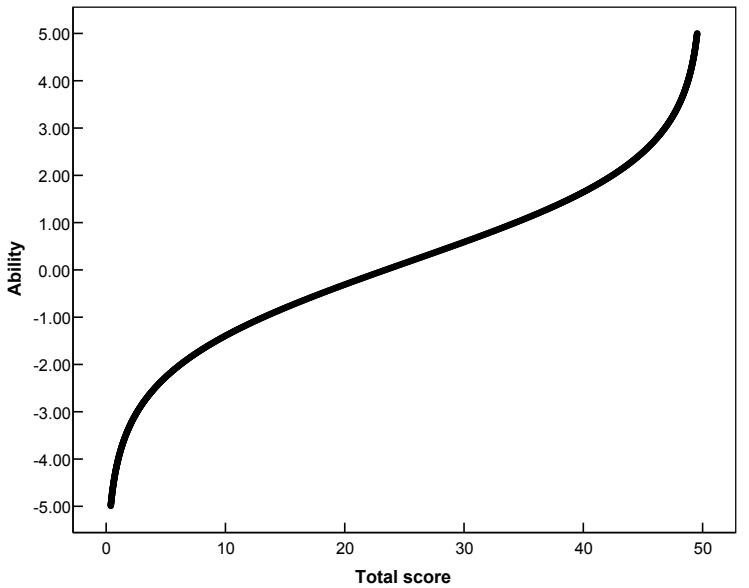

Figure 2. Converting from total score to ability (logits).

$$\mu(X_j \mid \theta) = \sum_{k=0}^{m_j} kP(X_{ij} = k) \cdot \qquad (6)$$

Note that in the *dichotomous case*, the probability of being wrong is associated with a score value (k) of zero. For this case, from (6), the expected score on the item is simply the probability of being correct. Having obtained the expected score on each item from (6), *the expected score on the total test* can be calculated.

Expected Score on the Total Test. Let Y denote the total test score. The expected total test score (for a given ability) is obtained by summing the expected item scores over the N items. This gives:

$$\mu(Y \mid \theta) = \sum_{j=1}^{N} \mu(X_j \mid \theta). \qquad (7)$$

The relationship in (7) is similar to that graphed in Figure 1, yielding the same integer scores, but also giving expected scores that can take fractional values. The conversions in Figures 1 and 2 are often used in Rasch modeling applications.

The Bookmark Method

Dichotomous Scoring. The Bookmark method will first be discussed for the case of dichotomously scored items, as this case is greatly simplified under Rasch modeling. The Bookmark method uses the Rasch model to arrange the items in order from easiest to hardest, these items being presented to the judges in an ordered booklet. The judges' task is to work through the booklet from the beginning, stopping at the item where the borderline student (at a given proficiency level) has a probability of success that just falls below a criterion probability of success. This criterion probability of success is called the *response probability (RP)*. It is often set at two-thirds (for example, Reckase, 2000; Mitzel et al., 2001) but is sometimes set at 0.5 (Wang, 2003). The judges place a bookmark *just before* this critical item. In this probabilistic sense, the borderline student will tend to be successful on items before the bookmark and unsuccessful on items after the bookmark.

The ordering of the items in the book requires the notion of the *bookmark difficulty location (BDL)*. Given the item difficulty δ_j, the *BDL* is the ability level required so that the probability of success on the item is equal to the *response probability*. Although this measure is called a difficulty location, it is defined as an ability level, the ability and difficulty being measured on the same logit scale.

Rearranging Equation (5), and letting P denote the probability of success on the item we get:

$$\theta_j = \delta_j - \ln\left(\frac{1-P}{P}\right), \qquad (8)$$

where ln is the natural logarithm. Substituting RP for P and BDL for θ gives:

$$BDL_j = \delta_j - \ln\left(\frac{1-RP}{RP}\right). \qquad (9)$$

For example, putting RP=2/3 gives $BDL_j = \delta_j + 0.69315$. That is:

$$BDL_j = \delta_j + constant. \qquad (10)$$

For the Rasch model with dichotomous items, the *BDL* differs only by a constant from the item difficulty. Thus Bookmarking with dichotomous items using the Rasch model is conceptually simple. As the *BDL* is equal to the item difficulty plus a constant, it will *rank the items in the same order as the item difficulty*.

If a response probability of 0.5 is used in (9) then:

$$BDL_j = \delta_j . \qquad (11)$$

Here the *BDL* is exactly equal to the item difficulty, which further simplifies the process. In this case, the concept of the *BDL* can be dispensed with altogether and the explanation of the method to the judges can be made solely in terms of item difficulties—a much easier process. The latter method is sometimes presented in a form known

as *Item Mapping* where the item difficulties are shown in graphical form in a histogram, with the histogram columns labeled with the item numbers. This compact presentation enables the judges to view all the data with relative ease. The judges must then place a bookmark between two columns of the histogram. See Wang (2003) and MacCann and Stanley (2006) for examples.

