SPRING 2021, VOL. 34 NO. 1; 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 three research notes and several announcements that may be of interest to the Rasch community.

The issue begins with a research note from David Andrich related to random distributions and the Rasch model. This is followed by a research note from Adrienne Walker related to person fit and score interpretations within the context of the COVID-19 pandemic. The third research note is Mike Linacre's survey of R packages that can perform Rasch analyses and some tips. Mike's research note is a great companion to the Fall 2020 research note by Govindasamy, Green & Olmos that also focused on Rasch modeling using R packages.

Following the research notes, we have included a reflection on the Virtual International Objective Measurement Workshop 2020 (IOMW) written by the IOMW organizing committee.

In other conference news, we have provided a list of presentations and other events that are scheduled during the American Educational Research Association (AERA) conference related to Rasch measurement theory.

The issue concludes with a list of recent publications in the Journal of Applied Measurement.

As always, we welcome your contributions to the next issue for RMT. 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: Hong Jiao
Secretary: Cari F. Herrmann-Abell
Treasurer: Matt Schulz
Program Chairs: Trent Haines &
Courtney Donovan

A Property of All Random Distributions: Relationship to the Rasch Model Distribution

It is surprising that a defining property of all *random* distributions is taken for granted and not explained in textbooks. It is particularly surprising given that: first, concluding that a distribution is random is central to statistical analyses of data; second, that there is a published definition.

A distribution, real or inferred, arises from the concept or reality of replicated outcomes. By *replication* it is meant that the same factors govern the outcomes from the set of possible outcomes, and that there is a probability, not certainty, for each outcome. The sum of the probabilities is 1. The typical introductory textbook examples involve tossing of the same coins or the same dice. In measurement, it arises from the concept that the same instrument is used to measure the same object on repeated occasions. The Gauss (normal) distribution was derived for exactly that concept: *''Laws of error,'' i.e., probability distributions assumed to describe the distribution of the errors arising in repeated measurement of a fixed quantity by the same procedure under constant conditions, were introduced in the latter half of the eighteenth century to demonstrate the utility of taking the arithmetic mean of a number of measurements or observed values of the same quantity as a good choice for the value of the magnitude of this quantity on the basis of the measurements or observations in hand.* (Eisenhart, 1983).

A random empirical distribution implies that all systematic factors that govern the distribution have been taken into account. Reciprocally, evidence of non-randomness in a real or inferred distribution implies that factors not accounted for have disturbed the outcomes.

What, then, is this defining property of random distributions? Descriptively, they must be *single-peaked* or *strictly unimodal*, that is, they cannot have two or more modes or peaks. However, strict unimodality is not enough - the transition between probabilities of adjacent counts or measurements must be *smooth*. For distributions to have smooth transitions between adjacent probabilities, they need to be *strictly log-concave* (SLC). The definition of SLC for continuous distributions is provided by Ibragimov (1956) and for discrete distributions by Keilson and Gerber (1971). Because this note relates the Rasch model to the definition of SLC, the concern here is with discrete distributions only.

Strict Log Concavity

Let P_x be the probability of a random variable of integer counts $x,\ x=0,1,2,3,...m$. Then the distribution of P_x is SLC if the SLC ratio ($SLCR$) satisfies:

$$SLCR = P_x^2 / (P_{x-1}P_{x+1}) > 1. \qquad (1)$$

If:

$$SLCR = P_x^2 / (P_{x-1}P_{x+1}) > 1\ ,$$

then:

$$\ln(P_x^2 / (P_{x-1}P_{x+1})) > 0$$

and:

$$\begin{aligned}
&\ln(P_x^2) - \ln(P_{x-1}P_{x+1}) > 0\\
&i.e.\ \ 2\ln P_x - (\ln P_{x-1} + \ln P_{x+1}) > 0\\
&i.e.\ \ 2\ln P_x > (\ln P_{x-1} + \ln P_{x+1})\\
&i.e.\ \ \ln P_x > (\ln P_{x-1} + \ln P_{x+1}) / 2,\ x = 1,2,3,..,m-1.
\end{aligned}$$

(2)

Thus, the logarithm of the probability of a count x is greater than the mean of the logarithms of the probabilities adjacent to x. This gives a concave distribution and suggests the name SLC. Because a probability and its logarithm are related monotonically, the distribution P_x is also strictly concave, and hence strictly unimodal. It is possible for two adjacent counts to have the same probability and the distribution still be SLC.

To illustrate strict log-concavity, Panel 1 of Figure 1 shows a binomial distribution with a maximum count of 4 in which the probability of a success is 0.75 and therefore the mean, $E[X]=3$. It also shows the SLC ratios, which are all positive. Panel 2 shows the log probability distribution in which the relevant means are also shown.

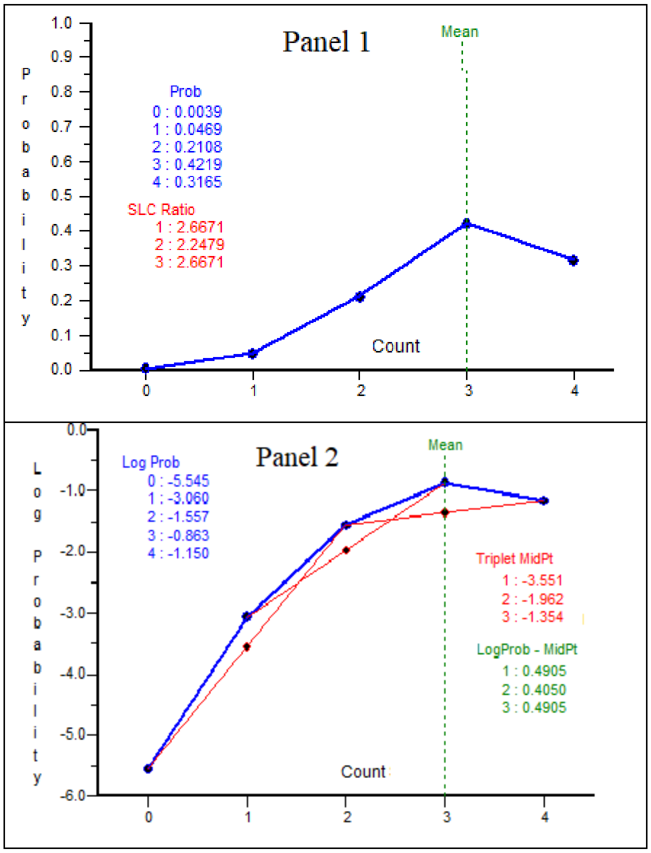


Figure 1. Binomial distribution (0.75, 4) (Panel 1) and log Binomial distribution (Panel 2) in which $E[X]=3$.

Although a SLC distribution is strictly unimodal, a strictly unimodal distribution is not necessarily SLC. The distribution in Panel 1 of Figure 2, which is both strictly unimodal and symmetrical, is not SLC. It is not SLC because of a relative deficiency in the probabilities of counts 1 and 3. Two of the three SLCRs, which are less than 1 (0.3749), reflect this for these counts. Panel 2 of Figure 2 shows where the log of a probability is less than the mean of logs of its adjacent probabilities. Complementary to these deficiencies, there is a surplus in the probability of 2 even though $E[X]=2$.

