

Maintaining consistent metrics in standard setting

Sandra Heldsinger

and

Stephen Humphry

Murdoch University, Western Australia

Mailing address

Stephen Humphry
Murdoch University
Murdoch 6150
Western Australia

Acknowledgements

This paper is based on PhD research by Dr Stephen Humphry completed under the co-supervision of Profs Andrich and Luo, and doctoral work by Sandra Heldsinger being completed under the supervision of Prof Andrich, at Murdoch University in Western Australia. Permission by the Department of Education and Training of Western Australia to use the data sets is acknowledged and support for the research within the Department has greatly benefited the work. The work has been supported by an Australian Research Council grant with the Australian National Ministerial Council on Employment, Education, Training and Youth Affairs (MCEETYA) Performance Measurement and Reporting Task Force, UNESCO's International Institute for Educational Planning (IIEP), and the Australian Council for Educational Research (ACER) as Industry Partners*.

*Report No. 8 ARC Linkage Grant LP0454080: Maintaining Invariant Scales in State, National and International Level Assessments. D Andrich and G Luo Chief Investigators, Murdoch University

Maintaining consistent metrics in standard setting

Abstract

Several methodologies have been devised to set cut scores in education. In this study two different methods are employed to set cut scores, one involving the use of a modified Angoff procedure for rating item difficulties and the other a method of pairwise comparisons of item difficulties. Item scale locations are inferred from both methods and these are compared with item locations obtained from student responses. Comparisons between the three resultant metrics reveal substantial differences between the dispersions of item locations, which result in divergent cut scores when the origins, but not units, of the metrics are equated in order to translate the cut score onto the student metric. In order to obtain convergent cut scores using the two methodologies, transformations must be made in order to account for the differences between the units. The implications of the findings are discussed.

Keywords: standard setting, metric, unit, Angoff procedure, pairwise comparison, cut score, Rasch model, BTL model, 2PL model

Maintaining consistent metrics in standard setting

Introduction

Standard setting exercises with respect to a particular assessment are commonly used in testing programs where there is a requirement to determine a point at which students are deemed to have demonstrated achievement of a standard (Reckase, 2006). This study examines a standard setting exercise for a reading assessment at year 7 undertaken in the Australian state of Western Australia. The relevant standard is referred to as the *benchmark standard*. Because the benchmark standard has been defined, both in terms of criteria and in terms of supporting exemplar materia, the context of the study differs from many other standard setting exercises. In particular, the task was not to define the standard, but rather to locate the qualitatively defined standard on an existing reading scale.

Two different standard setting methodologies are employed to translate the defined standard onto a student scale. One involves the use of a rating scale and the other a method of pairwise comparisons. The method in which rating scales are used is referred to as the *likelihood methodology*, and this is essentially a modified Angoff (1971) standard setting procedure. Modified Angoff standard setting procedures require judges to rate the probabilities of success of students on items (Reckase, 2000, 2006) and these ratings are used to infer cut scores. The second procedure, referred to as the *pairwise methodology*, involves judges comparing items in pairs and judging which item is the more difficult. In this methodology, judges compare items on the student reading test with items specifically chosen to illustrate the benchmark standard. The latter items are then scaled together with the items on the reading test to infer a location for the standard on the resulting metric.

The measurement models used in this study to analyse student responses and the data collected in the likelihood and pairwise methodologies permit direct comparisons between the item scale locations. The main findings are that: (i) the units of the metrics obtained from student responses, the likelihood methodology, and the pairwise methodology are substantially different; (ii) without accounting for the differences between the metrics, the cut scores obtained using the pairwise and

likelihood methodologies are substantially divergent; and (iii) when the difference between the metrics is accounted for, the cut scores converge.

These findings carries important implications for valid use of standard setting methods, particularly the Angoff and modified Angoff procedures widely used in licensure and certification testing (Clauser et al, 2002). Reckase (2006, p. 6) observes that a potential source of error in Angoff procedures is that panellists “may not be able to specify the probability of a correct response to an item for a person at the intended cut score with the necessary level of accuracy”. Several authors have in the past observed that judges have difficulty accurately judging likelihoods of success on items (Impara & Plake, 1996, 1997; Lorge & Kruglov, 1953; Shepard, 1995). The finding reported in this study confirms these observations. The study clarifies the implications of the inaccuracy of judgements for cut scores set using Angoff procedures by clarifying the effect of differences between the units of metrics of likelihood and student metrics on resultant cut scores. In particular, the pattern of inaccuracy observed in the likelihood judgements results in systematic distortion of the cut score if the resultant differences between the units of metric are not accounted for in setting the cut score.

