Article

Applied Psychological Measurement
2014, Vol. 38(6) 432–449
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621614529646
apm.sagepub.com
SAGE

Person Proficiency Estimates in the Dichotomous Rasch Model When Random Guessing Is Removed From Difficulty Estimates of Multiple Choice Items

David Andrich[1] and Ida Marais[1]

Abstract

Andrich, Marais, and Humphry showed formally that Waller's procedure that removes responses to multiple choice (MC) items that are likely to be guessed eliminates the bias in the Rasch model (RM) estimates of difficult items and makes them more difficult. The former did not study any consequences on the person proficiency estimates. This article shows that when the procedure is applied, the more proficient persons who are least likely to guess benefit by a greater amount than the less proficient, who are most likely to guess. This surprising result is explained by appreciating that the more proficient persons answer difficult items correctly at a greater rate than do the less proficient, even when the latter guess some items correctly. As a consequence, increasing the difficulty of the difficult items benefits them more than the less proficient persons. Analyses of a simulated and real example are shown illustratively. To not disadvantage the more proficient persons, it is suggested that Waller's procedure be used when the RM is used to analyze MC items.

Keywords

Rasch model, multiple choice tests, guessing, Raven's progressive matrices

Many educational and psychological assessments consist of multiple choice (MC) items, where there is a possibility that correct answers will be guessed. Depending on how difficult a person finds an item, if the person does not know the correct answer then he or she may eliminate some distractors and guess randomly among the rest, or if not able to eliminate any distractors, guess randomly among all of them. In this article, partial or complete random guessing will be referred to as random guessing. The previously implied proposition, and a theme of the article, is that a person will tend to guess a response as a function of how difficult the person finds an item, and not because of structural properties of the item. Thus, it is not assumed that guessing is present

[1]The University of Western Australia, Perth, Australia

Corresponding Author:
David Andrich, Chapple Professor, Graduate School of Education, The University of Western Australia, M428, 35 Stirling Highway, Crawley, Perth, Western Australia 6009, Australia.
Email: david.andrich@uwa.edu.au

or not, but that it is a matter of degree, which is a function of a person's proficiency relative to an item's difficulty.

Because it has a guessing parameter for each item, responses to MC items are analyzed, often using the three-parameter logistic (3PL), which includes an item guessing parameter. However, because of the proposition that the degree of guessing is a function of the relative difficulty of an item for a person, this article studies its consequences by analyzing responses using the dichotomous Rasch model (RM), which has only an item difficulty parameter. There are other reasons, compatible with the earlier proposition, for analyzing responses to MC items using the dichotomous RM. The most important is Rasch's (1961) theory of invariant comparisons—that if the data fit the dichotomous RM within a specified frame of reference, then the item parameter estimates are independent of any assumptions about the person distribution, and the person estimates are independent of which subset of items is used for assessment (Andrich, 2004). This feature is relevant in adaptive or tailored testing, where the focus is on the assessment of individuals and where, from a specified class of items, different persons are administered different items which are aligned relatively closely to each person's proficiency. In addition, if tailored testing is applied, it is expected that little or no random guessing will take place, and it seems at best redundant to have an item characterized partly by a guessing parameter. Finally, the model has the convenience that the person's total score on a set of items is sufficient for the person parameter estimate.

Random guessing violates the RM and will therefore bias the estimates of its difficulty parameters. Specifically, because the sufficient statistic for an item's difficulty is the number of correct responses, which guessing increases, the more difficult items will appear relatively easier than they would be without guessing. This violation of the RM does not preclude using it as a hypothesis that there is no guessing in the responses, seeking evidence to the contrary, and then considering how to deal with any guessing found. This, rather than focusing on modeling the data using the 3PL or a more complicated model (e.g., San Martin, del Pino, & De Boeck, 2006), is the approach taken in this study (Andrich, 2004).

Only a few studies have investigated the effect of guessing on item and person estimates in the RM. Using simulated data, and as might be expected, Waller (1973, 1976, 1989) showed that items that had correctly guessed responses appeared relatively easier than their simulated values and that persons with low proficiency, who were hypothesized to be more likely to guess, were estimated to be more proficient than their simulated value. Waller proposed removing responses likely to be guessed from the data and showed that the resulting analysis recovered simulated person and item locations more accurately. Waller's approach removes the response of a person to an item after it is established that the person has a low proficiency relative to the item's difficulty and, therefore, that the response is likely to have a guessing component. In addition to Waller's studies, it seems that only Wainer and Wright (1980) investigated the effect of guessing on the estimates of the dichotomous RM. They proposed an ''adjustment'' for guessing using the RM and a jackknife scheme, in which each item was omitted separately and proficiencies reestimated.

Recently, Andrich, Marais, and Humphry (2012) formalized Waller's approach, which is simpler than that of Wainer and Wright's, which permits estimating the magnitude and significance of the effect of guessing on the item difficulty estimates in the RM. Andrich et al. (2012) did not consider the consequent effects on the person estimates. The purpose of this study is to formalize the effects on *person* proficiency estimates in the RM when the effect of guessing is removed from the item parameter estimates using the procedure summarized above. The main new result, and perhaps surprising in the first instance because it is the least proficient persons who guess, is that although estimates of the least proficient persons are increased, the estimates of the most proficient persons are increased by an even greater margin. The reason for this effect is that the most proficient persons answer more difficult items correctly at a greater rate than the least proficient, even when the latter guess answers correctly, and therefore are affected more by changes in the locations of the difficult items than the less proficient. This effect is one of the main illustrations of the study.

The rest of the article is structured as follows: The section titled ''The Data Sets'' describes two data sets, one simulated and one real, which are analyzed for illustrative purposes. The ''Removing Random Guessing From the Item Parameter Estimates'' section summarizes the procedure used to correct the item parameter estimates. The section ''Proficiency Estimates'' describes the consequent effect on the person parameter estimates and the final section is a summary and discussion. Appendices A and B include elaborations of the article.

The Data Sets

Two data sets used illustratively in Andrich et al. (2012) are used in this study. They are a set of real data from the Raven's Advanced Progressive Matrices (RAPM) and a set of data simulated to parallel the real data. The RAPM is a non-verbal test of reasoning consisting of 36 MC items (Raven, 1940). Each item has a two dimensional matrix with a pattern in which the element in the lower right hand corner is missing. From six or eight alternatives, the respondent is required to choose one to complete the pattern. This advanced version is more difficult than Raven's Standard Progressive Matrices, the dimensionality of which was investigated by Van der Ven and Ellis (2000) using the dichotomous RM. The study by Andrich et al. (2012) of a sample of 469 persons with complete data on the RAPM showed that the items were relatively difficult for them and that there was substantial guessing in the most difficult items. Item 36 was so difficult that it was removed from all analyses.

The simulated data set had parameters similar to those from the RAPM data. The details of the item parameter estimates for the RAPM and the simulated data are shown for completeness in Table A1. In summary, the items were uniformly distributed in the range −3.0 to 3.0 logits with a mean of 0, and 470 persons were normally distributed with a mean of −0.75 and a standard deviation (*SD*) of 1.2 logits. By analogy to evidence from the RAPM, 23 of the most difficult items, in parallel to the RAPM data, were simulated to have guessing.