In contrast to the Rasch model, other IRT models do not have this desirable property—that the *BDL* ranks the items in the same order as the item difficulty. For example, with a 2-parameter model, the slope (discrimination) of the item characteristic curve may vary substantially from item to item. An item with a shallow slope may be more difficult for high ability students than an item with a steeper slope, even though the former may have a lower item difficulty index.

Finding BDLs for a mixture of Dichotomous and Constructed-Response Items. For constructed-response items, a scoring rubric is a brief description of what students are required to have demonstrated in their answers in order to obtain a particular mark. When a constructed-response item is included in a test, the *BDL of each mark level (k)* on the item is calculated. Both the item and the rubric corresponding to the particular mark level are then placed in the booklet in *BDL* order. Thus, for a given item, the rubric for a mark of 1 may appear fairly early in the booklet and the rubric for a mark of 2 would appear later in the booklet, perhaps separated by several other items.

The *BDL* for a constructed-response item is the ability for which the probability of obtaining a mark of *k or above* is equal to the *RP*. In short, at what value of θ does $P(X \geq k) = RP$? Beretvas (2004) has derived *BDL* formulas for various IRT models applied to constructed-response items, up to a maximum mark of 4, the latter requiring the solution of polynomial equations of the fourth power. Beyond this maximum, she notes that there is no algebraic formula that solves a polynomial of the fifth or higher power.

An easier and more generally applicable approach is to *construct a table* and estimate θ and *Y* through interpolation. As part of this process, it is also convenient to calculate the expected score on each item and the expected score of the total test for each θ value.

(i) Construct a table of θ values ranging from –5 to +5 logits with a step size of 0.01 logits, say.

(ii) Calculate $P(X = k)$, for each θ value, from (1).

(iii) Calculate the value of $P(X \geq k)$, for each θ value. For example, suppose the maximum possible score on an item is *m*. Then
$$P(X \geq k) = P(X = k) + P(X = (k+1)) + \ldots + P(X = m).$$

(iv) Calculate the expected score on each item, for each θ value, from (6).

(v) Calculate the expected score on the total test, for each θ value, from (7).

(vi) Treat the *RP* as though it were a value in the $P(X \geq k)$ data field, and linearly interpolate to obtain the estimate of θ.

(vii) Treat the *RP* as though it were a value in the $P(X \geq k)$ data field, and linearly interpolate to obtain the estimate of *Y*.

Once this method is programmed, it can be used for all items, both dichotomous and constructed response. For dichotomous items, from (5), step (iii) above requires the calculation of

$$P(X \geq 1) = P(X = 1) = \frac{e^{(\theta_i - \delta_j)}}{1 + e^{(\theta_i - \delta_j)}}. \quad (12)$$

The Bookmarking Process. The method outlined above was applied to the Economics data using a response probability of 0.5, with the judges targeting the borderline for minimal competence. The *BDL*s for the lower ability range are shown below in Table 2. The number following the underscore in the constructed response items indicates a particular mark level. For example, B5_1 represents obtaining a mark of 1 in item B5.

The heavy dashed line indicates where a judge has placed the bookmark—in a probabilistic sense, the minimally competent student would be just expected to gain 2 marks on item B9 but would probably get item A11 wrong.

The type of booklet that the judges receive is given in Figure 3, which shows a listing of example items. The dichotomous items appear once in the booklet. A constructed response item worth *m* marks appears *m* times in the booklet, for the mark values 1, 2, …, *m*. The constructed response item itself is listed along with its scoring rubric. At each listing, the appropriate part of the rubric is shaded to indicate the mark value reached at that *BDL* point. In the example above, a judge has estimated that the borderline candidates would have a greater than 0.5 chance (the response probability) of getting a mark of 2 on item B9. However, the judge believes that the borderline candidates have a less than 0.5 chance of getting item A11 correct. Hence the bookmark is placed *before* item A11. Referring back to Table 2, this corresponds to a *BDL* of –0.83 and a cutscore of 14.7.

Each judge would independently place a bookmark in the first stage of the standard setting. Then if a second stage were employed, there would usually be consultation between judges, after which they would be given the opportunity to independently change their bookmark positions. Once the bookmarks are finalized, the various *BDL*s are averaged across judges and this average is then converted to a cut score by the Figure 1 conversion table.