Following from this point, Green et al (2003) report studies in which convergence of results among multiple standard settings is used as evidence of validity of cut scores. They note that while convergence may occur to a reasonable degree when variations of the same methods are used, there are few reports of convergence when *different* procedures are used, as is the case in the present study. This criterion is consistent with general remarks by Bock & Jones (1968) regarding the use of independent methods of measurement to confirm the validity of each of the methods. The present study shows that differences between the units of different metrics largely account for the divergence of the cut scores obtained using the likelihood and pairwise methodologies. The implication is that differences between units should routinely be investigated and accounted for in standard setting methodologies.

The study also illustrates a more general point consistent with observations by Bock & Jones (1968) regarding experiment conditions for measurement; namely, empirical features of the format for collecting responses can have a pronounced impact on the

unit of a metric in any context for quantitative educational research. This broader point is also broached in the discussion.

Formats for data collection

The format for collection of student data was a typical one in which students responded to a number of dichotomous items on a reading test. The relevant test was part of the 2001 Western Australian Literacy and Numeracy Assessment (WALNA) program, administered to approximately 25,000 year 7 students in the state. The program includes the administration of reading, writing, mathematics, and spelling assessments in years 3, 5, and 7 by classroom teachers based on detailed administrative instructions.

The format for the collection of student response data is shown in Table 1. The items are denoted $i = 1, \dots, I$, the students $n = 1, \dots, N$ and the responses of students to items x_{ni} .

Table 1: Student data collection format

Student / Item	1, 2, 3, I
1	x_{11} x_{1I}
2		
:		
n	x_{n1} x_{nI}
:		
N	x_{N1} x_{NI}

The first methodology used for standard setting is the likelihood methodology. The likelihood methodology required expert judges to envisage a minimally competent year 7 student as described in the standard, and then estimate the likelihood of that student answering each item correctly. A minimally competent student is one who only just possesses the level of skills and understanding to demonstrate the qualitatively defined standard but is unlikely to have the capacity to perform tasks that are more demanding than those that are defined by the standard. To follow, such a student is referred to as a *benchmark student*.

Judges used the eleven-point scale similar to that proposed by Angoff (1971) to record their estimates of the likelihood of a minimally competent benchmark student correctly answering an item. The response format was presented as shown in Figure 1.

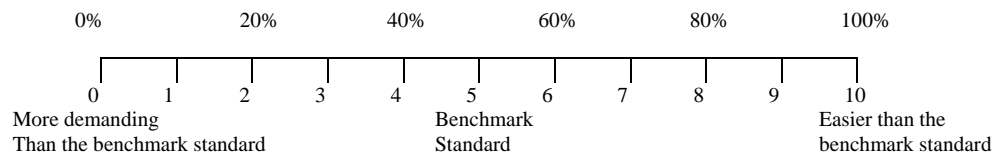


Figure 1: Likelihood rating scale

Judges were instructed that the minimally competent benchmark student should answer an item which is very close to benchmark standard correctly 50% of the time. If the skills needed to answer an item were more demanding than the benchmark standard, the chance that the benchmark student would to answer it correctly should be rated as less than 50%. Conversely, if the skills needed to answer an item were less demanding than the benchmark standard, the rating should be greater than 50%.

The format for data collection using the likelihood method is shown in Table 1. The responses are the rated likelihoods of success of a benchmark student on each item, which are integer values between 0 and 10. There were 25 judges, including 16 teachers and 11 non-teaching curriculum and assessment specialists.

Table 2: Likelihood data collection format

Judge / Item	1, 2, 3, <i>I</i>
1	$x_{11} \dots$	$\dots x_{1I}$
2		
:		
<i>g</i>	$x_{g1} \dots$	$\dots x_{gI}$
:		
<i>G</i>	$x_{G1} \dots$	$\dots x_{GI}$

The second standard setting methodology used in this study is based on the method of pairwise comparison; a method originally conceived and articulated by Thurstone (1927). Pairwise data were collected from the judges' comparisons of a series of pairs of items from an item bank. Judges recorded the item perceived as more difficult in

each pair. The item bank consisted of the reading test items as well as the reading items written specifically to exemplify the benchmark standard.