Because it appeared that the guessing was less prevalent by moderately proficient persons to moderately difficult items than implied by the 3PL, Andrich et al. (2012) used a simulation algorithm, a generalization of the 3PL, which permits modifying the degree of guessing as a function of the difference between a person's proficiency and an item's difficulty. This generalization is now summarized. The 3PL takes the form

$$\Pr\{X_{ni}=1\}=c_i+(1-c_i)P_{ni}=P_{ni}+c_i(1-P_{ni}), \tag{1}$$

where

$$P_{ni}=[\exp(\alpha_i(\beta_n-\delta_i))]/[1+\exp(\alpha_i(\beta_n-\delta_i)], \tag{2}$$

is the probability of a correct response without guessing, β_n is the proficiency of person n, and δ_i, α_i are respectively the difficulty and discrimination parameters of item i (Birnbaum, 1968). In the case $P_{ni}=0$, it is evident that the parameter c_i is the probability of a correct random guess. Because it has a discrimination parameter as well as a difficulty parameter, but no c_i parameter, Equation 2 is referred to as the two-parameter logistic (2PL) model. Algebraically, the dichotomous RM specializes further by setting $\alpha_i=1, \ \forall i$, giving

$$\Pr\{X_{ni}=1\}=[\exp(\beta_n-\delta_i)]/[1+\exp(\beta_n-\delta_i)]. \tag{3}$$

The generalization of the 3PL used in Andrich et al. (2012) takes the form

$$\Pr\{X_{ni}=1\}=P_{ni}+c_i(1-P_{ni})^y, \tag{4}$$

where y is any real number. In Equation 4, the probability P_{ni} of a correct response with no guessing is qualified by the term $c_i(1-P_{ni})^y$, where in the 3PL, $y=1$. Because $0<(1-P_{ni})<1$, the greater the value of y, the smaller the value of $(1-P_{ni})^y$. Therefore, for a given P_{ni}, which is a function of an item's difficulty relative to the person's proficiency, the effect of guessing is reduced. This interpretation is interpreted graphically in Figure A1. To mirror the RAPM data and for illustrative purposes, Andrich et al. (2012) chose $y=15$. Furthermore, they used the dichotomous RM (Equation 3) for P_{ni}. They also specified $c_i=1/7$ as a summary value for all items where in the real data some items had six and some had eight alternatives.

Removing Random Guessing From the Item Parameter Estimates

The procedure in Andrich et al. (2012) for removing the effects of random guessing is summarized below in part for completeness and in part to understand the reasons for the observed effects on the person proficiency estimates. All analyses used the RUMM2030 software (Andrich, Sheridan, & Luo, 2013), which provides consistent estimates (Zwinderman, 1995) using a pairwise conditional method of estimation of the item parameters (Andrich & Luo, 2003). Then taking the item parameters as known, the person parameters are estimated using a weighted likelihood method (Warm, 1989), which reduces the stretching bias of persons with scores toward the extremes found in maximum likelihood estimates.

The Tailored Analysis and Choice of Probability Cutoff

To operationalize Waller's procedure for removing random guessing, Andrich et al. (2012) used the following steps. First, all responses (which may have included guessed responses) were analyzed with the RM. This analysis was termed the *original analysis*. To the degree that guessing is present in the responses, to that degree there the responses will misfit the RM and the difficulty estimates will be biased. However, they will be biased in a predictable way. It is this reasoning that leads to the subsequent steps that identify the likely presence of guessing in MC items using the dichotomous RM. We broach fit again with the analysis of the illustrative examples.

Second, given estimates from the original analysis, if a person's correct response has a smaller probability than a specified value, whether the response was correct or not, it is converted to missing data. The effect is analogous to adaptive or tailored testing whereby students are not administered items that they are expected to find very difficult. In this case, the tailoring is carried out post hoc, rather than a priori. Accordingly, the analysis is referred to as a *tailored analysis* and the data as *tailored data*. The hypothesis of guessing is tested by comparing the difficulty estimates from the two analyses, which would be statistically insignificant if there was no guessing and the data fitted the dichotomous RM well. It is stressed that the processes does *not identify persons* who may or may not have guessed any response, rather the process identifies those *responses* likely to have a guessing component. It is also stressed that there is no way of telling whether or not any response was guessed. For example, a poorly proficient person may use specialized knowledge to answer an item correctly, but the response is still converted to be missing.

Specifically, Andrich et al. (2012) set a cutoff, which converted any response with a probability less than .3 of being correct to missing. This conservative value, relative to a theoretical value of $c_i = 1/7 = 0.143$, was chosen because if there is guessing in the original analysis, a person of low proficiency has an inflated estimate while a difficult item has a deflated estimate, making the probability of a correct response greater than with no guessing.

The Origin-Equated Analysis

In a RM analysis, an arbitrary identifying constraint, usually $\sum_{i=1}^{I} \hat{\delta}_i = 0$, on the item estimates is required. Other constraints can be imposed, for example, the difficulty of one item, or the mean difficulty of a subset of items, can be fixed. Because every analysis has such a constraint, the estimates from the original and tailored analyses for the items cannot be compared directly. The reason is that if difficult items appear even more difficult in a tailored analysis, then because of the constraint $\sum_{i=1}^{I} \hat{\delta}_i = 0$, the easy items will appear relatively easier. To set the same identifying constraint, and effectively the same origin, Andrich et al. (2012) reanalyzed the original data but anchored the mean of a subset of six easy items to their mean in the tailored analysis. The rationale, borne out in the examples, was that because the responses to the easiest items will have little if any guessing, a tailored analysis will make little change to the responses to these items, and therefore, little change to their *relative* difficulty estimates. The choice of six items was based on evidence from the data. Thus, if the items are relatively easier for the persons than those in the RAPM example, the mean of more easy items in the tailored analysis might have been chosen to equate the origin.

This provided a third analysis referred to in this article as the *origin-equated* analysis. Evidence of guessing then is that the estimates of the more difficult items will be more difficult in the tailored analysis than in the origin-equated analysis (complete data), while the very easy items will have similar difficulty estimates. Andrich et al. (2012) showed that the magnitude and significance of the difference in difficulty estimates could be assessed using a theorem by Andersen (1995, 2002). They also showed that the tailored analysis of the simulated data set, being minimally affected by guessing, provides unbiased difficulty estimates.

Summary of the Difficulty Estimates

Appendix A shows details of the item parameter estimates from the different analyses. However, for the purposes of this study, which focuses on the person estimates, it is sufficient to demonstrate the relationships graphically. Figure 1 (top) shows, for the simulated example, the estimated item locations from the origin-equated and the tailored analyses against the simulated item locations. The difficult items are clearly more difficult in the tailored analysis than in the origin-equated analysis, whereas the very easy items are equally difficult. The difficulties in the tailored analysis are closer to the identity line with simulated values. Because the more difficult items have more responses eliminated, their standard errors of estimates are greater, and this is reflected by the greater variation from the identity line of the more difficult items. Figure 1 (bottom) shows the estimated item locations from the origin-equated and tailored analyses for the RAPM. The difficult items are more difficult in the tailored analysis than in the origin-equated analysis, strongly indicating the presence of random guessing in the more difficult items. Furthermore, because the differences in difficulty estimates increases as a function of difficulty, it confirms the hypothesis that guessing is a matter of degree as a function of relative difficulty, and not simply present or not.