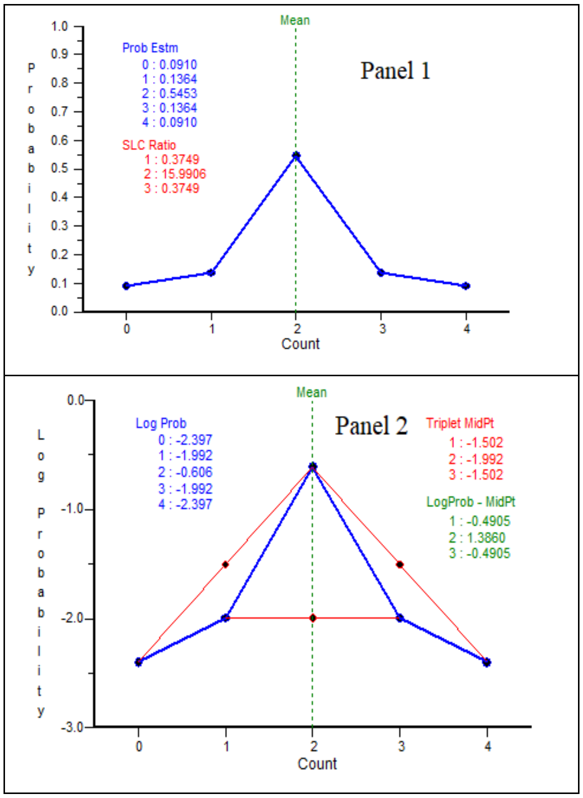


Figure 2. A strictly unimodal distribution (Panel 1) which is not SLC (Panel 2).

Because they are more constrained than those which are strictly unimodal, and it is a property of all random distributions, Andrich and Pedler (2019a) coined the term *randomly unimodal* for SLC distributions. If the distribution in Figure 2 were empirical, it could be inferred that at least one systematic factor is contributing to the surplus in the probability of the count of 2 and the deficiency in counts of 1 and 3. Perhaps some local dependence was present.

The Rasch Model Distribution

The Rasch model (RM) of concern here is that for ordered response categories. The model can be written in the form

$$P_x = \Pr\{x;\beta_n,\delta_i\} = \exp(-\tau_{i0}-\tau_{i1}-\tau_{i2}-...-\tau_{ix}+x\beta_n\}/\gamma_{ni} \qquad (3)$$

where x, $x=0,1,2,...,m$ is an integer variable that is the *count* of the number of thresholds $\tau_{i1},\tau_{i2}...,\tau_{ix}$ deemed to have been exceeded on item i, β_n is the measure of person n, and γ_{ni} is a normalising factor (Andrich, 1978). The thresholds are points of equal probability of two adjacent categories on the continuum of measurement, analogous to a single threshold in the case of dichotomous responses, with $\tau_{i0}\equiv 0$ for convenience of notation.

Unfortunately, the distributional meaning of Eq. (3) is generally ignored. It is the *inferred distribution of replicated* responses, that is, *as if the same person* (or another with the same person parameter), responded to the same item (or an item with the same parameters), with responses that are statistically independent. When data are analysed and the thresholds are estimated for an item, then for any person parameter value β_n, the distribution is an inferred distribution as if people with exactly that person parameter responded independently to that item. That is,

$$\hat{P}_x = \Pr\{x;\beta_n,(\hat{\tau}_i)\} = \exp(-\hat{\tau}_{i0} - \hat{\tau}_{i1} - \hat{\tau}_{i2} - ... - \hat{\tau}_{ix} + x\beta_n\} / \hat{\gamma}_{ni} \quad (4)$$

is an *inferred distribution of replicated responses* for any value β_n to an item with parameters $(\hat{\tau}_i)$.

Panel 1 of Figure 3 shows the typically represented category characteristic curves (CCCs) for a RM item with four thresholds and five ordered categories. In addition, and not typically shown, is the *inferred* distribution of replications for the value $\beta = 0$ highlighted with ●. Panel 2 of Figure 3 shows explicitly this single distribution for $\beta = 0$. Panel 2 also shows the probabilities of each count and the SLC ratios (2.7069, 2.4239, 2.4942) for adjacent counts – they are all greater than 1 and the distribution of inferred replications is SLC with smooth transitions between adjacent probabilities. Unlike the distribution in Figure 2, there is no evidence that there is a factor(s) that is disturbing the randomness of the distribution. The mean of this distribution is a non-integer, $E[X] = 2.3605$, while the mode is clearly $x = 2$.

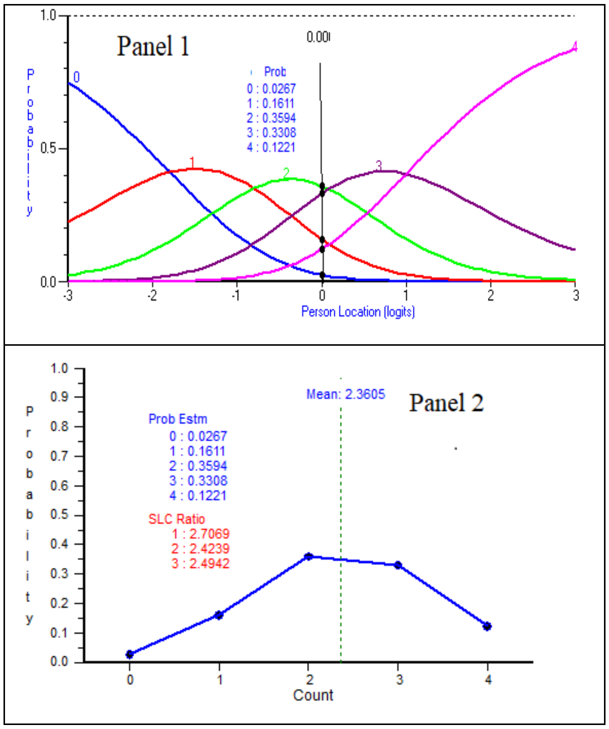


Figure 3. CCCs with ordered thresholds (-1.8, -0.8, 0.08, 1.0) and distribution (●) for $\beta = 0$ (Panel 1); the distribution for $\beta = 0$, mode 2, $E[X] = 2.3605$ and SLC (Panel 2).

Panel 1 of Figure 4 shows the CCCs of another item with the distribution highlighted (●) for $\beta = -0.430$. Panel 2 shows explicitly this single distribution for

$\beta = -0.430$. This distribution is clearly bimodal (not strictly unimodal) and therefore not SLC. This distribution is not a random distribution – there is some factor disturbing the distribution leading to a deficit of responses in the count 2 and a relative surplus in counts 1 and 3. The SLC ratio (0.3329) involving the count 2 is less than 1.