The format for pairwise data collection is shown in Table 3. A response $x_{ji} = 1$ indicates item j was judged more difficult than item i and a response $x_{ji} = 0$ indicates item j was judged more difficult than item i .

Table 3: Pairwise data collection format

Judge		Judge 1	Judge 2	Judge P
Item		1, 2, 3, i I
Judge 1	1	x_{11} x_{12} x_{13} x_{1I}
	2	x_{22} x_{23} x_{2I}
Judge 2	3	x_{33} x_{3I}
	:			
Judge P	:			
	J		 x_{JI}

There were 35 test items and 13 benchmark exemplar items. The number of possible comparisons if each item is compared with every other item is equal to the half the number of off-diagonal elements in Table 3. Here, this number is

$$\frac{I(I-1)}{2} = \frac{48(47)}{2} = 1128 \quad (1)$$

which is too large a number of comparisons for any one judge. A design was therefore constructed involving 27 judges, in which each judge made approximately 113 comparisons. The design was also constructed so that each comparison was made at least twice.

Models for data analysis

The student response data and likelihood data were both analysed using the Rasch model for measurement (Rasch, 1960, 1961). The Rasch model is generally stated as

$$\Pr\{X_{ni} = 1\} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)}, \quad (2)$$

where β_n is the location of person n and δ_i is the location of item i on a latent continuum.

In the case of the likelihood methodology, the parameter β_g takes the place of β_n in Equation (2). The parameter β_g represents the ability of a benchmark student perceived and internalised by judge g as described above. The likelihood ratings are treated as the estimates of probabilities of success of the benchmark student on each item. The item scale values are inferred from these probabilities using joint maximum likelihood (JML) estimations through application of customised software RUMMMM (Andrich & Luo, 1998).

Although the Rasch model is usually stated in the form shown in Equation (2), more generally the model can be stated as

$$\Pr\{X_{kni} = 1\} = \frac{\exp(\rho_k(\beta_n - \delta_i))}{1 + \exp(\rho_k(\beta_n - \delta_i))}, \quad (3)$$

where ρ_k is an arbitrary scale factor which must be of uniform magnitude within a *specified frame of reference* for measurement. The concept of a frame of reference was defined by Rasch (1977) and the general form of the model given in Equation (3) is referred to as the *frame of reference Rasch model* (FRM) (Humphry, 2006). Each of the formats for data collection shown in tables 1 to 3 represents a different frame of reference. In the case of the likelihood format for data collection, the scale parameter ρ_l takes the place of ρ_k in Equation (3) and β_g again takes the place of β_n .

The term ρ is a *scale factor* in the sense that it only enlarges or shrinks the person-item distance $\beta_n - \delta_i$ to $(\beta_n - \delta_i)/\rho$ (Andrich, 1995; Luo, 1998). This arbitrary scale factor was recognized by Rasch (1960, p. 121) when he identified the general class of models for measurement. Although the magnitude of ρ is arbitrary in the model, each set of data has its own natural unit of scale which is dependent on empirical features of the frame of reference in which data are collected. The same arbitrary scale factor $\rho \equiv 1$ is generally imposed in analysing data irrespective of empirical differences between the frames of reference. It is shown in this and the following section that this arbitrary constraint must be taken into account in order to maintain a consistent metric across the different formats for data collection.

The defining feature of Rasch models for measurement is that the parameters of the models have sufficient statistics. Sufficiency arises due to the fact that the person parameter can be eliminated in a comparison between two or more items as shown by Rasch (1960). Specifically, the probability of a correct response to one of two items conditional on a total score of 1 is

$$\Pr\{X_{knj} = 0, X_{kni} = 1 \mid r_n = 1\} = \frac{\exp(\rho_k(\delta_j - \delta_i))}{1 + \exp(\rho_k(\delta_j - \delta_i))}, \quad (4)$$

where $r_n = x_{ni} + x_{nj}$ is the total score of person n across items i and j .