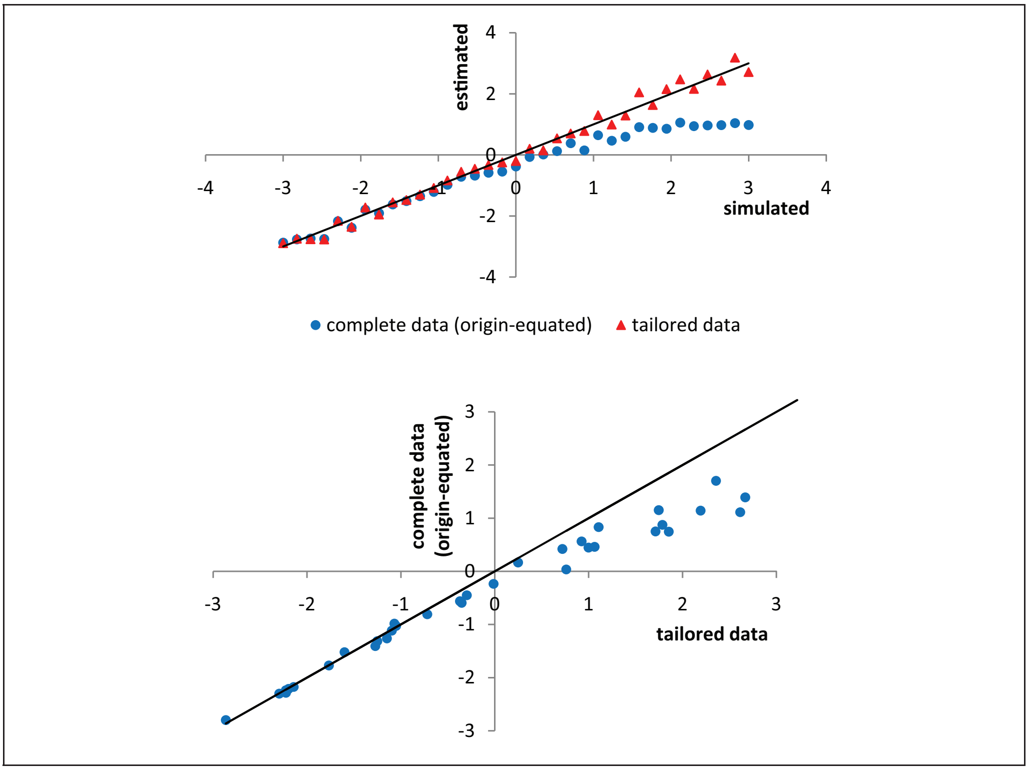

Figure 1. Item estimates from the tailored and origin-equated analyses against their simulated values (top), and item estimates from against origin-equated aginst the tailored analysis (bottom) for the RAPM, both showing the hypothesized identity line.

Note. RAPM = Raven's Advanced Progressive Matrices.

Proficiency Estimates

From Figure 1, the difficulty estimates in the tailored analysis are taken to be less biased than those of the origin-equated analysis. Accordingly, it is expected that the proficiencies estimated from the tailored analysis will also be less biased. However, when proficiency estimates of individual performances are reported, often for policy reasons, no responses of persons can be removed. Therefore, a fourth analysis was conducted. In this analysis, all of the original responses of all persons are used to estimate the person proficiencies, but the difficulties of all items are anchored to their estimates from the tailored analysis. This analysis is referred to as the *all-anchored* analysis. Therefore, in the origin-equated and all-anchored analyses, the complete set of responses is analyzed while in the tailored analysis a subset of responses is analyzed. Because meaningful comparisons require the same origin, we compare the person distributions of the tailored, origin-equated, and all-anchored analyses. In the latter two analyses, we take advantage of the sufficiency of the total score for person estimates, which, therefore, when all persons have responded to the same items can be compared.

Estimates From Total Scores to Proficiency Estimates

Because the simulated example is used to confirm the accuracy of the proposed procedure for correcting difficulty estimates for guessing, the consequent effect on the proficiency estimates

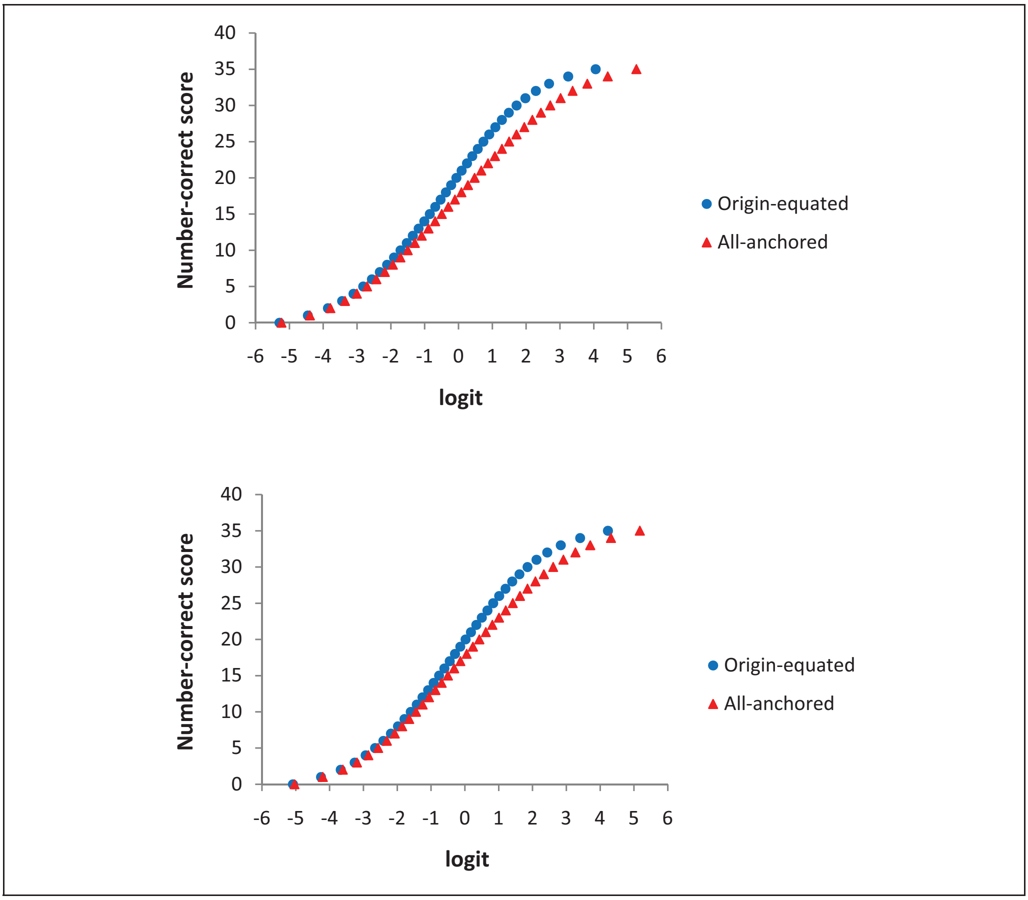

Figure 2. Number-correct score to logit conversion for the simulated example (top) and the RAPM (bottom) for the origin-equated and all-anchored analyses.
Note. RAPM = Raven's Advanced Progressive Matrices.

are summarized first. Figure 2 shows the estimated proficiencies for each total score from the origin-equated and the all-anchored analyses. It is evident that at the least proficient end of the continuum, there is little difference between the two, but that as proficiency increases, the all-anchored estimates for each total score show a systematically increasing difference. This is a consequence of the nonlinear way in which the difficulty estimates of items increase in the all-anchored, relative to the origin-anchored, analysis. The proficiency estimates of the RAPM show the same relationship.

