SUMMER 2021, VOL. 34 NO. 2; ISSN 1051-0796

RMT

RASCH MEASUREMENT TRANSACTIONS

Transactions of the Rasch Measurement SIG
American Educational Research Association

Overview of The Issue

In this issue of RMT, we have included one research note and several announcements that may be of interest to the Rasch community.

The issue begins with a research note from David Andrich and Sonia Sappl on the stability of person estimates under the Rasch model.

Following the research note are two announcements related *Journal of Applied Measurement* (JAM). The first is an update from Hak Ping Tam and Richard Smith related to editorial and publisher changes for the *Journal of Applied Measurement* (JAM). The second is a call for papers from George Engelhard, who will be co-editing an upcoming special issue in JAM on unfolding models along with Ye Yuan and Jue Wang.

The issue concludes with an announcement about an upcoming virtual conference on classroom assessment that will be hosted by the National Council on Measurement in Education (NCME).

As always, we welcome your contributions to the next issue for RMT. We would appreciate receiving your research note, conference or workshop announcement, etc. by October 1, 2021. Please contact us at the email addresses below if you wish to submit something for inclusion.

Sincerely,

Your RMT Co-editors, Leigh and Stefanie

Rasch Measurement Transactions
www.rasch.org/rmt
Copyright © 2021 Rasch Measurement SIG, AERA
Permission to copy is granted.

Editors: Leigh M. Harrell-Williams
& Stefanie A. Wind
Email submissions to:
Leigh.Williams@memphis.edu or swind@ua.edu
RMT Editors Emeritus: Richard M. Smith,
John M. Linacre, &
Ken Royal
Rasch SIG Chair: Jue Wang
Secretary: Eli Jones
Treasurer: Dandan Liao
Program Chairs: Manqian Lia &
Audrey Conway Roberts

Stability of Person Estimates Using the Rasch Model in the Presence of Varying Discriminations Among Items

With the popularisation and adaptation of models of modern test theory, introduced more or less independently in the middle part of the last century with the work of Lord (1952), Rasch (1960), and Birnbaum (1968), comparisons have been made among the efficacies in the application of different models. Although these comparisons were generalised for models with more than two ordered categories, many of the original comparisons were made between the models for dichotomous responses that are relevant for many tests of proficiency, in particular multiple-choice tests. This paper is concerned with tests composed of dichotomous items.

The models of most concern form a hierarchical structure in which the number of item parameters increases by one. Often the debates that seem to lead to controversy (Andrich, 2004) arise from the comparison between a general form of a model for dichotomous responses with three item parameters and one of its special cases with only one item parameter. The most general of the three forms of the model for multiple choice items, in which there is a potential for guessing, (Birnbaum, 1968) is given by

$$P\{x_{ni}=1;\beta_n,(\delta,\alpha,\gamma)\}=\gamma_i+1-\gamma_i[(\exp\alpha_i(\beta_n-\delta_i))/\Gamma_{ni}] \quad , \tag{1}$$

where $x_{ni} \in \{1,0\}$ for a correct and incorrect response respectively of person n responding to item i, β_n is the proficiency of person n and $(\delta_i,\ \alpha_i,\ \gamma_i)$ is a vector of item parameters respectively referred to as the difficulty, discrimination, and guessing propensity of an item, and $\Gamma_{ni} = 1 + (\exp\alpha_i(\beta_n - \delta_i))$. Eq. (1) is now known as the three-parameter logistic (3PL) model. A special case of this model is the two parameter logistic (2PL) in which there is no parameter γ_i, also known as a lower asymptote of the $P\{x_{ni} = 1\}$.. This takes the form

$$P\{x_{ni}=1;\beta_n,(\delta,\alpha)\}=(\exp\alpha_i(\beta_n-\delta_i))/\Gamma_{ni} \quad . \tag{2}$$

A further special case, but derived from a different rationale, is the model with no discrimination parameter for each item, which takes the form

$$P\{x_{ni}=1;\beta_n,\delta_i\}=(\exp(\beta_n-\delta_i))/\Gamma_{ni} \quad . \tag{3}$$