Data collected in the pairwise format were analysed using the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952; Luce, 1959), which is usually stated in the following form:

$$\Pr\{X_{ji} = 1\} = \frac{\exp(\delta_j - \delta_i)}{1 + \exp(\delta_j - \delta_i)}. \quad (5)$$

In this context, Equation (5) constitutes a statement of the probability that item j is judged more difficult than item i in a pairwise comparison between the items. Andrich (1978) showed that (i) the person parameter is eliminated experimentally in

the pairwise design and also that (ii) when the person parameter of the Rasch model is eliminated, and the logistic response function is substituted for the cumulative normal, the case V specialisation of the Thurstone model is identical to conditional form of the Rasch model shown in Equation (5). The conditional form of the Rasch model has the same form as the BTL model. This equivalence enables a direct comparison between item estimates obtained from application of the BTL to the pairwise response data with item estimates obtained by applying the Rasch model to the student response data.

Because the effect of empirical features of the data format on the unit of the metric is explicitly considered, a scale parameter is also incorporated within the BTL model. The resulting form of the model is

$$\Pr\{X_{mji} = 1\} = \frac{\exp(\rho_m(\delta_j - \delta_i))}{1 + \exp(\rho_m(\delta_j - \delta_i))}. \quad (6)$$

Clearly, Equations (4) and (6) have an identical structure. The only difference between the equations is the subscripts of the scale parameter ρ , which recognise differences between empirical factors inherent to each format for data collection. Three subscripts k , l , and m are used, where k denotes that student responses were instrumental to comparisons between items, l denotes that likelihoods of success were judged and m denotes that judges were instrumental to comparisons in the pairwise format. By subscripting the parameter in this way, allowance is made for the effects of empirical factors on the units of the different metrics. The item parameters are identical in the three formats, which permits comparison between the estimates of the item locations under the formats.

The compatibility of the models means that three sets of item parameters are reduced to one set of item parameters and three scale parameters. That is, the number of parameters is reduced substantially and there are three separate *compatible* models used to estimate any given item parameter. Bock and Jones (1968, p. 9) made the following remarks in relation to the use of different methods to obtain independent measurements.

It is the mark of a maturity of a science that the number of parameters with which it deals is small and the number of models large. In such a science, measurement can reach a high level of perfection because, first, there is generally more than one distinct model which can be said to estimate a given parametric value. This provides the possibility for independent measurement by different methods, which serves to confirm the validity of each. Second, the science may include models which make it possible to predict the effect of altered experimental conditions on the measurement procedure.

Thus, in this study the item parameters are estimated by different methods and using separate models, and results are compared in order to confirm the validity of each method.

With respect to the models used in this study, it is stressed that in Equation (3) the scale parameter ρ_k is not treated in the same way as the item discrimination parameter of Birnbaum's (1968) *two parameter logistic model*. Instead, it is treated as a parameter which pertains to empirical factors in terms of which a frame of reference for measurement, as a whole, is defined. In the terms used by Bock & Jones, this scale parameter recognizes the effects of *altered experimental conditions* on the measurement of the common items. This point is elaborated in the discussion.

Comparison of the metrics obtained from the two formats

RUMM2020 (Andrich, Sheridan & Luo, 1997-2006) was used to implement scaling based upon application of the Rasch model to student responses, and RUMMcc (Andrich, Sheridan, & Luo, 2003) was used to implement scaling of pairwise data based on the BTL model. Customised software was used to implement scaling based on application of the Rasch model to the likelihood ratings using joint maximum likelihood estimation. RUMM2020 implements pairwise conditional maximum likelihood estimation, described in Andrich & Luo (2003). Although unconditional JML estimation is used for the likelihood data in this study rather than conditional estimation, in theoretical terms the item estimates are effectively equivalent given a correction factor of $(I-1)/I$ for bias in the item estimates (Wright & Douglas, 1977). This enables direct comparison between the item estimates for the purpose of

common item equating. The correction factor in the present case is 34/35, which is inconsequential to the findings and hence not used.