Angoff Methods

The Angoff method derives from a brief comment and a footnote by Angoff (1971) in a chapter mainly devoted to test equating and calibrating issues. He considers whether a hypothetical minimally competent student could answer each dichotomously scored item in a test. If such a person could answer an item, a mark of 1 is given; otherwise, 0. The sum of such scores over the whole test gives the cutscore for minimal competence. In the footnote, he modifies this slightly to consider a *group of such persons* and the judge must estimate the proportion of this group that would be correct on an item. This simple system has evolved over time into a multi-stage process where, after initial independent judgments, the judges meet to discuss their decisions and receive feedback on the results.

In the NSW system, six judges generally comprise a standard-setting panel for a particular course, all being experienced teachers from a range of government and private schools. Prior to the standard setting, each judge can refresh their mental image of the borderline student at a given proficiency level through their study of the performance descriptors and the exemplars in the Standards Package. The judges are asked to base their estimates on *a group*

Table 2

Bookmark difficulty locations (BDLs) for Economics (RP=0.5)

Item	BDL (ability)	Equivalent mark (/50)
B5_1	–2.28	4.9
B8_1	–1.89	6.8
B2_1	–1.88	6.9
B9_1	–1.55	8.9
A10	–1.23	11.3
A4	–1.22	11.4
B6_1	–1.10	12.3
A19	–1.08	12.5
A3	–1.04	12.8
B10_1	–1.02	13.0
A5	–0.94	13.7
B9_2	–0.83	14.7
A11	–0.69	16.1
B7_1	–0.62	16.8
B8_2	–0.54	17.6
A20	–0.41	18.9
A1	–0.37	19.3
B5_2	–0.36	19.4
B10_2	–0.14	21.8

of borderline students, as in the Angoff footnote. For dichotomous items, the judges are asked to estimate the proportion of the group that would answer the item correctly. For constructed-response items they are asked to estimate the average score that the group would obtain on the item. When the Angoff ratings are finalized, they are summed over all the items on the test to produce a total cutscore for each judge. The final total cutscore is obtained by averaging over the results of the six judges.

There are a number of ways in which Rasch modeling can assist the judges in the Angoff method. These vary in the amount of direction that the judges are given and the way the data is displayed. In the next method to be described, the process fits well with a traditional Angoff multi-stage procedure, with few constraints being placed on the judges.

Judge Ratings versus Rasch Estimates. In Stage 1 of the Angoff method, the judges form their independent cutscores on each of the items in the form of proportions, for a given proficiency borderline. These proportions are summed to give a total cutscore for each judge and averaged

Part A Q5 (1 mark)
Economic growth in developing countries is most likely to be increased through:

(a) Reducing the level of foreign investment
(b) Increasing the level of education
(c) Reducing the level of aggregate demand
(d) Increasing the rate of taxation.

Part B Q9 Students gaining 2 marks (/5)
Define what is meant by the term "inflation" and list actions that the Federal Government could take to contain inflationary pressures.

Marking Guidelines Criteria	Marks
• Gives a clear and precise definition of inflation. • Lists three or more Government actions that would help contain inflation.	5
…….	4
…….	3
• Conveys some knowledge of the term "inflation" and lists only one action that the Government could use to contain it.	2
…….	1

Judge places bookmark here

Part A Q11 (1 mark)
If the Consumer Price Index increases from 150 to 157 in one year, the rate of inflation in that time interval is given by:

(a) 4.7%
(b) 7.0%
(c) 4.5%
(d) 5.7%.

Figure 3. Items and item rubrics listed in BDL order.

across the six judges to obtain a single total test cutscore. Then from Figure 2, this cutscore can be converted into a Rasch ability in logits. This Rasch ability can then be used to estimate how such students would perform on each of the items in the test (the expected scores).

As noted previously, for dichotomous items the *expected score on the item* (for a given θ) is simply the probability of getting it correct. This is given in Equation (5). For polychotomous items, the *expected score on the item* is given by Equation (6). This requires the prior use of Equation (1) to calculate the probability of gaining a particular score point, given θ.

From (5) and (6), a table can be constructed comparing the expected score on each item with the Angoff ratings of the judges. Both individual judge ratings and the mean of these ratings across judges can be compared to the Rasch expected values. The latter is shown in Table 3.

The proficiency borderline being estimated in this example is the highest band level (Band 6) in the NSW system. The total cutscore (average of six judges) was 40.48 (out of 50 items) which corresponds to an ability in logits of 1.71. As the test items varied in their maximum possible score (m), the difference between the judge rating and the Rasch expected value was divided by m.