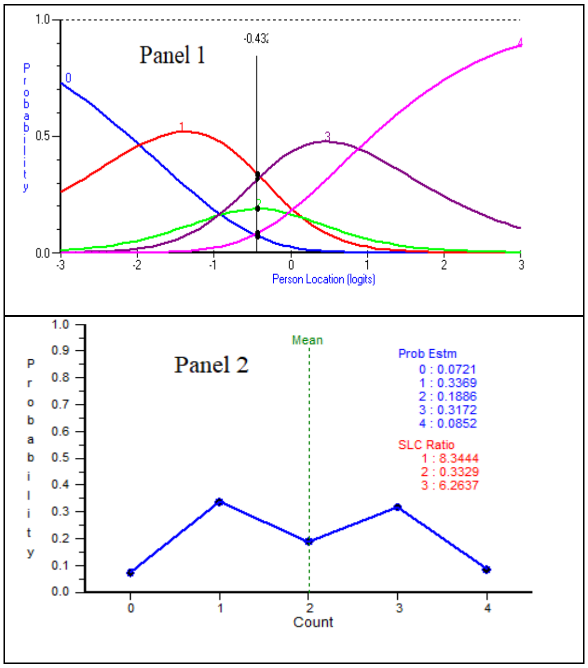


Figure 4. CCCs with disordered thresholds (1.97, 0.15, -0.95, 0.89) and distribution (●) for $\beta = -0.430$ (Panel 1); the distribution for $\beta = -0.430$, bimodal 1,3, $E[X] = 2.0$ and not SLC (Panel 2).

It is stressed that the distribution in Figure 4 is the *inferred distribution* as if the person with location $\beta = -0.430$ was assessed on repeated occasions with the same item, which implies the same threshold estimates – it is the inferred distribution of replicated outcomes.

To make the implications of the distributions in Figures 3 and 4 concrete, suppose that the category classification arose from diagnoses of X-Rays from a medical check-up and that the category classifications were: *Reassess in five years (0); Reassess in one year (1); Drug treatment (2); Minor surgery (3); Major surgery (4).* From the diagnosed treatments, there is a clear order in the degree of malady observed in the X-Ray.

Inferred Distribution of Figure 3

Consider first the distribution in Figure 3. Suppose that the threshold estimates arose from the training of say 20 novice radiologists with many relevant X-Rays, and that (i) the model-data fit was excellent, and (ii) all radiologists produced thresholds statistically equivalent to those depicted in Figure 3. As a result of the fit and invariance of estimates among the novice radiologists, they were all certified to read and assess the same kinds of X-Rays.

Suppose next that a case presents to one of these novice radiologists in which, from replicated expert assessments, the location estimate of the X-Ray was $\beta = 0$ as in Figure 3. Although the radiologist gives only one assessment, the distribution from which this one assessment is observed is that depicted in Panel 2 of Figure 3. The most likely assessment is *Drug treatment* $(x = 2)$ with the next most likely assessment being adjacent to it, *Minor surgery* $(x = 3)$. The third most likely assessment is *Reassess in 1 year* $(x = 1)$. Thus, the second and third most likely assessments are adjacent to the one that is most likely and the mean of the distribution is $E[X] = 2.3605$, somewhere between the most likely count $(x = 2)$ and the next most likely $(x = 3)$. A defining point of the SLC distribution is that as assessments deviate successively from the mode of 2, then their probabilities also successively decrease. The distribution suggests that there might be other factors playing a role leading to low probabilities of the other assessments, but because the distribution is SLC, the evidence is that these factors are random. The distribution seems tolerable.

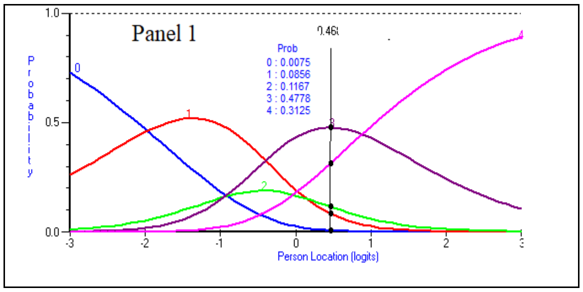
Inferred Distribution of Figure 4

Suppose next that the threshold estimates in Figure 4 also arose from the training of 20 novice radiologists with many relevant X-Rays, and that again (i) the model-data fit was excellent, and (ii) all radiologists produced thresholds statistically equivalent to those depicted in Figure 4. As a result of the fit and invariance of estimates among the novice radiologists, again they were all certified to read and assess the same kinds of X-Rays.

Suppose next that a case presents to one of these novice radiologists in which, from replicated expert assessments, the location estimate of the X-Ray was $\beta = -0.430$ as in Figure 4. Although the radiologist gives only one assessment, the distribution from which this one assessment is observed is that shown in Panel 2 of Figure 4. The two equally likely assessments are both most likely, that is *Reassess in one year* $(x = 1)$ and *Minor surgery* $(x = 3)$. The third most likely assessment is *Drug treatment* $(x = 2)$, the category in between the other two even though the mean of this distribution, $E[X] = 2.0$. This distribution is not random – some factor is forcing a bimodal distribution with the two modes on either side of the mean. While it is expected that there might be uncertainty in these assessments, it would not be considered acceptable that the uncertainty is of a kind that provides a bimodal distribution. It cannot be inferred from this distribution that the uncertainty is no more than random, and therefore it cannot be tolerated.

Before proceeding with the formalization of an SLC distribution and the RM, Figure 5 shows the distribution for another X-Ray location, $\beta = 0.46$, for the same item as that in Figure 4. Here the mean of the distribution, $E[X] = 3.0$. Because the value $\beta = 0.46$ was chosen, the distribution has greater probabilities for the higher counts and is strictly unimodal. However, Panel 2 of Figure 5 shows that the SLC ratio for a count of 2 is less than 1. In this distribution, there is a deficit in the count of $x = 2$ and because it is not SLC, it is not a random distribution.

It might be observed that the SLC ratios in Figures 4 and 5 are the same, $(8.3444,\ 0.3329,\ 6.2367)$, even though the locations of β are different, and therefore the distributions themselves are different. It is shown next that this invariance of the SLC ratio is a property of the RM.



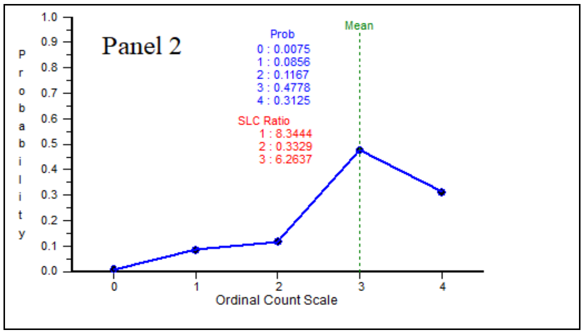


Figure 5. CCCs with disordered thresholds (1.97, 0.15, -0.95, 0.89) and distribution (●) for $\beta = 0.46$ (Panel 1); the distribution for $\beta = 0.46$ is unimodal, $E[X] = 3.0$, but not SLC (Panel 2).

The Rasch Model and SLC

Inserting Eq. (4) of the RM into Eq. (1) gives, on simplification,

$$SLCR = P_x^2 / (P_{x-1}P_{x+1}) = \exp(\tau_{x+1} - \tau_x),\ x = 1, 2, 3, \ldots, m-1. \qquad (5)$$

Eq. (5) has two important properties. First, the ratios are only a function of the thresholds, and are independent of the location of the person – the person parameter drops out in Eq. (5). Thus the SLCR is a property of the item, and not of any particular person location. That is the reason why the SLCR in Figures 4 and 5 are the same – it is because the thresholds are identical.

Second, it is evident that if

$\tau_{x+1} > \tau_x,\ x = 1, 2, 3, \ldots, m-1$, then

$\tau_{x+1} - \tau_x > 0$ and the $SLCR = \exp(\tau_{x+1} - \tau_x) > 1$.