Eq. (3) is also known as the dichotomous Rasch model. Andrich (2004) argues that the source of controversy in the choice of model to analyse tests with dichotomous responses arises from the different paradigms that govern the application of the models. In one paradigm, which arises from the standard applications of statistics (Andrich, 2013), the goal of modelling is to find a model that best accounts for the data in terms of tests of fit. This is the paradigm of item response theory (IRT) exemplified in Bock (1997), Hambleton (2000), and many others. In the second paradigm, which arises from the theory of measurement of Rasch (1960, 1961) and is elaborated in Andrich (2018), Wright (1997), and others, the model does not arise from the data, but from a criterion of invariance of comparisons for the achievement of measurement. Then fit to the relevant Rasch model acts as a form of quality control that reveals the degree to which measurement has been achieved. It also diagnoses anomalies, for example problems with some items that do not contribute to the measurement.

In many of the comparisons between the models, the criterion of the first paradigm (IRT) was invoked, that is, the quality of the fit of the data to the chosen model. If the model with the least number of parameters, the Rasch model, showed misfit, then a model which had more parameters was tested, and if that model showed better fit, then it was concluded that the Rasch model should be rejected (Andrich, 2013). This argument did not convince those who subscribed to the paradigm of Rasch measurement theory. These paradigms are not reviewed further here.

Because a method of controlling guessing within the Rasch paradigm has been described (Andrich & Marais, 2014; Waller, 1989), this paper is not concerned with the most general model available, presented in Eq. (1). Instead, it is concerned with some comparisons between data that arises from the 2PL model, that of Eq. (2), and the Rasch model, that of Eq. (3). However, rather than comparing tests of fit when estimating the person parameters, this paper compares the stability of the person estimates when the data are generated from items with variable discrimination as well as variable difficulty, and when the person parameters are estimated using the Rasch model.

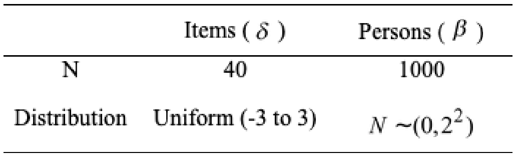
The purpose of this comparison is twofold. First, any model is a very strong summary of the relationships among the variables parameterised in the model, and no model fits perfectly. Therefore, although the Rasch model constrains the discriminations to be uniform, in particular $\alpha = 1$, in any analysis of most programs, it is most unlikely that all items in fact have exactly the same discrimination. Second, comparisons of the stability of the person parameter estimates when the data show different item discriminations, that is generated by the 2PL model, and are analysed using known item parameters according to the Rasch model and the 2PL model, seem not to have been performed. Instead, when comparisons have been made, they were, as indicated above, mostly in terms of tests of fit.

This paper shows the above comparison. To do so the paper uses a series of simulations and sets up a frame of reference for assessing the stability of the person estimates.

The Simulation Design

Table 1 shows the design of the simulations for the item and person parameters. There is nothing special about the item difficulties and the person distribution in these specifications, with the number of items (40) chosen to be reasonably reflective of proficiency tests, and their difficulties chosen to cover the range of proficiencies.

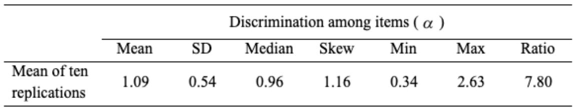
Table 1. *Location parameters for the simulated item and person distributions used in all ten replications*