This type of table can allow the judges to see where they have rated an item as too easy or too difficult compared to the Rasch estimates. For example, item A11 is underlined to indicate a relatively large discrepancy in the multiple choice items. The judges thought that about 83%

Table 3

Judges' ratings versus Rasch expected scores

	Item cutscores			
Item	Maximum	Judges	Rasch	Difference/m
A1	1	0.900	0.889	0.011
A2	1	0.800	0.822	–0.022
A3	1	0.908	0.940	–0.032
A4	1	0.917	0.949	–0.033
A5	1	0.875	0.934	–0.059
A6	1	0.800	0.781	0.019
A7	1	0.792	0.784	0.007
A8	1	0.767	0.743	0.024
A9	1	0.808	0.832	–0.024
A10	1	0.900	0.950	–0.050
A11	1	0.833	0.917	–0.083
A12	1	0.800	0.773	0.027
A13	1	0.800	0.804	–0.004
A14	1	0.700	0.711	–0.011
A15	1	0.717	0.727	–0.010
A16	1	0.808	0.819	–0.010
A17	1	0.725	0.767	–0.042
A18	1	0.808	0.817	–0.009
A19	1	0.875	0.942	–0.067
A20	1	0.817	0.893	–0.076
B1	2	1.567	1.570	–0.002
B2	2	1.600	1.847	–0.124
B3	2	1.583	1.594	–0.005
B4	2	1.533	1.207	0.163
B5	2	1.667	1.899	–0.116
B6	2	1.633	1.699	–0.033
B7	4	3.200	3.008	0.048
B8	4	3.267	2.917	0.087
B9	5	4.000	3.978	0.004
B10	5	4.083	3.970	0.023

of borderline Band 6 students would get this item right, as compared to the Rasch estimate of 92%. If the Rasch estimate is used as the criterion, then the judges have over-estimated the difficulty of this item for borderline Band 6 students. The judges could then examine this item, looking at the distractor options to try to understand this difference.

On the other hand, on item B4 (out of 2 marks), the judges have under-estimated the difficulty of this item for Band 6 students, compared to the Rasch estimate. The judges have expected the borderline students to average 1.5 (out of 2), whereas the Rasch model indicates that such students would average about 1.2. The judges could examine the marking rubric to try to determine why such capable students did not perform quite as well as expected.

A table such as Table 3 fits naturally into an Angoff multi-stage procedure. It provides statistical feedback in Stage 2, allowing the judges to discuss their results and possibly modify some of their decisions, although there is no compulsion for them to change their ratings.

Distribution of Total Cutscores. In the previous method, the judges performed their Angoff ratings in Stage 1 and these were then summed and averaged over the six judges to determine a total cutscore and hence a borderline ability estimate. In this method, the judges perform their Angoff ratings in Stage 1, but *these are not summed*. Instead, each item rating itself is used to determine a total cutscore. When the judges estimate the probability of success on a dichotomous item or estimate the average score on a constructed-response item, they are, in effect, setting an ability level which can be converted to a total cutscore.

For example, for a dichotomous item, the ability estimate is related to a judge's probability estimates from Equation (8). Let $\hat{P}$ represent a judge's estimate and substitute it in (8) in place of the probability of success. This gives the equivalent ability estimate as follows:

$$\hat{\theta} = \delta_j - \ln\left(\frac{1-\hat{P}}{\hat{P}}\right). \qquad (13)$$

This ability is then converted to a total cutscore from the relationship shown in Figure 1.

For constructed-response items, a table is formed (as outlined previously) in which the *expected item score* has been calculated for every value of θ. A judge's estimate of the borderline item cutscore is then treated as though it were an expected item score and the table is used to estimate θ by linear interpolation. For example, consider a graphical representation of this process as shown in Figure 4 below.

Figure 4 shows the relationship between θ and expected item score for item B10. This item has a maximum possible value of 5. It can be seen that a judge giving an item cutscore of 4.0 would be indicating an equivalent ability of about 1.8.

In this example, the proficiency level being targeted is the highest level, Band 6. In Table 4 below, the mean of the judges' cutscores is given for every item, along with the equivalent ability estimate and the equivalent total cutscore. It can be seen that there is a wide range of estimated total cutscores, from 27.5 to 45.0. Some of these would be regarded as wildly off-target by the judges. Yet in the normal Angoff procedure, without statistical feedback, these would be treated identically to all the others, as part of a sum of all item cutscores.