That is, if the thresholds are ordered, then the SLCR values are greater than 1, the distribution is SLC and there is no evidence that it is not a random distribution. On the other hand, if $\tau_{x+1} < \tau_x$, that is a pair of thresholds is disordered, then for some $x = 1, 2, 3, ..., m-1$, $\tau_{x+1} - \tau_x < 0$ and the $SLCR = \exp(\tau_{x+1} - \tau_x) < 1$. This implies that the distribution is not SLC for any person location β with some counts deficient and some in surplus relative to randomness. Moreover, there are values of β for which the distribution such as that shown in Panel 2 of Figure 4 will be bimodal.

The Rasch Model and the Binomial Distribution

For completeness, the binomial, which was used to introduce SLC distributions, is recast now into the form of the RM. The binomial takes the form

$$P_x = \binom{m}{x} \pi^x (1-\pi)^{m-x} \qquad (6)$$

where π is the probability of a success for a dichotomous response, m is the number of statistically independent replications, and x is the number of successes. Let $\pi = \exp(\beta) / [1 + \exp(\beta)]$. Then Eq. (6) becomes

$$\begin{aligned} P_x &= \binom{m}{x} \left(\frac{e^\beta}{1+e^\beta} \right)^x \left(\frac{1}{1+e^\beta} \right)^{m-x} \\ &= \binom{m}{x} \frac{e^{x\beta}}{(1+e^\beta)^m} \\ &= \exp\{\ln \binom{m}{x} + x\beta\} / (1+e^\beta)^m, \end{aligned}$$

$$(7)$$

where $\ln \binom{m}{x} = -\tau_0 - \tau_1 - \tau_2 - ... - \tau_x$ in the RM of Eq. (3). That is:

$$\ln \binom{m}{x} = -\tau_0 - \tau_1 - \tau_2 - ... - \tau_x .$$

$$(8)$$

After some tedious but routine simplification,

$$\tau_x = \ln[x / (m - x + 1)],\ x = 1, 2, ...m$$

$$(9)$$

For example, with $m = 4$ as in the example in Figure 1, $\tau_1 = -1.39$, $\tau_2 = -0.41$, $\tau_3 = 0.41$, $\tau_4 = 1.39$. Clearly, the thresholds are ordered. It can be shown that in general,

$$\tau_{x+1} - \tau_x = \ln[1/(1+x)] + \ln[1 + 1/(m-x)],\ x = 1, 2, ...m$$

$$(10)$$

and because each term in Eq. (10) is greater than 1, its value must be greater than 0, that is, it is positive and the thresholds are ordered.

The Rasch Model and Test of Fit

Because the thresholds of the binomial are defined, any test of fit will not involve the estimates of the thresholds – they are part of the distribution. However, with the general RM, the thresholds are estimated. Then the reason that tests of fit do not bear on whether the distribution, following the estimates of the thresholds, is random or not is that the values of the threshold estimates are used to recover the distribution, taking account of the loss of degrees of freedom. Thus if the threshold estimates are reversed, it is these reversed estimates that are used to recover the relevant frequencies in any test of fit. Thus with the RM, evidence of fit is not sufficient to ensure a random distribution (Andrich & Pedler, 2019b). And of course, evidence of naturally ordered thresholds, though necessary, is not sufficient to conclude that the distribution is no more than random.

It might be noted that typically, tests of fit involve standardised residuals, yet the above exposition was concerned with the distribution itself. Because a standardised residual is derived from a linear transformation given the mean and standard deviation, the distribution of the residuals is identical to the distribution itself. Therefore, if the distribution is bimodal, the distribution of the residuals is also be bimodal, and the residuals cannot be considered random.

In summary, if the threshold estimates in the RM are not in their natural order, then there is at least one $SLCR < 1$, $x = 1, 2, 3, ..., m-1$ and the resultant distribution is not a random distribution. Therefore there is evidence that all the factors that affect the distribution have not been taken into account by the model. If the system of ordered classification is of the kind shown above with the X-Rays, I would suggest that no-one would want to have his or her X-Ray assessed by a radiologist who has CCCs, following training, of the kind shown in Figure 4.

David Andrich

The University of Western Australia

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-574.

Andrich, D. & Pedler, P. (2019a). A law of ordinal random error: the Rasch measurement model and random error distributions of ordinal assessments. *Measurement*, *131*, 771–781.

Andrich, D. & Pedler, P. (2019b). Modelling ordinal assessments: fit is not sufficient.

Communication in Statistics, 48, 2932–2947. https://doi.org/10.1080/03610926.2018.1473595

Eisenhart, C. (1983). Law of error I: Development of the concept. In S. Kotz & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 4, pp. 530-547). Toronto: Wiley.

Ibragimov, I. A. (1956). On the composition of unimodal distributions. *Theory of probability and its applications*, Volume I, 255 – 260.

Keilson, J. & Gerber, H. (1971). Some results for discrete unimodality. *The Journal of the American Statistical Association, 66,* 386 – 389.

Person Fit for COVID-19 Score Interpretation

The learning disruptions due to the COVID-19 pandemic have serious implications for the interpretation and use of standardized test scores for 2020-2021, and potentially beyond. Even if the many challenges of test administration (e.g., remote vs. in-person) and scoring (e.g., item stability and equating) during the 2020-2021 school year can be overcome and test scores are reported, there is still the question of how best to interpret and use these scores. In this note I argue that person fit information can inform how to use test scores.

The idea of using person fit to inform score interpretation and use is based on the premise that all scores, or estimated measures, obtained during a given testing event are not equally useful, and that the degree of model-data fit observed between the item responses and estimated measure that is derived from the scoring model dictates how much the estimated measure can be trusted (and used) as a measure of a person's achievement. This premise holds for aggregated measures, for example subgroups of students, and importantly, for individual measures.

In the Rasch framework, person fit procedures flag students that provided unusual patterns of item responses given their total score. In general, misfit suggests lack of model-data fit to the Rasch statistical requirements. Extreme misfit implies that the measured construct may be different for the misfitting student than for non-misfitting students (Glas & Khalid, 2017; Meijer & Sijtsma, 2001). Practically, this means that the intended score interpretation, which is consistent with the developmental foundations of the test and the interpretational materials that were developed for it (e.g., the achievement level descriptors) may not hold for misfitting students. In other words, for misfitting students the meaning behind the test score is unknown.

Person fit information can assist stakeholders in knowing how much trust to afford the score for determining what the student should do next. Given that scores from 2020-2021 are judged appropriate enough to report, educators, parents, and students may be looking for additional information regarding the interpretation and use of these scores. For many students, the score will still provide an interpretable and useful piece of information, given the pandemic context (non-misfitting students). For other students, additional evidence will be needed to confirm the student's standing on important skills (misfitting students). Person fit procedures can help distinguish between these two outcomes and can be added to existing operational procedures, quality checks, and score interpretation guidance. Now may be a good time for testing professionals and stakeholders to explore the utility of person fit as a model-fit indicator in score interpretation practice.

A. Adrienne Walker
Georgia Department of Education

References

Glas, C. A. W., & Khalid, N. (2017). Person Fit. In W. van der Linden (Ed.), *Handbook of item response theory*, (Vol. 3, pp.107-127). Boca Raton, FL: CRC Press.