	Items (δ)	Persons (β)
N	40	1000
Distribution	Uniform (-3 to 3)	$N \sim (0,2^2)$

Table 2 shows the summary statistics for the discriminations among the ten replications. These were generated with the following rationale. Proficiency items are designed to assess the same variable, while capturing different aspects of the variable. In any pilot study where data are obtained from dichotomous items, the items with a discrimination close to 0 or less than 0 would be re-examined and any observed problems corrected. If the study design is based on the Rasch model and paradigm, items with very high discrimination, which might show potential local dependence, would also be examined. Thus the empirical work would be designed to have items with relatively high and homogeneous discriminations. Then, in the final application of a test, if an item showed very low discrimination it would be removed. Thus there is a lower bound for the item discrimination. On the other hand, a similar a-priori constraint is not present on the upper bound of the discrimination, unless the item shows very high discrimination and it is evident that it is a result of local dependence on one or more other items. The method for choosing the discriminations among items was designed to reflect this asymmetrical feature of the distribution of discriminations. A normal distribution of parameters with a mean of 0 and variance of 0.5^2 was generated for the natural logarithm of the discrimination, $ln(\alpha)$: $N \sim (0, 0.5^2)$, for each replication of simulated responses for the same distribution of person parameters, and then exponentiated to give the distribution of α. Table 2 shows the summary statistics for α among the ten replications. It is evident that the distribution is positively skewed around a value close to 1. The Table also shows the ratio of the maximum and minimum discrimination, suggesting the range in actual discriminations is relatively realistic.

Table 2. *Summary statistics for the simulated discrimination values among the items with* varying discrimination, ln(α): N ~ (0, 0.5^2)*, and the ratio of the maximum to minimum discrimination, summarised over ten replications*



	Discrimination among items (α)						
	Mean	SD	Median	Skew	Min	Max	Ratio
Mean of ten replications	1.09	0.54	0.96	1.16	0.34	2.63	7.80

The Comparisons

The form of the comparisons is between the known, simulated person value, β_n, for each person and the estimated value, $\hat{\beta}_n$, under different estimation procedures. This comparison is summarised by the root mean square, $RMS = \sqrt{\sum_{n=1}^{N}(\beta_n - \hat{\beta}_n)/N}, N = 1000$.

In addition, to understand the stability of the person estimates, similar comparisons were made for the item difficulty estimates from the dichotomous Rasch model, with

$RMS = \sqrt{\sum_{i=1}^{I}(\delta_i - \hat{\delta}_i)/I}, I = 40.$

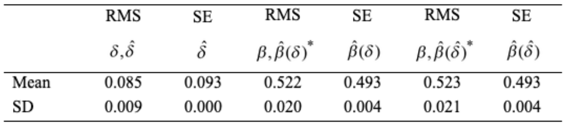
Summary statistics for the items over the ten replications are shown in Table 3.

The Frame of Reference for the RMS

To provide a frame of reference for the order of magnitude of a *RMS* that can be produced with different estimation procedures, a series of simulations were made where the items satisfied the dichotomous Rasch model (α = 1). Then the person parameters were estimated under two conditions, first when the known item parameters were used and second, when the item parameters were estimated.

Table 3 shows the RMS values for both items and persons when the data are generated according to the dichotomous Rasch model (α = 1). The RMS for items compares the known difficulties of the items with the estimated difficulties, which are obtained from the RUMM2030Plus software (Andrich, Sheridan, & Luo, 2020). This software uses the pairwise, conditional method of estimation for the item parameters in which the person parameters are eliminated (Andrich & Luo, 2003). Maximum likelihood estimates (MLE) of the person parameters are then obtained given the item difficulties. The person parameters were obtained from two sets of item parameters: first, from values known from the simulation and second, from estimated values. First, it is evident that the estimates of the item difficulties are very close to their known values, with the order of magnitude of the RMS being that of the standard error of the estimate. Second, it is evident that the RMS for the person locations are of the same order of magnitude for both sets of item parameters, and of the order of magnitude of the standard errors, approximately 0.5 logits. This last value provides the frame of reference for the magnitude of the RMS when the data are generated by the 2PL model, that is, $\alpha \neq 1$ in the data.

Table 3. *RMS with the simulated values for the item and person estimates from data generated and analysed according to the dichotomous Rasch model* (α = 1)*, summarised over ten replications*



	RMS	SE	RMS	SE	RMS	SE
	$\delta,\hat{\delta}$	$\hat{\delta}$	$\beta,\hat{\beta}(\delta)^*$	$\hat{\beta}(\delta)$	$\beta,\hat{\beta}(\hat{\delta})^*$	$\hat{\beta}(\hat{\delta})$
Mean	0.085	0.093	0.522	0.493	0.523	0.493
SD	0.009	0.000	0.020	0.004	0.021	0.004

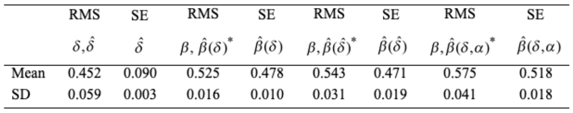
$\hat{\beta}(\delta)^*$, $\hat{\beta}(\hat{\delta})^*$ are β estimates with known and estimated item difficulties, $\delta, \hat{\delta}$ respectively.