Summary of the Proficiency Distribution

Table 1 shows the proficiency means and standard deviations from the three analyses described above for both the simulated and the RAPM data. In addition, for the simulated example, these values for the generating parameters and the actual simulated values, which are slightly different, are provided. Finally, for the simulated example, the regression of residuals (differences between simulated values and the estimated values) on the simulated values, are shown.

Table 1. Proficiency Estimates From Different Analyses and Regression of Residuals on the Generating Parameters for the Simulated Example.

			Estimates in analyses		
	Generating[a] parameters	Simulated values	Tailored (tailored data)	Origin-equated (complete data)	All anchored (complete data)
Simulation					
Person *M*	−0.750	−0.790	−0.780	−0.957	−0.584
Person *SD*	1.200	1.290	1.401	1.055	1.251
(*SE M*)			(0.060)	(0.060)	(0.060)
Sample size	470	470	466	470	470
Slope	0.000		−0.011	0.254	0.113
(*SE* slope)			(0.021)	(0.016)	(0.028)
Intercept	0.000		−0.017	0.367	−0.118
(*SE* intercept)			(0.031)	(0.024)	(0.018)
RAPM					
Person *M*			−0.742	−0.906	−0.632
Person *SD*			1.163	0.998	1.122
Sample size			462	469	469

Note. RAPM = Raven's Advanced Progressive Matrices.
[a]Person distribution is normal.

Because no bias in the person estimates implies that this regression line should have a zero intercept and a zero slope, it was calculated as an indicator of bias in the estimates. The mean square difference between the simulated value and the estimate could also have been calculated as an index of bias, but because the tailored analysis reduces the number of responses for the less proficient persons, it increases the standard error of their estimates. As a result, the bias and reduced precision are confounded in the mean square index.

Tailored Analysis

First, Table 1 shows that the regression of residuals in the simulated example from the tailored analysis are as expected (intercept −.017, slope −.011). Taking plus or minus two standard errors as a confidence interval around the value of 0.0, neither is statistically significant. Second, given the known standard deviation of the simulated values (1.290), the standard deviation of the mean is given by $1.290/\sqrt{470} = 0.060$. Again, taking two standard errors as the confidence interval around the simulated mean −0.790, it is evident that −0.780 falls well within this range. The standard deviation (1.401) is slightly greater than the simulated value (1.290), but this can be accounted for by the estimated values having an error component, which, when added to the simulated value, increases the standard deviation. Subtracting the average error variance available from the weighted likelihood estimates, the estimated true standard deviation is 1.279, which is even closer to 1.290. The conclusion is that the tailored difficulty estimates are unbiased and relatively precise and that they are the most correct estimates to define the scale.

The tailored analysis, therefore, also provides a frame of reference for the analyses of the RAPM. In the tailored analysis, seven persons had no response to any item, indicating that they were very poorly proficient relative to the difficulty of the test. (One person had no correct answer to any item even before the tailored analysis.) Of the remaining 462 persons, their mean (−0.742) is substantially less than the constrained mean item difficulty of 0.0, and it is this relative difficulty that is considered to have engendered guessing. The total number of responses deleted in the tailored analysis was 7,144 from a possible 16,380 responses, excluding the one person who had no correct answer in the original data. All items had at least one response removed, with six items having only 16 removed. These were the easiest items used to set the origin. The most difficult item had 466 responses removed. For completeness, Figure A2 shows that the persons are aligned to the easier end of the continuum. Clearly, with a lower cutoff for the probability of a successful response, say, .2, the number of responses deleted would have been less. However, as indicated already, that might have left more potentially guessed responses in the data.

Origin-Equated Analysis

In contrast to the tailored analysis, in the origin-equated analysis the intercept and slope of the residuals regressed on the simulated proficiencies are significantly different from zero. In addition, the mean proficiency is significantly less than the mean of both the simulating parameters and of the tailored analysis. This is initially counterintuitive. It would be expected that with complete data and therefore the presence of guessing, the mean would be greater in the origin-equated analysis. It is not, however, because the difficulties of a substantial number of items in the origin-equated analysis are regressed to lesser difficulties and the reward for a correct response to these items is also regressed. For the same reason, the standard deviation of the proficiencies is also smaller than that of the simulating parameters and of the tailored analysis. In the RAPM, the pattern is the same, with the mean and standard deviation of the origin-equated analysis both smaller than that in the tailored analysis.

All-Anchored Analysis

As indicated earlier, in any individual reporting of proficiency, all responses of a person are generally used. In this case, the relevant analysis is the all-anchored analysis, in which the items are anchored to the tailored item estimates, but where all responses of each person are included. Because of the presence of guessing in the data, the slope and intercept of the regressed residuals for the simulated example in Table 1 are significantly different from zero and, therefore, biased. As a result, the proficiency estimates are also biased. However, the degree of bias is substantially smaller than in the origin-equated analysis, 0.113 compared with 0.254 for the slope, and −0.118 compared with 0.367 for the intercept.

The person mean in the all-anchored analysis (−0.584) is greater than both the simulated mean (−0.790) and the mean in the origin-equated analysis (−0.957). The former is greater because it includes guessed responses, and the latter is greater because the item difficulties are greater than in the origin-equated analysis. The standard deviation in the all-anchored analysis (1.251) is slightly smaller than the simulated value (1.290), but is noticeably greater than the origin-equated analysis (1.055). These results point to the all-anchored analysis being preferred over the origin-equated analysis when all responses are analyzed. Incidentally, because the total score in the RM is the sufficient statistic, persons who have completed the same items and have the same total score have the same proficiency estimates, irrespective of the pattern of responses, although if the responses fit the model, they will be probabilistically Guttman-like (Andrich, 1985). This is not the case with the 3PL, where each pattern of responses generates a different estimate, and where, as shown by Chiu and Camilli (2012), a correct response for a person of low proficiency does not obtain the same credit as a correct response for a person of high proficiency.