Meijer, R. R., & Sijtsma, K. (2001). Methodology Review: Evaluating person fit. *Applied Psychological Measurement*, 25(2), 107-135. doi:10.1177/01466210122031957

R Statistics Rasch Packages: A Survey

In the course of writing a chapter for Cano, *et al.* (2020), I investigated the R Statistics packages, shown in Table 1, that implement Rasch models. Winsteps and Facets (my own software) are included for comparison. Using the "Knox Cube Test" dichotomous and "Liking for Science" polytomous datasets and others, item estimates were obtained from the packages. Findings:

1. R Statistics packages are free, but a usual consequence is that the documentation is somewhat skimpy and support almost non-existent.
2. These R Statistics packages were easy to use after the first time. The same .rdata dataset could be used with them all: column (item) labels, no row labels, no maximum or minimum possible item or person scores, scored responses start at zero, no unobserved intermediate categories†. The same R command, "summary(…)", reports the item estimates for them all. The most challenging part is to find the correct estimation command in the package's documentation. To help with this, a typical command for each package is shown in Table 1.
3. Each set of item estimates had its own logit scaling. However most sets of estimates, including those from Winsteps and Facets, could be linearly transformed into the same scale.
4. The additional capabilities of each package vary widely. *Advice:* check that the package provides whatever other output you require, such as person estimates *(thetas)* and fit statistics.
5. The sets of PMLE item estimates from "pairwise" and "sirt" differed noticeably from the other sets of estimates and from each other. Accordingly, I implemented Bruce Choppin's (1985) PMLE, producing a third set of estimates. Figure 1 plots the PMLE estimates against the "eRm" CMLE estimates. It shows the discord. After adjusting the mean item difficulties to zero, all 3 sets of PMLE estimates have trendlines close to the identity line with the CMLE estimates. However, the vertical spread of the PMLE estimates is wider than the

Table 1. *Software Implementing the Rasch Model.*

R Statistics Package	Year of PDF	Models	Item Estimation	Typical R command
eRm	2020	DRM, RSM, PCM, Others	CMLE	res <- RM(dataset)
ltm	2018	DRM, PCM, Others	MMLE	res <- rasch(data = dataset)
mixRasch	2015	DRM, RSM, PCM, Others	JMLE	res <- mixRasch(dataset, …)
pairwise	2020	DRM, PCM	PMLE	res <- pair(daten=dataset)
pcIRT	2019	DRM, Others	CMLE	res <- DRM(dataset)
RM.weights	2015	DRM, PCM, Others	CMLE	res <- RM.w(dataset, …)
sirt	2020	DRM, PCM, Others	PMLE	res <- rasch.pairwise(dataset)
TAM	2020	DRM, RSM, PCM, Others	JMLE, MMLE	res <- tam.jml(dataset)
Other Software				
Winsteps	2020	DRM, RSM, PCM, Others	JMLE	
Facets	2020	DRM, RSM, PCM, Others	JMLE	

Note: In the model column, DRM = Dichotomous Rasch Model, RSM = Rating Scale Model, PCM = Partial Credit Model. For the item estimation column, CMLE = Conditional Maximum Likelihood Estimation, JMLE = Joint Maximum Likelihood Estimation, MMLE = Marginal Maximum Likelihood Estimation, PMLE = Pairwise Maximum Likelihood Estimation.

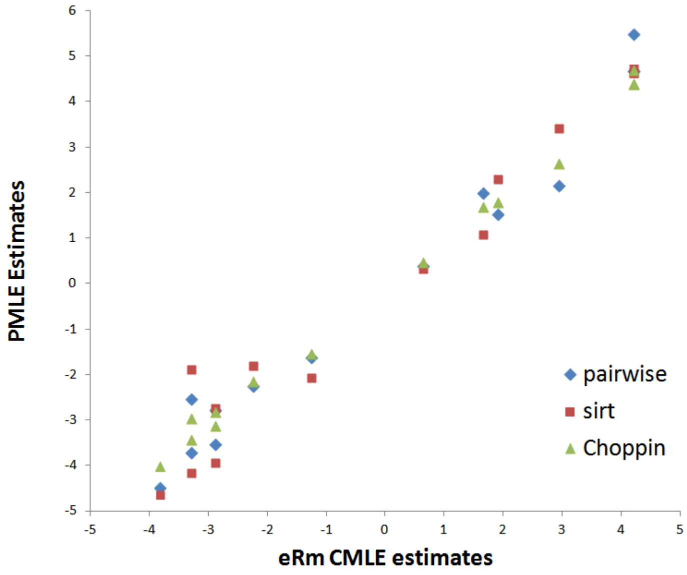


Figure 1. Scatterplot of PMLE item estimates against CMLE estimates for the Knox Cube Test dataset.

spread caused by the influence of response patterns on PMLE estimates. Choppin's method plots closest to the CMLE estimates. *Advice:* Figure 1 suggests that the "pairwise" and "sirt" PMLE estimates are idiosyncratic.

6. The R packages proved unreliable on occasion. With standard datasets, some crashed. Others produced noticeably incorrect estimates. These failures are not reported here in detail because I have emailed the package authors. Hopefully the problems are remedied by the time you read this. Packages with more recent PDF documents are more likely to have developers who will respond to bug reports and feedback. *Advice:* Obtain item estimates from at least two packages and cross-plot them. Investigate off-diagonal estimates to decide which set of estimates is more reasonable for your data. If you are planning a series of Rasch analyses using R, always have two packages available for any analysis.

John Michael Linacre
mike@winsteps.com

References

Cano S., Marquis P., Regnault A., Fisher W.P. Jr. (2020, forthcoming). Person-Centered Outcome Metrology. Cham: Springer.

Choppin, B. (1985). A fully conditional estimation procedure for Rasch model parameters. *Evaluation in Education, 9(1),* 29-42.