In the algorithms used in each piece of software, the scale parameter is arbitrarily defined as $\rho \equiv 1$ and hence differences between the relative magnitudes of the scale factor across the different formats are absorbed into the item estimates produced in the software. In order to recognise that the scale factor is absorbed into the scale locations given the algorithms, we define

$$\delta_{mi} \equiv \rho_m \delta_i \quad (7)$$

and

$$\delta_{ki} \equiv \rho_k \delta_i, \quad (8)$$

where δ_{ki} is the scale location of item i obtained using the response data of students and δ_{mi} is the scale location of item i obtained from pairwise comparisons by judges. In stating these definitions, it is assumed the locations are referenced to a common origin by constraining the sum of the locations of common items to be 0. Similarly, we define $\delta_{li} \equiv \rho_l \delta_i$ and $\beta_{lg} \equiv \rho_l \beta_g$.

It follows from the definitions in equations (7) and (8) that

$$\frac{\sqrt{V[\delta_m]}}{\sqrt{V[\delta_k]}} = \frac{\rho_m \sqrt{V[\delta]}}{\rho_k \sqrt{V[\delta]}} = \frac{\rho_m}{\rho_k}, \quad (9)$$

where $V[\delta]$ is the variance of the *common items*; that is, items that were administered to students, rated by the judges in the likelihood format, and compared by the judges in the pairwise format. The standard deviations of the item estimates of the 35 common items obtained from the three formats for data collection are shown in Table 4. It is evident that the standard deviations are very different from each other.

Table 4: Standard deviations of scale locations collected under each format

$\sqrt{V[\hat{\delta}_k]}$	$\sqrt{V[\hat{\delta}_l]}$	$\sqrt{V[\hat{\delta}_m]}$
1.16	0.50	2.51

The estimates of the ratios of the likelihood and pairwise scale parameters to the student scale parameter are shown in Table 5.

Table 5: Ratios of scale parameters

$\hat{\rho}_l / \hat{\rho}_k$	$\hat{\rho}_m / \hat{\rho}_k$
0.50	2.16

The first ratio shown in Table 5 reflects that the standard deviation of the item estimates of the common items obtained from the likelihood data is 0.50 times the standard deviation of the estimates of the same items obtained from the students' responses. The second ratio reflects that the standard deviation of the item estimates of the common items obtained from the pairwise judgements is 2.16 times the standard deviation of the estimates of the same items obtained from the students' responses.

As elaborated to follow, there is a systematic pattern of inaccuracy in the ratings in the likelihood format. This pattern affects the inferred item estimates and hence the ratio $\hat{\rho}_l / \hat{\rho}_k$ inferred from the differences between the dispersions of the item estimates. Consequently, the difference between the natural units of the likelihood and student metrics is likely to be at least partly an artefact of treating ratings as though they are literally probabilities of success, rather than it being a genuine difference in the level of precision as would be expected given a genuine difference between the natural units of the metrics arising due to empirical features of the formats. Nevertheless, the effect of likelihood judgements on the cut score is real irrespective of the source of the difference between the units. It is noted that the pairwise format does not involve ratings, and the differences between the natural units of the pairwise and student metrics is a genuine reflection of differences between the formats for data collection.

Given that both the likelihood and pairwise methodologies depend upon equating to the scale derived from student data, it is important to examine the relationships between the variables. The correlations between item estimates derived from the data obtained within the different formats are shown in Table 6. The correlation matrix shows that there was a very high correlation between the scale generated from the judges' likelihood estimates and the scale generated from the judges' pairwise comparison of items ($r = 0.95$). Given many judges were involved in both methodologies these results indicate the judges had a clear and consistent interpretation of relative item difficulties, across the two methodologies. The likelihood scale and the pairwise scale had similar correlations with the student achievement scale ($r = 0.80$ and 0.74 respectively), indicating that while judges had a consistent perception of relative item difficulties their perceptions diverge to a degree from the *actual relative* difficulties of the items for students.

Table 6: Correlation matrix of scale locations

	Student Data	Likelihood	Pairwise
Student Data	1.00	0.80	0.74
Likelihood	0.80	1.00	0.95
Pairwise	0.74	0.95	1.00

Locating the benchmark standard on the student scale

The benchmark locations were firstly estimated on the metrics derived from the likelihood and pairwise standard setting methodologies, referred to as the *likelihood metric* and *pairwise metric* respectively.