As in the simulated example, the mean of the all-anchored analysis in the RAPM is greater than both the tailored and origin-equated analyses, where in the latter, again counter-intuitively the mean is less than in the tailored analysis. The standard deviation of the all-anchored analysis is also greater than that of the origin-equated analysis, and closer to that of the tailored analysis.

Thus, the first demonstration of the article is that (a) if guessing is removed from responses, unbiased estimates of the item difficulties are obtained and (b) if all responses are retained for the proficiency estimates of persons in the RM, then the mean and the standard deviations of the proficiencies are greater relative to those when guessing is not removed. However, as shown in the following, this effect does not arise from a uniform change across the continuum.

Cumulative Frequencies and Cut Points

In many large-scale educational assessments, cut points are set to indicate a minimum benchmark for achievement, and if this benchmark is not reached, remedial action is indicated. To not focus solely on reaching a minimum achievement standard, higher cut points that recognize meeting excellent benchmarks are also set. Because the RAPM and the simulated examples are relatively difficult, suppose for illustrative purposes that the lower and upper cut points are set at −2.0 and 0.0 logits, respectively. Figure 3 shows the cumulative percentage of persons in the simulated and RAPM examples. Consistent with Figure 2, the percentage of persons below any proficiency estimate in the origin-equated analysis is either equal to or greater than that in the all-anchored analysis, and the percentage difference increases as the proficiency level increases.

Applying the cut points graphically, the percentage of persons below the lower benchmark for the origin-equated and all-anchored analyses are both approximately 13, while the percentages below the upper benchmark are 89 and 72, respectively. Therefore, the percentages *above* the upper benchmark are 11 and 28, respectively, which is a substantial underestimate in the origin-equated analysis. Applying the same relative cut points, the same interpretation can be given for the cumulative percentages in the RAPM shown in Figure 3 (bottom), which are close to those of the simulated example. This closeness is not surprising, given that the simulated example was based on the parameters of the tailored analysis of the real data, and the degree of guessing was also simulated to be similar.

Thus, the second main demonstration of the article is that the greater mean in the all-anchored analysis than the origin-equated analysis does not result from a uniform shift in the proficiency estimates; the estimates of the most proficient are increased by a greater amount than of the least proficient. This nonlinear effect is demonstrated in the cumulative frequencies for both the simulated and RAPM examples in Figure 3. The implication is that if the effects of random guessing are not removed in estimating the item difficulties, then the relative achievements of the more proficient persons are under recognized in the RM.

Fit of Responses to the Model

As already noted, guessing violates the dichotomous RM. Consequently, the fit in the tailored analysis in which guessing is removed is hypothesized to be better than in the origin-equated analysis. For an illustrative general test of fit for this study, persons were divided into eight class intervals, and for each class interval an approximate χ^2 statistic based on the differences between the observed and expected frequency was calculated for each item, and then pooled over items. Table 2 shows the values and the probabilities for these statistics. As expected, for the simulated data, the tailored analysis (p = .227) shows satisfactory fit, while the origin-equated analysis fit (p = .000) is relatively unsatisfactory. These statistics show the same pattern

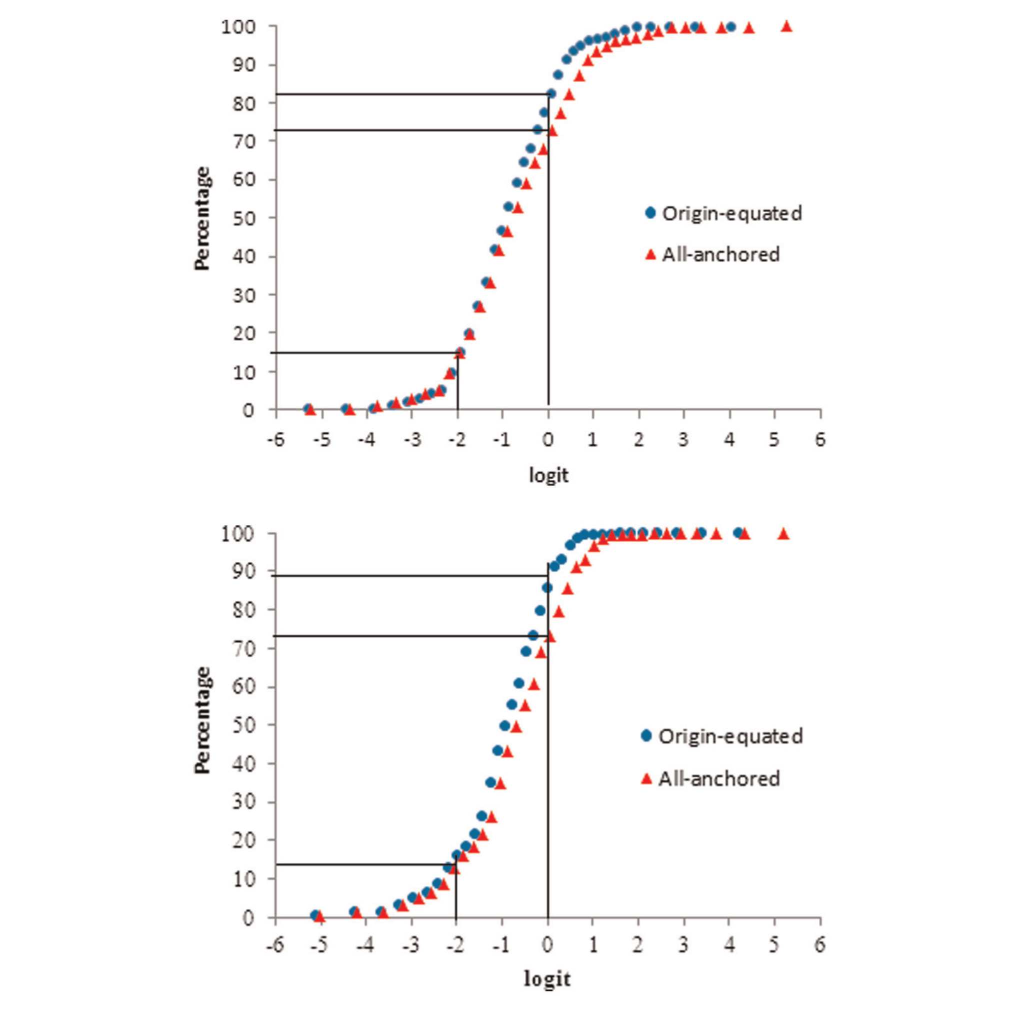

Figure 3. Cumulative percentage of person estimates for the simulated example (top) and the RAPM (bottom) for the origin-equated and all-anchored analyses.
Note. RAPM = Raven's Advanced Progressive Matrices.

in the RAPM: The tailored analysis shows relative fit ($\chi^2 = 0.049$), while the origin-equated analysis shows relative misfit ($p = .000$).

Number of Alternatives

The RAPM data were chosen to build on the analyses of the same data as in Andrich et al. (2012), in which the method of removing the effects of guessing from difficulty estimates in the RM was first formalized. These data have noticeable guessing, despite a larger than usual number of distractors, and so are ideal to illustrate a novel method for dealing with guessing. However, many traditional achievement tests have only four alternatives in their MC items and are designed not to be very difficult for persons to whom they are administered. For illustration

Table 2. General Approximate χ^2 Fit Statistics of Four Analyses for Simulated and RAPM Data.