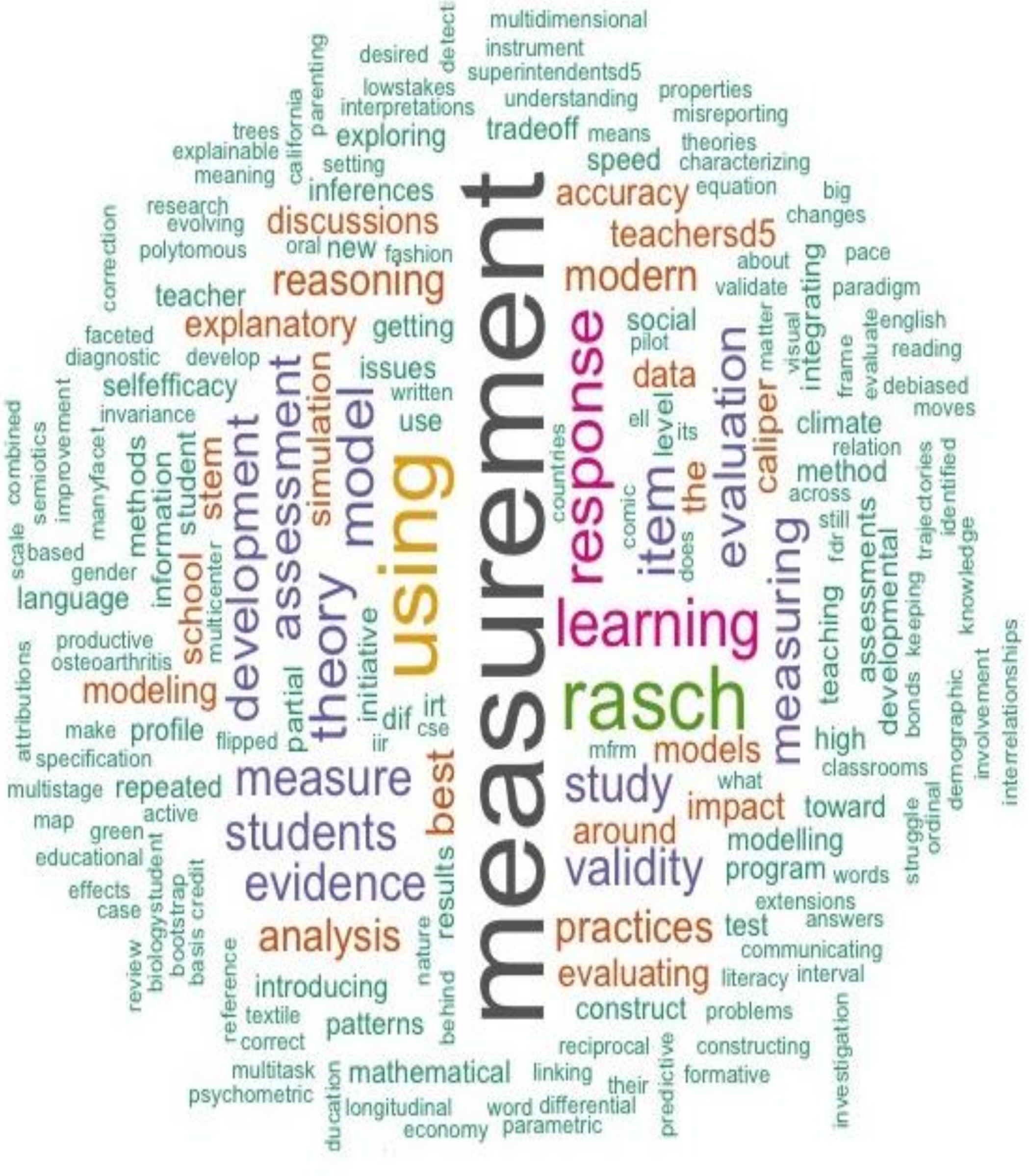
† Winsteps can output this type of .rdata file by means of its Output Files menu, IPMATRIX=, "Response value after recounting from zero" with STKEEP=No.

Reflections on the 2020 Virtual IOMW 2020 Conference

The Virtual IOMW 2020 Conference (International Objective Measurement Workshop) was held from February 4 to 6, 2021, along with two workshops (one on confirmatory mixture Rasch models and another on the BEAR Assessment System Software-BASS). We were very fortunate to have Luca Mari from Carlo Cattaneo University LIUC in Italy and Neal Kingston from the University of Kansas as our keynote speakers. In his talk entitled "Measurement, computation, simulation, etc. Is there still a difference in the 'big data' age?", Luca scrutinized measurement through the lens of the philosophy of measurement, while Neal shared with us his work on learning maps: "Measurement issues associated with learning map assessments”.

We had 56 presentations (5 spotlight talks, 20 podiums, 14 roundtables and 17 posters). The recording of these presentations and slides will be made available at our website at: https://www.iomw.org. The word cloud below, based on the presentation titles, illustrates the topics discussed at the conference. If we mix up the top 3 most frequent words, we get “Measurement using Rasch”. No surprise!

It was truly an international conference with participants joining from 15 countries in different time zones. Some presented their work at wee hours (even at 3am their local time, such dedication!).



Taking advantage of the virtual setting, we had a special IOMW edition of Jeopardy during the happy hour. We did a sing-along with Coldplay's "The Scientist" at the closing session. We thank all the participants who made the gathering possible and special.

We are exploring an opportunity to publish some of the work presented at the conference as an edited volume. Our next conference will be in 2023 (hopefully in-person or hybrid), and we are surveying our members about what they would like (the location, the timing, etc.). If you are interested in getting involved in the next IOMW, please contact Mark Wilson at MarkW@berkeley.edu.

Perman Gochyyvev, Veronica Santelices, Yukie Toyama, Mark Wilson

Rasch Measurement SIG Announcements

SIG Business Meeting Speaker

Dr. Carol Myford (pictured to the right), Emerita Professor at the University of Illinois at Chicago, was the winner of the Benjamin Drake Wright Senior Scholar Award for 2020. She will be the speaker at the AERA 2021 Rasch Measurement SIG Business meeting. Her talk is "How Has Training in Rasch Measurement Evolved over the Years, and What Might It Look Like in the Future? (With Sincere Apologies to Marty McFly (Michael J. Fox) and Doc Brown (Christopher Lloyd))". The abstract for her talk is below.

Abstract for talk: Please join me in my snazzy virtual DeLorean for a wild ride as we time travel back to the 1980s to visit Rasch measurement training. With the aid of 1.21 gigawatts of power that will make it possible for us to travel 88 miles per hour, we will activate the flux capacitor to scream forward to the present to consider what training looks like today. Finally, barring any pesky starter problems or other technical glitches that might hinder our time travel, we will humbly contemplate what training might look like in the future.

2022 SIG Program Co-chairs

Dr. Audrey Roberts, an Assistant Professor in the School of Educational Foundations at Bowling Green State University, will join Dr. Manqian (Mancy) Liao, a psychometrician at Duolingo, as program co-chairs for the 2022 SIG Program.

Jue Wang, SIG Chair

Rasch-Related Events and Presentations at the 2021 American Educational Research Association (AERA) Conference

In this section, we have included a list of Rasch-related events and presentations that are scheduled during the upcoming 2021 American Educational Research Association (AERA) conference.

Rasch Measurement SIG Business Meeting:

- *Time:* Saturday, April 10, 6:15 to 8:15 p.m. EDT
- *Speaker: Dr. Carol Myford*
- *Title: How Has Training in Rasch Measurement Evolved over the Years, and What Might It Look Like in the Future? (With Sincere Apologies to Marty McFly (Michael J. Fox) and Doc Brown (Christopher Lloyd))*

Rasch Measurement SIG Sessions:

- Rasch Modeling Methodologies
    - *Time:* Friday, April 9, 4:10 to 5:40 p.m. EDT
    - *Papers:*
        - Does sparseness matter? Comparing Generalizability Theory and many-facet Rasch measurement in sparse rating designs – *Stefanie A. Wing, The University of Alabama – Tuscaloosa; Eli Andrew Jones, The University of Memphis; Sara Bernice Grajeda, University of Delaware*
        - Exploring the impact of missing data on residual-based dimensionality analysis for measurement models – *Stefanie A. Wind, The University of Alabama – Tuscaloosa; Randall E. Schumacker, The University of Alabama*
        - Rasch/Guttman-Based Scenario (RGS) Scales: Development and benefits – *Larry H. Ludlow, Boston College; Katherine Ann Reynolds, Boston College; Maria Eugenia Baez Cruz, Boston College; Wen-Chia Claire Change, University of Nebraska – Lincoln*
        - Using Rasch analysis for determining the cut score of a computer science placement exam – *Steven McGee, The Learning Partnership; Everett V. Smith, EVS Psychometric Services, LLC; Andrew Rasmussen, Chicago Public Schools; Jeremy Gubman, Chicago Public Schools*