The benchmark exemplar item difficulties generated from the pairwise comparisons items ranged from -4.39 to $+0.55$. That is, on the pairwise metric, the easiest benchmark item location was -4.39 and the most difficult benchmark location was $+0.55$. In order to translate the benchmark definition to a single scale location, the average of all the benchmark exemplar item locations was computed. The resulting location was $\hat{\beta}_m = -1.55$ on the pairwise metric.

The location of the benchmark on the likelihood metric represents the mean of the judge's perceptions of the benchmark ability, as described above. The resulting benchmark location was $\bar{\hat{\beta}}_l$ on the likelihood metric.

The benchmark locations on these metrics were translated onto the student metric using common item equating in two forms: (i) equating only the means of the scale locations of common items; and (ii) equating both the means and standard deviations of the common items.

The benchmark locations on the student metric, equating only for the origin, are shown in Table 7.

Table 7: Benchmark locations and corresponding cut scores after equating for the origins of the metrics

Methodology	Scale location	Raw Score
Likelihood	-0.179	16.08
Pairwise	-1.810	7.09

The difference between the raw scores is 9.01 of a possible total of 35. Thus, the methodologies produce substantially different locations of the standard on the student metric when the origins, but not units, of the pairwise and likelihood metrics are equated with the origin of the student metric.

The benchmark locations on each of the metrics were subsequently also equated with the student metric for differences between the units and origins. Letting $\bar{\hat{\beta}}_l$ represent the estimated benchmark location on the likelihood metric, the location was transformed onto the student metric as follows:

$$\bar{\hat{\beta}}_k = (\bar{\hat{\beta}}_l - \bar{\hat{\delta}}_l) \left(\frac{\rho_k}{\rho_l} \right) + \bar{\hat{\delta}}_k ; \quad (10)$$

where $\bar{\hat{\beta}}_k$ is the estimated benchmark location on the student metric, $\bar{\hat{\delta}}_l$ is the mean item difficulty estimate on the likelihood metric, and $\bar{\hat{\delta}}_k$ is mean item difficulty

estimate on the student metric. Note that in the transformation made earlier for only the difference between the origins, the ratio ρ_k / ρ_l was defined to be 1, which implies the assumption that there is a common unit underlying the metrics.

Similarly, the estimate of the benchmark location from the pairwise metric, equating for differences between units and origins, was obtained using the following transformation:

$$\hat{\beta}_k = (\hat{\beta}_m - \bar{\hat{\delta}}_l) \left(\frac{\hat{\rho}_k}{\hat{\rho}_m} \right) + \bar{\hat{\delta}}_m ; \quad (11)$$

where $\hat{\beta}_m$ is the estimated benchmark location on the pairwise metric and $\bar{\hat{\delta}}_m$ is the mean item difficulty estimate on the likelihood metric.

Equations (10) and (11) have the same form as the metric transformation referred to by Baker (1983, p. 100). However, in the present context, $\hat{\rho}_k / \hat{\rho}_l$ is a ratio between scale parameters pertaining to a frame of reference as a whole, rather than a ratio between average item discriminations as was the case in the context referred to by Baker (1983).

The benchmark locations and corresponding raw scores obtained using the transformations shown in Equations (10) and (11) are shown in Table 8.

Table 8: Benchmark locations and corresponding cut scores after equating for the units and origins of the metrics

	Scale location	Raw Score
Likelihood	-0.46	14.64
Pairwise	-0.86	12.15

There is a substantial change in the locations of the cut scores shown in tables 7 and 8. Moreover, there is considerably closer correspondence between the estimated benchmark locations when transformations are made to equate for differences between both the units and origins of (i) the likelihood and student metrics and (ii) the

pairwise and student metrics. In terms of the raw cut score, the difference is reduced from 9.01 to 2.49. This is the main finding of the study.

Qualitative analysis was also employed to substantiate that the cut score is a reasonable representation of the defined benchmark standard. Exposition of the qualitative analysis is however beyond the scope of this paper.