A useful way to view the distribution of total cutscores is by boxplot, as shown in Figure 5. This shows a negatively skewed distribution with a long tail, culminating in the outrider, case 25, which corresponds to item B5. The median total cutscore is 40.1; the mean is 38.6, with the middle 50% of the cutscores lying between 35.6 and 41.3. The presentation of total cutscore equivalents as a boxplot gives a vivid way of displaying item estimates that seem to be anomalous. It also indicates that there is a distribution of total cutscore estimates, suggesting that there are alternative ways of arriving at a final estimate other than the usual Angoff procedure of summing the item cutscores. For example, one could simply take the median of this distribution as a suitable total cutscore. This median of 40.1 is reasonably consistent with the total cutscore obtained from the orthodox Angoff procedure, where each judge's ratings are summed and then averaged over the six judges.

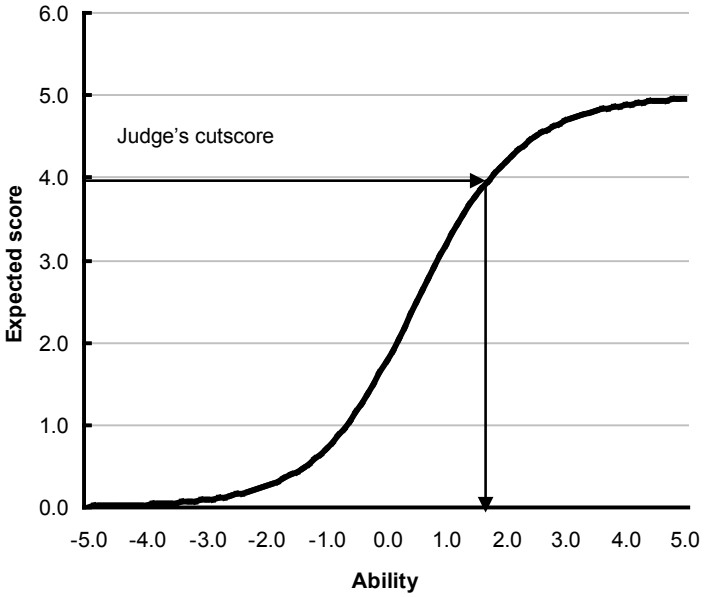

Figure 4. Estimating ability from item cutscore for item B10.

Table 4

Total cutscore equivalents for the judges' item ratings

Question	Angoff cutscore	Equivalent ability	Equivalent total cutscore
A1	0.90	1.825	41.3
A2	0.80	1.564	39.4
A3	0.91	1.253	36.7
A4	0.92	1.180	36.1
A5	0.88	1.005	34.4
A6	0.80	1.824	41.3
A7	0.79	1.754	40.8
A8	0.77	1.841	41.4
A9	0.81	1.549	39.3
A10	0.90	0.968	34.0
A11	0.83	0.922	33.5
A12	0.80	1.873	41.6
A13	0.80	1.683	40.3
A14	0.70	1.655	40.1
A15	0.72	1.661	40.1
A16	0.81	1.641	40.0
A17	0.73	1.488	38.8
A18	0.81	1.650	40.0
A19	0.88	0.871	33.0
A20	0.82	1.084	35.1
B1	1.57	1.701	40.4
B2	1.60	0.571	29.8
B3	1.58	1.682	40.3
B4	1.53	2.481	45.0
B5	1.67	0.359	27.5
B6	1.63	1.464	38.6
B7	3.20	1.972	42.3
B8	3.27	2.301	44.1
B9	4.00	1.741	40.7
B10	4.08	1.844	41.4

Under that method, the total cutscore would be 40.5, as shown in the previous section.

Judges Choose a Set of Expected Scores. In this method, the Rasch model is used to provide maximum information to the judges, so that the standard-setting process can be expedited. Corresponding to every total score, there is an ability estimate (see Figure 2). For each of these ability estimates, the expected score on each item can be calculated (see Equation (6)). This data can be given to the judges in a table. In such a table, the total score values would be the column headings and the rows would represent the items. The columns themselves would comprise the expected item scores, given the total score. The task of the judges would be *to select the column that best reflects their Angoff ratings over all items*. Once the column is chosen, the total cutscore is given by the score in the column heading.

For example, consider an Angoff procedure where the proficiency level being estimated is the highest level, Band 6. Then a table would be provided to the judges, showing a range of total score columns within which the cutscore would be expected to lie. An example of such a table is shown in Table 5, with the column that a judge has selected being enclosed in a box.

This table includes the total score at the top of each column and the equivalent percentile. It is probably best to suppress this information initially, as it may bias the judges' decisions (see for example, Reid, 1991). Before beginning the task, the judges would familiarize themselves with the proficiency standard through the Standards Package and refresh their knowledge of the test items. Each judge would then select the column that best reflects how the borderline students would be likely to perform across all items. In such a decision, the judges can consider item data simultaneously by ranging over a column and checking their subjective estimates against the values in the column. Undoubtedly there will be cases where the judges are surprised at the expected scores allocated to some items, relative to those of other items in the same column. However, all they are after is the best fit to their judgments.