- Rasch Modeling in Educational Settings
    - *Time:* Saturday, April 10, 10:40 to 12:10 p.m. EDT
    - *Papers:*
        - Examining raters effects using the many-facet Rasch Model in mathematics teacher classroom performance assessment – *Ianrong Ii,*

Florida State University; Robert Schoen, Florida State University; Insu Paek, Florida State University

- Measuring office environmental satisfaction among Saudi faculty members: A Rasch analysis – *Ahlam Alghamdi, Kent State University*
- Self-efficacy scale of Confucius Institute Chinese teachers: A Rasch-based measurement instrument development – *Huadong Yin, Capital Normal University; Ren Liu, University at Buffalo – SUNY; Xiufeng Lie, University at Buffalo – SUNY*
- The Smartphone Addiction Scale for Children—short form: Bifactor modeling and Rasch analysis – *Ilker Soyturk, Kent State University; Busra Basak Ozyurt Soyturk, Marmara University*
- Treatment outcome measurement using the Homework Problems Checklist for high school students with attention deficit hyperactivity disorder – *Qiong Fu, Lehigh University; George J DuPaul, Lehigh University; Steven W Evans, Ohio University*

- Rasch Modeling Applications
  - *Time:* Sunday, April 11, 9:30 to 10:30 a.m. EDT
  - *Papers:*
- Questionable cultural comparisons: A Rasch analysis of the Teacher Self-Efficacy Subscale of the Teaching and Learning International Survey – *Janine Alisha Jackson, Morgan State University*
- Using many-facet Rasch measurement to investigate construct-irrelevant variance for contextualized constructed response assessment – *Xiaoming Zhai, University of Georgia; Kevin Haudek, Michigan State University; Christopher D. Wilson, Biological Sciences Curriculum Study; Molly A.M. Stuhlsatz, BSCS Science Learning*
- Using Rasch measurement to examine the psychometric properties of a listening test used for immigration – *Angel Arias, University of Ottawa; Jean-Guy Blais, University of Montreal*
- Validating educational assessments: The Callysto Computational Thinking Test – *Connie Yuen, University of Alberta; Chang Lu, University of Alberta; Florence A. Glanfield, University of Alberta; Maria Cutumisu, University of Alberta; Catherine Adams, University of Albert*

Other Rasch-related Sessions:

- Rasch Modeling: Methodology and Application (Division D Roundtable)
  - *Time:* Saturday, April 10, 2:30 to 3:30 p.m. EDT
  - *Papers:*
- Validity and test length reduction strategies for complex assessments – *Lance Kruse, North Carolina State University; Gregory Stone; Toni A. Sondergeld, Drexel University; Jonathan David Bostic, Bowling Green State University*
- Comparing maximum likelihood estimation methods in the Rasch Model with sample sizes and test lengths – *Jiaqi Zhang, University of Cincinnati*
- An application of many-facet Rasch Model approach to assessing

creativity in science – *Haiying Long, The University of Kansas; Jue Wang, University of Miami*
    - Using Rasch to explore students' understanding of energy: A modeling-based intervention study – *Ayca Karasahinoglu Fackler, The University of Georgia; Daniel K Capps, University of Georgia; Johnathan Todd Shemwell, The University of Alabama*
    - Synthesis of articles in the Journal of Applied Measurement: 2000-2019 – *Cheng Hua, The University of Alabama*

- Practical Guidance and Best Practices for Gathering Validity Evidence Based on Assessment Type (Division D Working group roundtable)
    - *Time:* Sunday, April 11, 10:40 a.m. to 12:10 p.m. EDT
    - *Papers:*
    - Validation evidence for forced-choice and mixed-format knowledge assessments – *Cai F. Herrmann-Abell, BSCS Science Learning*
    - Validity evidence for Likert/rating scale instruments – *Leigh M. Harrell-Williams, The University of Memphis*
    - Validity evidence for rater-mediated performance assessments – *Stefanie A. Wind, The University of Alabama – Tuscaloosa*
    - Validation evidence for observation protocols – *Eli Andrew Jones, The University of Memphis*

- Computer and Internet Applications in Education Poster Session
    - *Time:* Thursday, April 8, 3:00-4:00 p.m. EDT
    - *Papers:*
    - Assessment of information and communication technology in tertiary education applying structural equation modeling and Rasch model – *A.Y.M. Atiquil Islam, East China Normal University*

- Research on Evaluation Session (Paper session)
    - *Time:* Monday, April 12, 11:10 a.m. to 12:40 p.m. EDT
    - *Papers:*
    - Using Rasch measurement theory for responsive program evaluation – *Albert Anthony Clairmont, University of California – Santa Barbara; Daniel Katz, University of California – Santa Barbara; Milk Wilton, University of California – Santa Barbara*

- Classroom Observation SIG Roundtable
    - *Time:* Thursday, April 8, 2:00 to 3:00 p.m. EDT
    - *Papers:*
    - Are teacher candidate ratings reliable? What many-facet Rasch measurement says about preservice teacher supervisor ratings – *Eli Andrew Jones, The University of Memphis; Stefanie A. Wind, The University of Alabama – Tuscaloosa; Anna Hart, Columbus State University; Jan Burcham, Columbus State University; Thomas Dailey, Columbus State University*

- Consequences of Testing and Rasch-Based Differential Item Functioning Analysis in Validation Research (Test

Validity Research and Evaluation Poster Session)

  - *Time:* Friday, April 9, 9:30 to 10:30 a.m. EDT
  - *Papers:*
    - ▪ Demographic interactions of differential item functioning in attitudinal scales – *Nana Amma Asamoah, University of Arkansas; Ronna C. Turner, University of Arkansas; Wen-Juo Lo, University of Arkansas; Brandon Crawford, Indiana University – Bloomington; Kristen Jozkowski, Indiana University; Sara McClelland, University of Michigan*

- Item Response Theory with Complex Data (Division D Paper Session)
  - *Time:* Sunday, April 11, 4:10 to 5:40 p.m. EDT
  - *Papers:*
    - ▪ Investigating latent class characteristics with covariates by using mixture Rasch model – *Selay Zor, University of Georgia; Allan S. Cohen, University of Georgia; Brian A. Bottge, University of Kentucky; Linda Joy Gassaway, University of Kentucky*

- Novel Approaches for Model Fit (Division D Paper Session)
  - *Time:* Saturday, April 10, 4:10 to 5:40 p.m. EDT
  - *Papers:*
    - ▪ A mixture Rasch facets model for rater's illusory halo effects – *Kuan-Yu Jin, Hong Kong Examinations and Assessment Authority; Ming Ming Chiu, The Education University of Hong Kong*

- Pre-K—8 Focused Science Teaching and Learning (Division C Paper Session)
  - *Time:* Sunday, April 11, 9:30 to 10:30 a.m. EDT
  - *Papers:*
    - ▪ Development of an instrument to evaluate junior school students' chemical context-based thinking skill – *Shaohui Chi, East China Normal University; Zuhao Wang, East China Normal University; Weilei Quan*
    - ▪ Research on chemical information processing capability of junior high school students – Yongchao Fu, East China Normal University; Zuhao Wang, East China Normal University; Shaohui Chi, East China Normal University