	Original	Tailored	Origin-equated	All-anchored
Simulation	596.432 (*df* = 245, *p* = .000)	251.938 (*df* = 236, *p* = .227)	596.432 (*df* = 245, *p* = .000)	1851.548 (*df* = 245, *p* = .000)
RAPM	481.837 (*df* = 245, *p* = .000)	261.175 (*df* = 225, *p* = .049)	481.838 (*df* = 245, *p* = .000)	1080.343 (*df* = 245, *p* = .000)

Note. RAPM = Raven's Advanced Progressive Matrices.

of the process of removing guessing in these circumstances, Appendix B summarizes such an example and shows the graph of the distribution of the persons relative to the items and the key graph comparing item difficulties for identifying guessing. The same patterns consistent with the hypothesis that guessing is a function of the relative difficulty of an item for the proficiency of a person, though with an expected smaller effect because the items are relatively easy for the persons, are demonstrated clearly. Of course, whether or not there are effects due to guessing is an empirical matter in any data set, and if the items are very easy for the persons, it is expected that there will be relatively little guessing.

Summary and Discussion

The aim of this study was to determine the effects of removing guessing in the manner reported in Andrich et al. (2012) on *person* estimates in the dichotomous RM. The first conclusion is that in parallel with the effect on item difficulties, if the effect of guessing is not removed and the correct origin for comparison applied, then counterintuitively, the mean of the estimates is less than their actual mean. This result was not observed by Waller, who concluded that the mean was greater when guessing was not removed. However, he compared only the original analysis with the tailored analysis and did not equate the origin to be identical in the two analyses. In addition to the mean being less than the actual mean, the proficiencies are regressed toward their mean. As with the difficulty estimates, the article demonstrates that the regression effect on the proficiency estimates is not uniform across the continuum. In particular, the estimates of higher proficiency persons are regressed more when the effect of guessing is *not* removed from the difficulty estimates than those of lower proficiency persons, making their proficiency estimates less than their actual proficiencies.

This non-uniform effect can be explained by the observation that although the low proficiency persons tend to guess on the very difficult items, the more proficient persons answer these more difficult items at a greater rate than the less proficient persons, even though it is the latter who guess more and, therefore obtain a greater benefit from having difficult items with their real difficulties rather than regressed difficulties. Thus, not removing guessing when MC items are analyzed with the dichotomous RM will disadvantage the achievement of the more proficient persons. This may have substantial policy implications in large-scale national and international assessments.

Finally, the study confirms the rationale that in at least two real data sets (the RAPM and the example in Appendix B), guessing can be interpreted as a function of the difficulty of each item relative to the proficiency of each person, rather than it being a general property of an item or a complicated function of proficiency.

Appendix A

Further Comments on Guessing as a Function of Relative Person and Item Locations

This Appendix provides additional comments on the theme that guessing is a function of relative proficiency and difficulty, and some additional information on the simulated and Raven's Advanced Progressive Matrices (RAPM) examples. A corollary of the proposition that the degree of guessing by a person is a function of the relative difficulty of an item for that person is that if an item produces guessing from relatively proficient persons who have a high probability of answering the item correctly, then the item itself needs to be modified to eliminate such guessing. The idea that an item has structural flaws of a kind that induces guessing even from very proficient persons seems untenable as an inherent property of an item.

Although not routinely noted, the hypothesis that the degree of guessing in a response increases as the difficulty of the item increases, relative to the proficiency of a person, is consistent with the formulation of the three-parameter logistic (3PL). The 3PL has a lower asymptote, in which guessing increases at the lower end of the proficiency continuum. Nevertheless, it has a parameter as a structural guessing property of an item. Figure A1 shows the item characteristic curves (ICCs) for (a) the dichotomous Rasch model (RM) with which the responses are analyzed, (b) the 3PL, and (c) the particular generalization of the 3PL that was used to simulate the data in the article according to Equation 4. It is also evident that the effect of guessing is reduced as a function of proficiency for both cases, but is more pronounced at higher levels of proficiency in the 3PL than with its generalization of Equation 4 with $y = 15$.

San Martin, del Pino, and De Boeck (2006) provide a more complicated analysis of the nature of guessing. They formalize a model in which a person's proficiency affects guessing by a property of an item and by a general discrimination on the person proficiency parameter. The emphasis in their work, as in the 3PL, is on modeling the given data, which include correctly guessed responses, by proposing a model even more complicated than the 3PL. In the present article, the model is fixed to the simplest of models, the dichotomous RM, and the focus is on editing the data in such a way that it is unlikely to include correctly guessed responses.

Item difficulty estimates from different analyses. Table A1 summarizes the item parameter estimates for the original, tailored, and origin-equated analyses of the simulated and RAPM data sets. In the tailored analysis of the simulated data, four persons had all their responses converted to missing data and had no items to which they had a response, leaving 466 for the analysis. In the RAPM, seven persons had all their responses converted to missing data, leaving 462 for analysis.

In the simulation example, the mean of the item difficulties was 0.0, and the mean of the estimates in the tailored analysis is constrained to 0.0. From Table A1, first it is evident that the mean of the six easiest items in the tailored analysis (−2.613) is very close to the mean of their generating values (−2.559). This confirms that the mean of these items can be used to define the constraint of the origin in other analyses. Second, the mean difficulty estimates in the origin-equated analysis (–2.613), which includes the presence of random guessing, is 0.515 logits less than in the tailored analysis, in which the presence of random guessing is removed. In addition, it shows that the standard deviation of the former (1.321) is less than that of the latter (1.843). This is consistent with the rationale that the relative difficulties of the easy items should be similar in the two analyses, but that the estimates of the more difficult items will become even more difficult in the tailored analysis, thus producing a greater

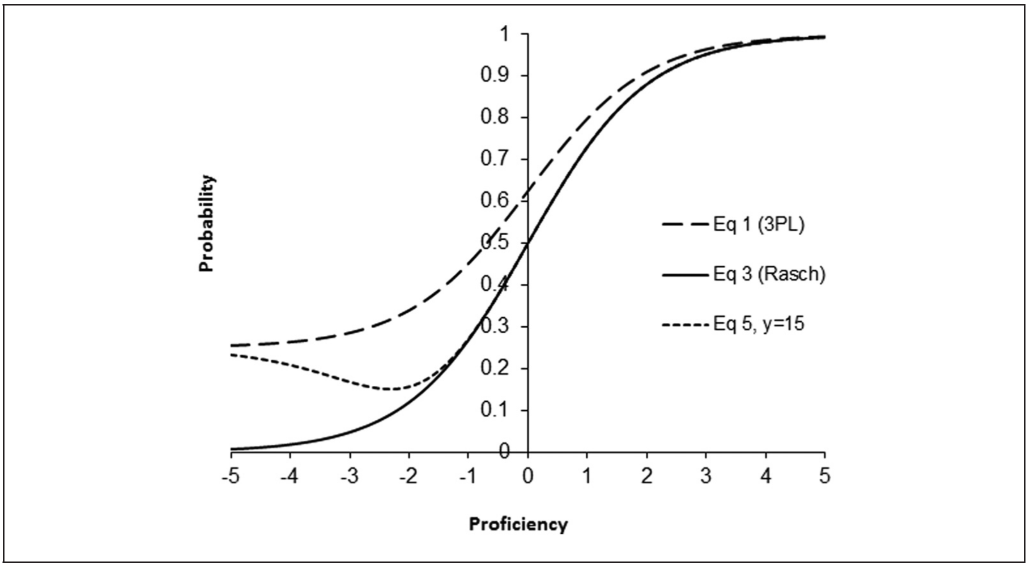

Figure A1. Item characteristic curves for the 3PL, the dichotomous RM, and the generalized guessing model with $y = 15$.
Note. 3PL = three-parameter logistic; RM = Rasch model.