This method would result in a considerable savings of time compared to the somewhat tedious item-by-item judgment of the normal Angoff method. In addition, it provides a way of

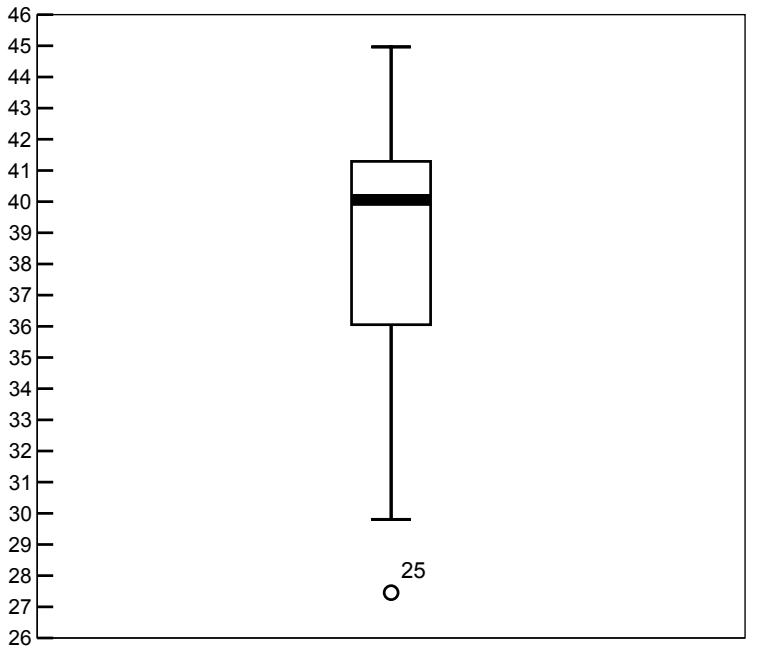

Figure 5. Total cutscore distribution from item cutscores.

removing the subjectivity of the Angoff ratings. The fallibility of such ratings has been remarked on in several studies (Bejar, 1983; Goodwin, 1999; Mills and Melican, 1988; Shepard, 1995). As these subjective Angoff ratings are replaced by Rasch expected scores, all the judgments are consistent, flowing from a clearly stated model. It has yet to be tested in practice whether such a process would result in greater consistency between judges than the usual Angoff method.

The procedure described above provides considerable direction and structure to the judges. However, the implementation of this method may be varied to suit the needs of the educational system. An alternative implementation would be to have the judges *supply independent Angoff ratings for a sample of the items before looking at Table 5*. The judges may be more confident about making Angoff ratings on some items than others, and may want to focus on the former items. Both multiple-choice and some constructed-response items could be included in this sample. Each judge would enter their initial item cutscores on a recording sheet before looking at the Table 5 data. They would then compare the ratings on the recording sheet with the expected scores in the table and *select the column that best matched their ratings*.

A third option would simply be to use Table 5 at Stage 2 in the normal Angoff method. In Stage 1, the judges would independently work their way through the test, estimating cutscores for all the items. In Stage 2, they would be given Table 5 and asked to discuss their ratings with the other judges in the light of the Table 5 data. As a result of this discussion they would be given an opportunity to revise their ratings. In this approach, the judges' ratings are regarded as the primary data and the Table 5 data as secondary. It would be up to the judges as to how much they revise their ratings as a result of studying Table 5.