RMS from Variable Item Discriminations

The *RMS* values of the item and person parameter estimates shown in Table 4 were obtained from data generated by the 2PL model with parameters summarised in Tables 1 and 2. The item difficulties were estimated using the dichotomous Rasch model in which the discrimination of all items is constrained to 1, then three different MLE of the person parameters were obtained. First when the item parameter difficulties were taken as known and the Rasch model was assumed (all α = 1), second when the item difficulty estimates from the Rasch model were used, and third when the known item difficulty and discrimination parameters were used and the 2PL model was assumed.

First, it is evident that the *RMS* value for the item difficulty estimates is substantially larger compared to when the data were generated according to the Rasch model, 0.452 (Table 4) compared to 0.085 (Table 3). This result reflects that the item difficulty estimates are affected by differences in the item discriminations. The order of magnitude of the RMS is close to half a logit. Second, the three *RMS* values for the three person estimates are of the same order of magnitude, approximately half a logit, which is equal to the value when the data fit the Rasch model.

Table 4. *RMS with the simulated values for the item and person estimates from data generated by the 2PL model* $(\alpha \neq 1)$ *and analysed according to the dichotomous Rasch model* (α = 1)*, summarised over ten replications*



	RMS	SE	RMS	SE	RMS	SE	RMS	SE
	$\delta,\hat{\delta}$	$\hat{\delta}$	$\beta,\hat{\beta}(\delta)^*$	$\hat{\beta}(\delta)$	$\beta,\hat{\beta}(\hat{\delta})^*$	$\hat{\beta}(\hat{\delta})$	$\beta,\hat{\beta}(\delta,\alpha)^*$	$\hat{\beta}(\delta,\alpha)$
Mean	0.452	0.090	0.525	0.478	0.543	0.471	0.575	0.518
SD	0.059	0.003	0.016	0.010	0.031	0.019	0.041	0.018

$\hat{\beta}(\delta)^*$, $\hat{\beta}(\hat{\delta})^*$ are β estimates with known and estimated item difficulties, $\delta, \hat{\delta}$ respectively, respectively, assuming the dichotomous Rasch model; $\hat{\beta}(\delta,\alpha)^*$ are β estimates with known item difficulty and discrimination parameters.

Specifically, using the known item difficulties but ignoring differences in discrimination when estimating person parameters with the Rasch model, the *RMS* is of the same order of magnitude as when using both known item difficulties and discriminations to estimate person parameters with the 2PL model with which the data were generated. Typically, of course, even though the item discriminations are not expected to be exactly the same in empirical data, the Rasch model estimates item difficulties assuming identical discriminations, which are then used to estimate the person parameters. Thus the RMS for the Rasch model estimates

$(\beta, \hat{\beta}(\hat{\delta}))$ is perhaps the most realistic value to compare with that generated from the 2PL model estimates $\beta, \hat{\beta}(\delta, \alpha)$. Again, there seems to be no difference in the order of magnitude between these two sets of estimates, and these are of the order of magnitude of the standard errors of measurement for an individual.

Some Inferences from the Simulation Study

The first inference from the above study is that in the presence of different discriminations, and although the item difficulty estimates are affected noticeably, the person parameter estimates are very stable. In particular, the order of magnitude of the difference between known and estimated values is the same as when the data fit the model perfectly. Thus even though the person estimates for the same total score are the same in the Rasch model and variable in the 2PL model depending on the pattern of responses, the summary accuracy of the person estimates remains stable. This may result from the compensation that occurs where some items have higher and some have lower discriminations. In the Rasch model, the implied single discrimination is a kind of average discrimination among all the items. The second inference is that estimates of person parameters using the 2PL model and the Rasch model to analyse dichotomous responses, when the data follow the 2PL model, are effectively the same. This may be relevant for comparing studies involving person estimates when some studies have used the 2PL model and some have used the Rasch model.