spread of item difficulties. Finally, it is also evident that the standard deviation of the estimates in the origin-equated analysis, which is the same as in the original analysis (1.321), is smaller than the simulated value (1.808), reflecting regression to the mean, and that the standard deviation in the tailored analysis (1.843) is slightly greater than the simulated value. Because, unlike the generating values, the estimates incorporate error, this slightly greater standard deviation from the tailored analysis is expected. Thus, the effects observed in the estimates are directly a result of the particular misfit between the responses and the model, a misfit whose source in the simulated data is known.

In the RAPM, the difficulty estimates show identical relationships. Thus, as in the simulated example, in the RAPM, the mean item difficulty of the estimates (−0.392) is smaller in the origin-equated analysis than in the tailored analysis (0.000), and the standard deviation in the former (1.302) is also smaller than in the latter (1.715). Figure A2 shows the person and item distributions on the same scale confirming that the items are relatively difficult for the group of persons, thus engendering guessing by the least proficient.

Accounting for random guessing using the 3PL. Because it includes a parameter that can be interpreted as accounting for guessing, data analysts may apply the 3PL model to multiple choice (MC) items. One point made in the article is that such a parameter seems redundant in adaptive or tailored testing where difficulties of the items are aligned to the proficiencies of the persons, which minimizes the effect of random guessing. A second concern with the 3PL is the unstable estimation of the guessing parameter, which therefore requires large samples, especially including persons of low proficiency. However, because of misalignment, this simultaneously reduces the efficiency and precision of estimates of these less proficient persons (Han, 2012; San Martin et al., 2006). It is evident that in the method for dealing with guessing described in Andrich, Marais, and Humphry (2012) and elaborated in the present study, that it is not necessary to have extremely large samples.

Table A1. Item Means and Standard Deviations for the Three Analyses in Both the Simulated and RAPM Examples.

		Estimates in analyses		
	Simulation parameters	Original (complete data)	Tailored (tailored data)	Origin-equated (complete data)
Simulation				
M 6 easiest	−2.559	−2.099	−2.613	−2.613
M All	0.000	0.000	0.000	−0.515
SD All	1.808	1.321	1.843	1.321
Sample size	470	470	466	470
RAPM				
M 6 easiest		−1.857	−2.249	−2.249
M All		0.000	0.000	−0.392
SD All		1.302	1.715	1.302
Sample size		469	462	469

Note. RAPM = Raven's Advanced Progressive Matrices.

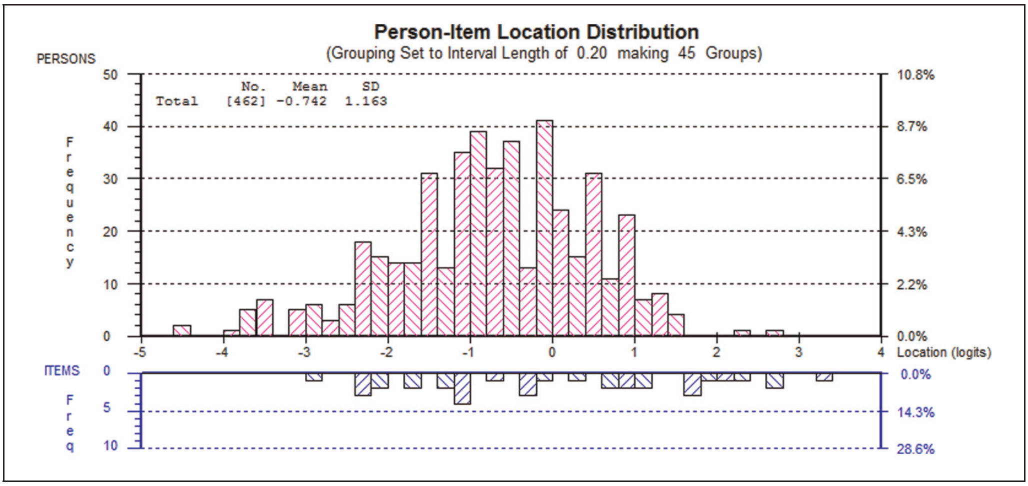

Figure A2. RAPM: Distribution of person and item estimates from the tailored analysis.
Note. RAPM = Raven's Advanced Progressive Matrices.

Another potential issue with using the 3PL is the result reported recently by Chiu and Camilli (2012) that a correct response by a less proficient person receives less credit than it does from a more proficient person. This can be understood by the ICC curve of the 3PL in Figure A1, which has greater guessing implied at the less proficient end of the continuum. However, it is not the statistics that may be an issue, but the perceived equity. It would seem difficult to justify that a correct response to an item by a less proficient person is of lesser value than that of a more proficient person. With the dichotomous RM in which, for the same set of items, all persons with the same total score obtain the same estimate, there is no such perceived issue of equity.

Appendix B

A Standard Example With Four Alternatives

Most educational assessments have MC items with four alternatives. An example is the Grade 3 Reading Assessment from the 2009 National Assessment Program–Literacy and Numeracy (NAPLAN), a high-profile, high-stakes, and large-scale assessment program in Australia. It consisted of 35 items of which 33 were of standard MC format with four possible answers. Figure B1 shows the dichotomous RM item and person estimates from an original analysis of the calibration sample data with 9,352 students. In contrast to the person-item alignment of the RAPM data shown in Figure A2, Figure B1 shows that the items were not very difficult for most students. The mean of the student proficiencies was positive, 0.867, relative to the mean difficulty of the items of 0.0. Nevertheless, there are persons whose proficiency estimate is of the order of -1.0 logits or less and, because their probability of a correct response is less than .20, who would most likely guess to items whose difficulty is of the order of .5 or greater.