Table 5

Expected item scores for each Total Score ability

Score:		38	39	40	41	42	43	44	45	46
Percentile:		69.0	73.0	76.7	80.5	84.2	87.8	90.9	93.8	96.2
Qst	Max									
A1	1	0.85	0.87	0.88	0.90	0.91	0.92	0.93	0.95	0.96
A2	1	0.77	0.79	0.81	0.83	0.85	0.87	0.89	0.91	0.93
A3	1	0.92	0.93	0.94	0.94	0.95	0.96	0.97	0.97	0.98
A4	1	0.93	0.94	0.95	0.95	0.96	0.96	0.97	0.98	0.98
A5	1	0.91	0.92	0.93	0.94	0.95	0.95	0.96	0.97	0.98
A6	1	0.72	0.75	0.77	0.79	0.82	0.84	0.86	0.89	0.91
A7	1	0.73	0.75	0.77	0.80	0.82	0.84	0.87	0.89	0.91
A8	1	0.68	0.70	0.73	0.76	0.78	0.81	0.84	0.86	0.89
A9	1	0.78	0.80	0.82	0.84	0.86	0.88	0.90	0.92	0.93
A10	1	0.93	0.94	0.95	0.95	0.96	0.97	0.97	0.98	0.98
A11	1	0.89	0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.97
A12	1	0.71	0.74	0.76	0.78	0.81	0.83	0.86	0.88	0.90
A13	1	0.75	0.77	0.79	0.82	0.84	0.86	0.88	0.90	0.92
A14	1	0.64	0.67	0.70	0.73	0.75	0.78	0.81	0.84	0.87
A15	1	0.66	0.69	0.71	0.74	0.77	0.80	0.82	0.85	0.88
A16	1	0.77	0.79	0.81	0.83	0.85	0.87	0.89	0.91	0.93
A17	1	0.71	0.73	0.76	0.78	0.80	0.83	0.85	0.88	0.90
A18	1	0.77	0.79	0.81	0.83	0.85	0.87	0.89	0.91	0.93
A19	1	0.92	0.93	0.94	0.95	0.95	0.96	0.97	0.97	0.98
A20	1	0.86	0.87	0.89	0.90	0.91	0.92	0.94	0.95	0.96
B1	2	1.43	1.49	1.54	1.60	1.65	1.70	1.75	1.80	1.84
B2	2	1.80	1.82	1.84	1.86	1.88	1.89	1.91	1.93	1.94
B3	2	1.47	1.52	1.57	1.62	1.67	1.72	1.76	1.81	1.85
B4	2	1.05	1.11	1.18	1.24	1.31	1.38	1.46	1.54	1.62
B5	2	1.87	1.88	1.89	1.91	1.92	1.93	1.94	1.95	1.96
B6	2	1.61	1.65	1.68	1.72	1.75	1.78	1.81	1.85	1.88
B7	4	2.74	2.85	2.96	3.06	3.17	3.28	3.39	3.49	3.60
B8	4	2.71	2.79	2.88	2.96	3.05	3.15	3.25	3.36	3.48
B9	5	3.74	3.84	3.93	4.03	4.13	4.23	4.33	4.44	4.55
B10	5	3.66	3.79	3.91	4.03	4.15	4.27	4.38	4.49	4.60

Discussion

This paper has put forward several ways in which Rasch modeling can be used in standard setting. It has long been noted that the most popular procedure, the Angoff method, shares a natural affinity with IRT procedures with both having a common view of a continuum of achievement and a probabilistic definition of performance on an item (Kane, 1987; van der Linden, 1982). The Rasch procedures given here can be used to support a normal Angoff multi-stage method, but can also be used to radically change the Angoff operating procedures. The support to the usual Angoff procedures would occur in Stage 2, where the judges discuss their Stage 1 decisions in the light of statistical data.

In the method entitled *Judge ratings versus Rasch estimates*, each judge's Stage 1 decisions are summed and the totals are averaged across the judges to get a single cutscore. This score defines a Rasch ability and this ability is used to estimate expected scores on the items. The judges can then see the extent to which their ratings differ from the Rasch expected scores and can change some of their ratings, if desired, following the group discussion.

In the *Distribution of Total Cutscores* method, a judge's rating for each item is regarded as a probability of success (or as an expected score for constructed-response items), and is used to calculate the equivalent person ability and hence the equivalent total score. A distribution of these total scores could be used as Stage 2 feedback to the judges to identify those item ratings that give wildly inappropriate total score estimates. The judges could then revise their ratings on these items. Alternatively, the distribution of total scores *itself* could be used to obtain an Angoff cutscore. Several possibilities are available, depending on the viewpoint of the educational system. One could test for and eliminate outriders and calculate the mean, or calculate a trimmed mean, or use some other statistic such as the median, to gain a cutscore estimate. These could then be averaged across the judges.

The method where *Judges choose a set of Expected Scores* has the potential to greatly streamline the judging process. In this approach, the Rasch model is used to calculate a set of expected item scores for every total-score ability point. The judges simply choose the column of expected scores that best reflects their item ratings. This could be implemented in several ways depending on the requirements of the educational system. If the judges are quite familiar with the items and are confident with their ratings, they could quickly "home in" on the appropriate column and confirm their ratings against the column values. Alternatively, the judges could select a sample of items that they were confident of judging and enter a Stage 1 set of independent ratings on a recording sheet, before looking at the table of expected values. By comparing their ratings with the expected values for the sample of items, they could then select the appropriate set of expected values. Provided an educational system has the time to prepare such a table of expected values, this method could greatly ease the cognitive load on the judges.