A question that then arises is, are there any advantages in using the 2PL model rather than the Rasch model in the construction and application of measuring instruments in education and the social sciences? The advantages of using the Rasch model as an empirical framework and criterion for measurement have been well documented, namely the invariance of comparisons within a frame of reference, the sufficiency of the total score for a person estimate, the provision of measurement of the form of the natural sciences, and of course its relative simplicity (Wright, 1997). In addition, the difficulties of the items can be placed on the same continuum more or less like the markings on a measuring instrument, whereas in principle, the two item parameters of the 2PL model need to be considered jointly. Typically, the benefit of the 2PL model is that the data fit the model better, which is inevitable given the greater number of parameters, with data that are unlikely to fit any model perfectly. However, although better fit will be shown by the 2PL because it has more parameters estimated, overall it does not recover the known person estimates better than the Rasch model even when the data are generated by the 2PL model. Further studies need to be made to confirm the generality of the result of this example and they might show differences at a finer level of analysis, for example the impact on the variances.

References

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*, 7–16.

Andrich, D. (2013). The legacies of R. A. Fisher and K. Pearson in the application of the polytomous Rasch model for assessing the empirical ordering of categories. *Educational and Psychological Measurement*, *3*(4), 553–580.

Andrich, D. (2018). Advances in social measurement: A Rasch measurement theory. In F. Guillemin, A. Leplège, S. Briançon, E. Spitz, and J. Coste (Eds.), *Perceived health and adaptation in chronic disease: Stakes and future challenge* (Chapter 7, pp 66–91). CRCS Press, Taylor and Francis.

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

Andrich, D. & Marais, I. (2014). Person proficiency estimates in the dichotomous Rasch model when random guessing is removed from difficulty estimates of multiple choice items. *Applied Psychological Measurement*, *38*(6), 432–449.

Andrich, D., Sheridan, B.S. & Luo, G. (2020). RUMM2030Plus: An MS Windows computer program for the analysis of data according to Rasch Unidimensional Models for Measurement. Perth, Western Australia: RUMM Laboratory.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–545). Reading, Mass.: Addison-Wesley.

Bock, R.D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16*(4), 21–33.

Hambleton, R.K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care*, *38* (Suplement II), II-28-II-42.

Lord, F.M. (1952). A theory of test scores. *Psychometric Monograph No.7, Psychometric Society.*

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. Expanded edition (1980). Reprinted (1993) Chicago: MESA Press.

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.

Waller, M.I. (1989). Modeling guessing behavior: A comparison of two IRT models. *Applied Psychological Measurement, 13*(3), 233–243.

Wright, B.D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33–45.

David Andrich and Sonia Sappl

The University of Western Australia

Update on Journal of Applied Measurement: New Publisher and Editorial Team

We are pleased to announce that the Journal of Applied Measurement has a new home and a new editorial team. The December 2020 (JAM V21, N4) issue of JAM was the last under founding editor and editorial board. Beginning with Volume 22, JAM will be located at **National Taiwan Normal University (NTNU),** located in Taipei, Taiwan. The new editor is Prof. Hak Ping Tam. A new editorial board is being developed to assist the editor. The journal will now be printed in Taiwan, so there may be changes to the postage fees associated with mailing the journal. For information about subscription and manuscript submission procedures, please contact the new editorial team at jamntnu@gmail.com or visit the new JAM website at http://www.jamntnu.net/ , which is currently under construction and will be functional in the near future.

Despite a change of publishers, JAM will continue to serve the Rasch community in particular and the measurement community in general. Suggestions of potential topics for Special Issues that are of academic or common interests are especially welcome. The current JAM website will be revised to reflect this change. All previous JAM information, including the table of contents for the first 21 volumes of JAM and the list of abstracts for articles published in those 21 volumes, will still be available at http://jampress.org/ . The pdfs of JOM issues will also remain on the website. The website will continue to support the activities of JAM Press. All information about books published by JAM Press and JAM Press materials will continue to be available for purchase.