Data were analyzed in a similar way as described in this study. Responses were tailored with a cutoff of 0.3. With four alternatives, the probability of a totally randomly guessed response is .25. Therefore, 0.30 is a conservative choice, though not as conservative as for the RAPM data, where there were six or eight options. Figure 1 shows the plot of the MC item estimates from the *origin-equated analysis* against estimates from the *tailored analysis*. Even though the items were not very difficult for the majority of persons and a relatively less conservative cutoff was chosen for converting responses to missing data than with the RAPM, Figure B2 shows that although easy items were estimated to be the same in both analyses, the more difficult items were estimated to be even more difficult in the *tailored analysis*. As expected from the hypothesis that the degree of guessing is a function of how difficult a person finds an item, the effect does not appear as great as with either the simulated or the RAPM data in the body of the article; nevertheless, it is clear. Moreover, the example further confirms the general hypothesis that the degree of random guessing is a function of the relative difficulty of each item to the proficiency of each person, rather than a general property of an item or even a property of a person. The point of this demonstration is not to give a best set of estimates, which would require

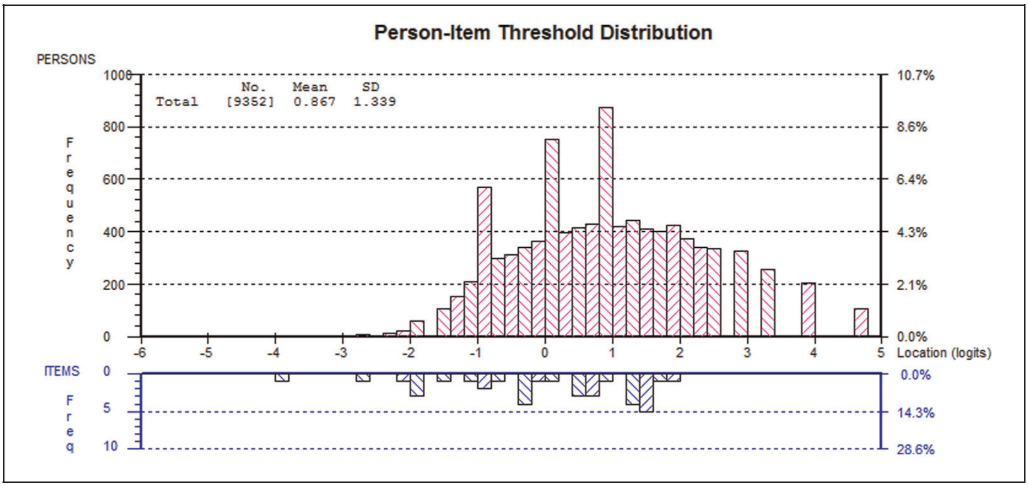

Figure B1. NAPLAN Grade 3 reading assessment 2009—calibration sample: Rasch item and person estimates from the original analysis.

Note. NAPLAN = National Assessment Program–Literacy and Numeracy.

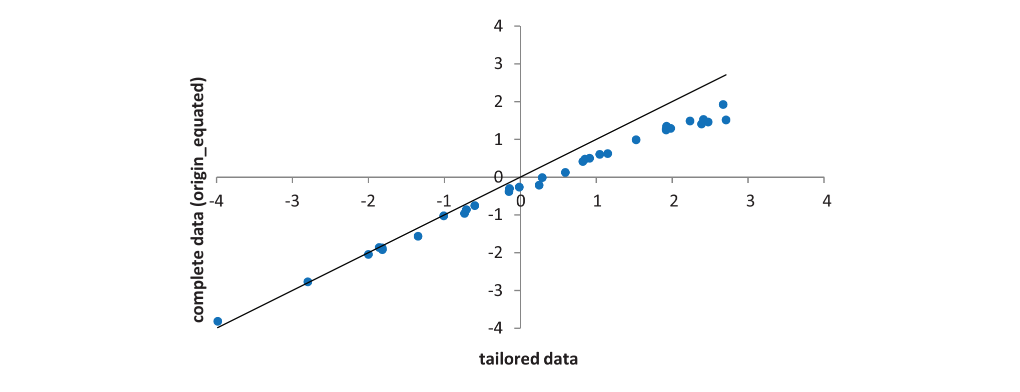

Figure B2. NAPLAN Grade 3 reading assessment 2009—calibration sample: Plot of the MC item estimates from the *origin-equated analysis* against those from the *tailored* analysis.
Note. NAPLAN = National Assessment Program–Literacy and Numeracy; MC = multiple choice.

further interactive analyses, including studying of differential item functioning and so on, but to illustrate that the hypothesis that guessing is a function of how difficult a person finds an item is confirmed and the consequences of dealing with it with the dichotomous RM, which has only difficulty and proficiency parameters. It further implies that there may be data sets which, because of the relative proficiencies of persons, have no guessing.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported in this article was supported, in part, by an Australian Research Council Linkage grant with the School Curriculum and Standards Authority of Western Australia and the Australian Curriculum Assessment and Reporting Authority as Industry Partners, and by Pearson plc.

References

Andersen, E. B. (1995). Residual analysis in the polytomous Rasch model. *Psychometrika*, *60*, 375-393.

Andersen, E. B. (2002). Residual diagrams based on a remarkably simple result concerning the variances of maximum likelihood estimators. *Journal of Educational and Behavioral Statistics*, *27*, 19-30.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Brandon-Tuma (Ed.), *Sociological methodology* (pp. 33-80). San Francisco, CA: Jossey-Bass.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, *42*, 7-16. (Reprinted in E. V. Smith, & R. M. Smith *Introduction to Rasch measurement: Theory, models and application*, pp. 143-166, Minnesota: JAM Press)

Andrich, D., Marais, I., & Humphry, S. M. (2012). Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. *Journal of Educational and Behavioral Statistics*, *37*, 417-442.

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

Andrich, D., Sheridan, B. E., & Luo, G. (2013). *RUMM2030: Rasch unidimensional models for measurement*. Perth, Western Australia, Australia: RUMM Laboratory.

Birnbaum, A. S. M. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-424). Reading, MA: Addison-Wesley.

Chiu, T., & Camilli, G. (2012). Comment on 3PL IRT adjustment for guessing. *Applied Psychological Measurement*, *37*, 76-86.

Han, K. T. (2012). Fixing the c parameter in the three-parameter logistic model. *Practical Assessment, Research & Evaluation*, *17*, 1-24.

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

Raven, J. C. (1940). Matrix tests. *Mental Health*, *1*, 10-18.

San Martin, E., del Pino, G., & De Boeck, P. (2006). IRT-models for ability-based guessing. *Applied Psychological Measurement*, *30*, 183-203.

Van der Ven, A. H. G. S., & Ellis, J. L. (2000). A Rasch analysis of Raven's standard progressive matrices. *Personality and Individual Differences*, *29*, 45-64.

Wainer, H., & Wright, B. D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika*, *45*, 373-391.

Waller, M. I. (1973). *Removing the effects of random guessing from latent trait ability estimates* (Unpublished doctoral dissertation). The University of Chicago, Chicago, IL.

Waller, M. I. (1976). *Estimating parameters in the Rasch model: Removing the effects of random guessing* (Research Bulletin). Princeton, NJ: Educational Testing Service.

Waller, M. I. (1989). Modeling guessing behavior: A comparison of two IRT models. *Applied Psychological Measurement*, *13*, 233-242.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427-450.

Zwinderman, A. H. (1995). Pairwise parameter estimation in Rasch models. *Applied Psychological Measurement*, *19*, 369-375.