Implications for standard setting procedures

As noted in the introduction, the main finding carries important implications for valid use of standard setting methods; particularly the widely used Angoff and modified Angoff procedures. Interrogation of the data revealed that the difference between the scale parameters of the student and likelihood metrics results from inaccurate judgements of the probabilities of success of a benchmark student on the items. This was confirmed by selecting approximately 700 students with raw scores on the reading test of 14 and 15 and comparing actual percentages correct with the mean likelihoods across judges using the rating scale. These raw scores are closest to the cut score of 14.64 shown in Table 8 above. The data revealed a systematic tendency for judges to overrate the likelihood of success for a benchmark student on relatively difficult items and underrate the likelihood of success for a benchmark student on relatively easy items. This observation is consistent with observations made by Lorge and Kruglov (1953). The systematic inaccuracy led to a distortion of the benchmark cut score when it was not accounted for. In general, compression of the likelihood metric relative to the student metric will result in a benchmark location that is compressed toward the mean item location. Because the benchmark standard is below the mean item location in the present study, the difference between the likelihood and student metrics led to inflation of the cut score when the effects were not taken into account. This inflation can be seen by comparing the likelihood cut score in Table 7 with that shown in Table 8.

Shepard (1995) reported on a series of exercises in which judges using the Angoff method incorrectly estimated item difficulty. She concluded that the Angoff method may not provide valid cut scores because judges were unable to perform the fundamental task required of them; i.e. to estimate the probability a student would

answer an item correctly. This study confirms Shepard's conclusion and shows the specific way in which the cut score is affected by inaccurate estimations of likelihoods when the units of the metrics are not accounted for when estimating the cut score.

Discussion

Differences between the units of different metrics carry important implications for quantitative research if they are not accounted for within the approach to measurement adopted by the researcher. The role of the unit in educational measurement is generally left implicit, and although a relationship between discrimination and the unit of a metric is often noted (eg. Brink, 1971; Wood, 1978; Andrich, 1988; Embretson & Reise, 2000), the relationship is not generally considered explicitly in a manner that permits investigations regarding the influence of empirical factors on the unit. In this paper, the influence of empirical factors has been explicitly parameterised in the models given in equations (4) and (6) to account for the effects on the units of the different metrics and the subsequent consequences for inferences regarding the location of the standard on the student metric.

In retrospect, it is not surprising that different formats for data collection produce different units of the different metrics, and that different cut scores result. It is likely that many anomalous results reported in the literature on standard setting results, including divergent results, arise from differences between the units of the metrics in the various response formats of judges relative to the format of data collection for students.

Before concluding, it is noted that the widespread use of Birnbaum's (1968) two parameter logistic model suggests a widespread perspective that scale parameters of item response models should be referenced only to individual items, and therefore only to item characteristics. Indeed, this is likely to be a significant reason that differences between units of metrics associated with different formats for data collection are not generally recognised and taken into account – the items are generally common across formats and the structure of the two parameter logistic model implies it is unnecessary to equate for differences between units when there are common items that are assumed to have common discriminations. The effects of

empirical factors other than item characteristics on the unit have, however, been observed in this study. The broader implication of the findings reported in this paper is that in general, the scale parameter should be conceptualised in broader terms as pertaining to *any empirical factor that has material relevance to the responses collected within a frame of reference*. This is consistent with the observations of Bock & Jones (1968) cited above regarding the effects of *altered experimental conditions* on the measurement procedure. The broader conceptualisation of the scale parameter is described more fully by Humphry (2005, 2006).

Conclusion

In practice in the Australian context, as in various other contexts, the benchmark location is translated from the likelihood to the student metric by equating the mean of common items. This practice inherently assumes the units of the metrics are the same. However, in this study the standard deviation of the item estimates on the student scale locations was approximately *twice* that of the likelihood scale locations, but approximately *half* that of the pairwise item locations. The substantial difference between the standard deviations reflects a difference between the units of the metrics. This difference means the units and origins of the likelihood and pairwise metrics must be equated to the student metric in order to establish the cut score on the student metric. Accordingly, the likelihood and pairwise metrics were transformed so that both origins and units of the metrics were equivalent to that of the student metric.

The transformation resulted in cut scores that were substantially closer to one another than when only the origins of the metrics were equated. It was also substantiated qualitatively that the resulting cut score is a reasonable representation of the defined benchmark standard. The findings have important implications for valid use of standard setting procedures and, accordingly, for obtaining convergent results using different procedures that involve substantive empirical differences between formats for data collection. The implication of the findings reported in this paper is that differences between the units of metrics associated with different formats for data collection should routinely be investigated and accounted for in the process of standard setting.