As the now former editor, I would like to thank all of the people who made JAM possible for the last 21 years, including the editorial board, the subscribers, and contributing authors, who believed enough in JAM's mission to share the journey with us. Special thanks to Judy E. Teska, who did all the layout work in preparing every article that appeared in all 21 volumes and supervised all of the communications with the various printers that printed the final JAM journals. Judy maintained the highly professional style that JAM achieved during that time.

Hak Ping Tam

Richard M. Smith

Journal of Applied Measurement
Call for Papers for Special Issue on Unfolding Models

The Journal of Applied Measurement (JAM) is planning a special issue featuring research on unfolding models. The editors are dedicating this issue to the publication of exemplars of important scholarship in the area of unfolding models.

Papers are sought that present interesting and innovative approaches to unfolding models. Papers on research, theory, and practice related to unfolding models in a variety of contexts will be considered for inclusion in this Special Issue.

Based on peer review, eight to ten of the most competitive papers will be published in this special issue of JAM. Submissions will be refereed according to standard procedures for JAM.

The special issue editors are George Engelhard (The University of Georgia), Ye Yuan (The University of Georgia), and Jue Wang (The University of Miami).

Timeline: Submission of manuscripts (Fall 2021 with final submission date of January 1, 2022), peer review of manuscripts (Spring 2022), and author revisions of manuscripts (Summer 2022).

Please submit your manuscripts and any questions to George Engelhard (gengelh@uga.edu).

George Engelhard, Jr.

The University of Georgia

2021 Georg William Rasch Early Career Publication Award Recipient

Dr. Wen-Chia Chang, Research Fellow affiliated with the International Coalition for Multilingual Education and Equity at the University of Nebraska Lincoln, is the recipient of the Georg William Rasch Early Career Publication Award for 2021. This award recognizes individuals for outstanding publications of Rasch measurement research.

Dr. Chang received her Ph.D. from the Boston College, Lynch School of Education and Human Development, Department of Measurement, Evaluation, Statistics, and Assessment in May 2017. Dr. Chang's program of research focuses on evaluation and measurement issues related to teaching and teacher education for equity and social justice. She is the co-author of a book, *Reclaiming Accountability in Teacher Education* (Teachers College Press, 2018) that won four awards, including the AACTE Best Book award in 2020. Her recent research projects include applying an argument-based framework to scale validation using a mixed-methods approach.

Dr. Chang has been nominated based on the following paper.

Chang, W. C. C., Ludlow, L. H., Grudnoff, L., Ell, F., Haigh, M., Hill, M., & Cochran-Smith, M. (2019). Measuring the complexity of teaching practice for equity: Development of a scenario-format scale. *Teaching and Teacher Education*, *82*, 69-85.

Here are the highlights of this article for interested readers.

- Equity-centered teaching practice promotes student learning and challenges inequity.
- Six interconnected principles of teaching practice for equity are introduced.
- The application of Rasch measurement and Guttman facet theory is illustrated.
- Detailed generalizable procedure for developing scenario-format items is presented.
- Score interpretation and implications to teacher education research are discussed.

As the recipient of the Early Career Award, Dr. Chang will be delivering the keynote address at the Rasch SIG Business meeting at 2022 AERA in San Diego, California.

Congratulations to Dr. Chang!

Jue Wang

AERA Rasch SIG Chair

Conference Announcement: NCME Special Conference on Classroom Assessment

The 4th National Council on Measurement in Education (NCME) Special Conference on Classroom Assessment will be held virtually on October 21 and 22nd, 2021.The conference will bring together K-12 teachers, school and district leaders, higher education faculty, and researchers to engage in dialogue, discussion, and learning to strengthen the practice and potential of classroom assessment and shape classroom assessment research needed to address current challenges faced by educators. Using a variety of session formats, the conference will blend the dynamic realities of the classroom with research and theory. For more information, please see: https://www.ncme.org/meetings/upcoming-events

Tonya R. Moon
University of Virginia