- Topics in Contemporary Program Implementation, Evaluation, and Measurement (Division H Poster Session)
  - *Time:* Monday, April 12, 11:10 a.m. to 12:10 p.m. EDT
  - *Papers:*
    - ▪ Building a Culturally Responsive Instruction Scale from an Observation Protocol (CRIOP) – *Shannon O. Sampson, University of Kentucky; Katherine Leung Robershaw, University of Kentucky; Susan Cantrell, University of Kentucky*

- Exploring Student and Teacher Cognition in Mathematics Education (Research in Mathematics Education SIG Paper Session)
  - *Time:* Saturday, April 10, 4:10 to 5:40 p.m. EDT
  - *Papers:*

    - Changes in upper elementary students' early algebra knowledge sophistication: Results from a computer game-based intervention – *Christopher Engledowl, New Mexico State University; Mohammad Saleh Al-younes, New Mexico State University; Barbara Chamberlin, New Mexico State University*

- Future Directions of Educational Measurement in International Large-Scale Assessments (Large Scale Assessment SIG Paper Session)
    - *Time:* Saturday, April 10, 10:40 a.m. to 12:10 p.m. EDT
    - *Papers:*
    - The effects of incorrect answer substitution in PIRLS – *Andrés Christiansen, Katholieke Universiteit Leuven; Rianne Janssen, KU Leuven*

- Identity as an Outcome and Unfolding Process in Science Education (Division C Roundtable Session)
    - *Time:* Saturday, April 10, 4:10 to 5:40 p.m. EDT
    - *Papers:*
    - The construct of researcher identity for secondary school students – *Linda Morell, University of California – Berkeley; Shruti Bathia, University of California – Berkeley; Ben Koo, University of California – San Francisco; Mark R. Wilson, University of California – Berkeley; Perman Gochyyev, University of California, Berkeley; Rebecca Smith, University of California – San Francisco*

- Imagining and Examining STEM Practice and Preparation (Division K Paper Session)
    - *Time:* Saturday, April 10, 4:10 to 5:40 p.m. EDT
    - *Papers:*
    - Examining elementary preserve teachers' emotions about teaching science and mathematics – Mihwa Park, Texas Tech University; Raymond Flores, Texas Tech University

- Learning, Community, and Relations (Division C Poster Session)
    - *Time: Monday, April 12, 11:10* a.m. to 12:10 p.m. EDT
    - *Papers:*
    - Examining the effects of a peer-learning research community on the development of research identity – *Ben Koo, University of California – San Francisco; Shruti Bathia, University of California – Berkeley; Linda Morell, University of California – Berkeley; Perman Gochyyev, University of California, Berkeley; Mark R. Wilson, University of California – Berkeley; Rebecca Smith, University of California – San Francisco*

- Measurement and Assessment in Higher Ed Paper Session
    - *Time: Friday, April 9, 10:40* a.m. to 12:10 p.m. EDT
    - *Papers:*
    - The Boston College Living a Life of Meaning and Purpose (BC-LAMP) portfolio: A reexpression and extension of the Claremont Purpose Scale – *Larry H. Ludlow, Boston*

College; Theresa O'Keefe, Boston College; Olivia Szendey, Boston College; Ella Anghel, Boston College; Henry I. Braun, Boston College; Burt Howell, Boston College; Christina Matz-Costa, Boston College

- Preservice Courses, Student Teaching Experiences, and Beginning Teacher Outcomes (Division L Paper Session)
    - *Time:* Saturday, April 10, 4:10 to 5:40 p.m. EDT
    - *Papers:*
  - How do we know if new teachers are prepared? Considering different predictors of instructional readiness – *Kavita Kapadia Matsko, Northwestern University; Matthew Ronfeldt, University of Michigan*

- Professional Ethics for Future Teachers: Fostering Educational Responsibility
    - *Time:* Saturday, April 10, 2:30 to 4:00 p.m. EDT
    - *Papers:*
  - Adaptation, Piloting, and Validation of a Test of Ethical Sensitivity in Teaching – *Bruce Maxwell, University of Montreal*

- Text Comprehension: What We Know About the Dance Between Reader, Text, and Task in Reading Comprehension (Division C Paper Session)
    - *Time:* Monday, April 12, 2:50 to 4:20 p.m. EDT
    - *Papers:*
  - Measuring learning in a full-book automated reading tutor: The interaction of reader and text characteristics – *Kathleen M. Sheehan, Retired*

- The Impact of the Learning Environment on Student and Educator Outcomes (Division L Poster Session)
    - *Time:* Thursday, April 8, 5:00 to 6:00 p.m. EDT
    - *Papers:*
  - Classroom emotional climate: Its assessment and associations with student attitudes – *Felicity McLure, Curtin University; Barry J. Fraser, Curtin University; Rekha Bhan Koul, Curtin University*

- Toward an Equity-Minded Pedagogy and Praxis (Division J Paper Session)
    - *Time:* Monday, April 12, 2:50 to 4:20 p.m. EDT
    - *Papers:*
  - Examining an adaptive equity-oriented pedagogical competency instrument: A validated measure that promotes college student success – Andrew Estrada Phuong, *University of California – Berkeley; Judy Nguyen, Stanford University; Christopher Todd Hunn, University of California – Berkeley; Fabrizio Daniel Mejia, Berkeley University of California*

- When School and Community Climate Intercede in the Educational Process (Stress, Coping, and Resilience SIG Paper Session)
    - *Time:* Saturday, April 10, 2:30 to 4:00 p.m. EDT
    - *Papers:*
  - Identifying cultural differences in stress-related measures in German and Singaporean social work students – *Richard G. Lambert, University of North Carolina – Charlotte; Andrea*

D. Schwanzer, Hannover University of Applied Sciences; Annette Ullrich, Duale Hochschule Baden-Württemberg Stuttgart; Boon Kheng Seng, Singapore University of Social Sciences

List of Recent Publications in Journal of Applied Measurement

Vol. 21, No. 4, Winter 2020

Rasch/Guttman Scenario (RGS) Scales: A Methodological Framework
Larry H. Ludlow, Maria Baez-Cruz, Wen-Chia Claire Chang, and Katherine A. Reynolds

The Effect of Interactions between Item Discrimination and Item Difficulty on Fit Statistics
Stephen Mark Humphry and Ken Bredemeyer

Comparing Causes of Dependency: Shared Latent Trait or Dependence on Observed Response
Christine E. DeMars

Using the Rasch Model to Measure Comprehension of Fraction Addition
Marius Lie Winger, Julie Gausen, Eivind Kaspersen, and Trygve Solstad

The Development of the Mental Toughness Situational Judgment Test: A Novel Approach to Assessing Mental Toughness
Nicholas M. Flannery, Neil M. A. Hauenstein, and E. Scott Geller

A Psychometric Replication of Fan (1998) Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics
Nicholas Marosszeky, E. Arthur Shores, Michael P. Jones, and Rassoul Sadeghi

Diabetes Distress in Emerging Adults: Refining the Problem Areas in Diabetes—Emerging Adult Version using Rasch Analysis
Katherine Wentzell, Judith A. Vessey, Lori Laffel, and Larry Ludlow

Evaluating the Impact of Multidimensionality on Type I and Type II Error Rates using the Q-Index Item Fit Statistic for the Rasch Model
Samantha Estrada

Development of a Short Form of the CPAI-A (Form B) with Rasch Analyses
Yixiao Dong, Weiqiao Fan, Fanny M. Cheung, and Mengting Li