References

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, 2, 449-460.

Andrich, D. (1988). *Rasch models for measurement*. Beverly Hills: Sage Publications.

Andrich, D. (1995). Hyperbolic cosine latent trait models for unfolding direct-responses and pairwise preferences. *Applied Psychological Measurement*, 19(3), 269-290.

Andrich, D., Sheridan, B. & Luo, G. (1997-2006). *RUMM2020*. RUMM Laboratory, Perth, Australia.

Andrich, D., Sheridan, B. & Luo, G. (2003). *RUMMcc*. RUMM Laboratory, Perth, Australia.

Andrich, D. & Luo, G. (1998). *RUMMmm*. RUMM Laboratory, Perth, Australia.

Andrich, D. & Luo, G. (2003): Conditional Pairwise estimation in the Rasch model for ordered response categories using principle components. *Journal of Applied Measurement*, 4(3), 205-221

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.

Baker, F. (1983). Comparison of ability metrics obtained under two latent trait theory procedures. *Applied Psychological Measurement*, 7(1), 97-110.

Bock, R.D. & Jones, L.V. (1968). *The measurement and prediction of judgement and choice*. San Francisco: Holden-Day.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. & Novick, M.R. (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bradley, R.A. and Terry, M.E. (1952): *Rank analysis of incomplete block designs, I. the method of paired comparisons*. Biometrika, 39, 324-345.

Brink, N.E. (1971). Effect of item discrimination in the Rasch model. *Proceedings of the Annual Convention of the American Psychological Association*, 6(1), pp. 101-102.

Clauser, B.E., Swanson, D.B. & Harik, P. (2002). Multivariate generalizability analysis of the impact of training and examinee performance information on judgments made in an Angoff-style standard setting procedure. *Journal of Educational Measurement*, 39(4), 26-9290.

Embretson, S. and Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Green, D.R., Trimble, S.C., & Lewis, D.M. (2003). Interpreting the results of three different standard setting procedures. *Educational Measurement: Issues and Practice*, 22, 22-32.

Humphry, S.M. (2005). *Maintaining a common arbitrary unit in social measurement*. PhD Thesis <http://wwwlib.murdoch.edu.au/adt/browse/view/adt-MU20050830.95143>

Humphry, S.M. (2006). Maintaining a common unit in social measurement. Manuscript submitted to *Applied Psychological Measurement* in June 2006.

Impara, J.C. & Plake, B.S. (1996). *Teachers ability to estimate item difficulty: A test of assumptions in the Angoff standard setting method*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Impara, J.C. & Plake, B.S. (1997). Standard setting: An Alternative Approach. *Journal of Educational Measurement*, 34, 353-366.

Lorge, I. & Kruglov, L.K. (1953). *The improvement of the estimates of test difficulty. Educational and Psychological Measurement*, 13, 34-36.

Luce, R.D. (1959): *Individual Choice Behaviours: A Theoretical Analysis*. New York: J. Wiley.

Luo, G. (1998). A general formulation for unidimensional unfolding and pairwise preference models: Making explicit the latitude of acceptance. *Journal of Mathematical Psychology*, 42, 400 – 417.

Rasch, G. (1977). On Specific Objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *The Danish Yearbook of Philosophy*, 14, 58-93.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology, pp. 321-334 in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV*. Berkeley: University of Chicago Press, 1980.

Reckase, M.D. (2000). *The ACT/NAGB standard setting process: how “modified” does it have to be before it is no longer a modified-Angoff process?* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, April 24 to 28, 2000.

Reckase, M.D. (2006). A conceptual framework for psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice*, 25(2), 4-18.

Shepard, L.A. (1995). Implications for standard setting of the NAE evaluation of the NAEP achievement levels. In *Proceedings of the Joint Conference on Standard Setting for Large Scale Assessments* (pp. 143-1600). Washington DC: National Assessment Governing Board and National Center for Educational Statistics.

Thurstone, L.L. (1927). A law of comparative judgement. *Psychological Review*, 34, 278-286.

Wood, R. (1978). Fitting the Rasch model – A heady tale. *British Journal of Mathematical and Statistical Psychology*, 31, 27-32.

Wright, B.D. and Douglas, G.A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement*, 1, 281-294.