In comparison to the Angoff method, the Bookmark method was explicitly designed to make use of IRT theory. Under the Rasch model, the method becomes very simple if the items are dichotomously-scored. In this case, the bookmark difficulty location (*BDL*) is equal to the item difficulty plus a constant. Hence the *BDL* order and the item difficulty order *are the same* for the purpose of presenting items in the standard setting booklet. If a response probability of 0.5 is used, the situation is even simpler. In this case, the concept of a bookmark difficulty location (*BDL*) is not required, as the *BDL is exactly equal to the item difficulty*. Thus the process of arranging the items in the standard-setting booklet can be explained solely in terms of item difficulties, a great gain in simplicity and much easier for the judges to follow. In addition, for the item mapping presentation, this allows the easy representation of the items in histogram form, which conveniently fits the data on a single page. The judges merely have to place their bookmark between the col-

umns of the histogram, which are labeled with the item numbers.

If the Rasch model is not used in the dichotomous Bookmark procedure, the situation becomes more complex, as the other IRT methods include item discrimination in their models, resulting in different slopes for the item characteristic curves. These IRT models will not generally arrange the items in the same *BDL* order as the Rasch model and the item orders will generally differ from model to model. *Even within the same IRT model,* the model may arrange the items in different orders, depending on the response probability being considered. The choice of response probability, whether 0.80, 0.67, or 0.5, can affect the ordering of the items, due to its interaction with the different slopes of the item characteristic curves. The effect of this on the reporting of standards, where items are used as exemplars of the types of questions students can typically answer at a given proficiency level, seems to be problematic. If items are used as exemplars in this way, then the response probability should be published with the item, as a different response probability may result in a different set of items at that proficiency (also, see the discussion in Beretvas, 2004).

For constructed-response items, the same problems occur for non-Rasch models. However, these problems will now occur for Rasch models. Within an item, for each score level, there is a curve that shows the probability of gaining that score or above as a function of ability. These curves do not cross within the item. However, *between items*, there is the likelihood that such curves could have different slopes, resulting in the problems described above. More experience in analyzing constructed-response data is required to form a clear view of the extent of this occurrence. For the data set analyzed in this paper (comprising mixed item types), the Spearman's rho correlation between *BDL*s from an *RP* of 0.5, and from an *RP* of two-thirds, was 0.990. For a broader spectrum of data, it is uncertain how much the rank order would typically change over differing response probabilities, and whether any controllable characteristics of the constructed response items can be identified that are associated with the change in rank order. It is also uncertain what impact this effect would typically have on the final cutscore. For Rasch practitioners in standard setting, this is an area where more research is needed to understand the scope of these problems.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123-140.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika*, *47*, 105-113.

Andrich, D. (1988). A general form of Rasch's extended logistic model for partial credit scoring. *Applied Measurement in Education*, *1*, 363-378.

Andrich, D., Sheridan, B., Lyne, A., and Luo, G. (2000). *RUMM: A Windows-based item analysis employing Rasch unidimensional measurement models*. Perth, WA, Australia: Murdoch University.

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

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

Beretvas, S. N. (2004). Comparison of bookmark difficulty locations under different item response models. *Applied Psychological Measurement*, *28*, 25-47.

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

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

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.

Kane, M. T. (1987). On the use of IRT models with judgmental standard setting procedures. *Journal of Educational Measurement*, *24*, 333-345.

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

MacCann, R. G., and Stanley, G. (2006). The use of Rasch modeling to improve standard setting. *Practical Assessment Research & Evaluation*, *11*(2). Retrieved 20 November 2006, from: http://pareonline.net/getvn.asp?v=11&n=2

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Masters, G. N., and Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika*, *49*, 269-272.

Mills, C. N., and Melican, G. J. (1988). Estimating and adjusting cutoff scores: Future of selected methods. *Applied Measurement in Education*, *1*, 261-275.

Mitzel, H. C., Lewis, D. M., Patz, R. J., and Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum.

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

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

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.)

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, *19*, 49-57.

Reckase, M. D. (2000). Survey and evaluation of recently developed procedures for setting standards on educational tests. In *Student performance standards on the National Assessment of Educational Progress: Affirmation and improvements* (pp. 41-69). Washington, DC: National Assessment Governing Board.

Reid, J. B. (1991). Training judges to generate standard-setting data. *Educational Measurement: Issues and Practice*, *10*, 11-14.

Shepard, L. A. (1995). Implications for standard setting of the National Academy of Education evaluation of the National Assessment of Educational Progress achievement levels. *Proceedings of Joint Conference on Standard Setting for Large-scale Assessments* (pp 143-160). Washington, DC: National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES).

van der Linden, W. J. (1982). A latent trait method for determining intrajudge consistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement*, *4*, 295-308.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement*, *40*, 231-253.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*, 97-116.

Wright, B. D., and Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, *29*, 23-48.

Wright, B. D., and Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.