Spring 2016, VOL. 29, NO. 4; ISSN 1051-0796

RMT

RASCH MEASUREMENT TRANSACTIONS

Research Notes and Editorials Include:

Transactions of the Rasch Measurement SIG
American Educational Research Association

Is JMLE Really that Bad? No, it's Actually Rather Good!

Mesbah (2012) reports that "The table [18.1] illustrates the advantages of MML and CML estimates compared to JML estimates because the confidence interval of the JML estimates is ten times larger than the confidence of the other estimates. Furthermore, the implementation in PROC LOGISTIC as shown here, is quite time consuming."

In Table 18.1, NLMIXED, GENMOD, GLIMMIX and LOGISTIC are the SAS procedures used for estimation. Est. is the average of the item estimates for each method. We expect this to approximate the True value. Lower is the lower confidence interval, 1.96 standard deviations below the Est. Upper is the upper confidence interval, 1.96 standard deviations above the Est.

Let's assume that the MML: NLMIXED signs were accidentally reversed in Table18.1. Then reformat Table 18.1 into Table 2.

Diff. is the difference between the Est. mean estimate and the True value. It does not exceed 0.11 except for JML: LOGISTIC. S.D. is the standard deviation of the estimates. It does not exceed 0.09 except for JML:LOGISTIC.

These results appear to be satisfactory for all methods except JML:LOGISTIC where the Est. (mean of the estimates) is too central, but the S.D. of the estimates is large.

Is JMLE really that bad?

Mesbah (2012) provides details of the data simulations and estimation procedures underlying Table 18.1. I replicated his method using the capabilities of Excel and Winsteps. 100 dichotomous Rasch datasets were simulated each based on a newly generated set of 100 abilities, distributed N(0,1) logits. 10 item difficulties were set at the 5%, 15%,… 85%, 95% quantiles of the unit-normal distribution. Item estimates were obtained very quickly for the 100 datasets using the batch-mode capabilities of Winsteps. The resulting estimates were summarized by Excel into this Table 3.

We can see that the S.D.s for JMLE: Winsteps are similar to those of the other estimation methods. "Theory S.D." is the expected value of S.D. according to Rasch theory for JMLE. It is also similar. Now all S.D. values for all methods and Rasch theory are in the range 0.07 - 0.10.

In Table 3, the JMLE Est. values are somewhat less central than the True values. This is the expected, but usually inconsequential, estimation bias when there are few items. In the right-most columns of Table 3 the Wright & Douglas (1977) bias adjustment has been applied to the JMLE estimates. For 10-item estimation, the adjustment is to multiply the estimates by (10-1)/10 = 0.9. After applying the bias adjustment, the JMLE Est. values are now closer to the True values than for any other estimation method!

Conclusion: The investigations of 50 years of Rasch estimation are confirmed. All these estimation methods (CMLE, MMLE, JMLE) are equally good for practical work when implemented proficiently. However, their capabilities do differ considerably in other areas, such as the analysis of datasets with missing data.

A better conclusion for Mesbah (2012) would have been: "Don't use this PROC LOGISTIC method for estimating JMLE."

John Michael Linacre

References

Linacre, J. M. (2015). Winsteps® 3.91.0 Rasch measurement computer program. Beaverton, Oregon: Winsteps.com

Mesbah M. (2012). Software for Rasch Analysis. Chapter 18 in *Rasch Models in Health.* K. B. Christensen, S. Kreiner, M. Mesbah, Eds., John Wiley & Sons Inc.

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

Table 18.1: Comparison of item parameter estimates obtained using four different methods in Mesbah (2012).

		MML: NLMIXED			CML: GENMOD			MML: GLIMMIX			JML: LOGISTIC		
Item	True value	Est.	lower	upper	Est.	lower	upper	Est.	lower	upper	Est.	lower	upper
β1	-1.65	1.6	1.42	1.78	-1.64	-1.79	-1.48	-1.6	-1.78	-1.42	-0.93	-2.34	0.47
β2	-1.04	0.99	0.83	1.15	-1.02	-1.16	-0.88	-0.99	-1.15	-0.83	-0.58	-1.99	0.82
β3	-0.67	0.75	0.59	0.9	-0.78	-0.92	-0.64	-0.75	-0.9	-0.59	-0.45	-1.85	0.96
β4	-0.39	0.32	0.17	0.47	-0.35	-0.49	-0.22	-0.32	-0.47	-0.17	-0.21	-1.61	1.2
β5	-0.13	0.03	-0.12	0.18	-0.07	-0.2	0.06	-0.03	-0.18	0.12	-0.05	-1.45	1.36
β6	0.13	-0.23	-0.38	-0.08	0.2	0.06	0.33	0.23	0.08	0.38	0.1	-1.3	1.5
β7	0.39	-0.39	-0.54	-0.23	0.35	0.22	0.49	0.39	0.23	0.54	0.19	-1.22	1.59
β8	0.67	-0.65	-0.8	-0.49	0.62	0.48	0.75	0.65	0.49	0.8	0.34	-1.07	1.74
β9	1.04	-1.06	-1.22	-0.9	1.03	0.88	1.17	1.06	0.9	1.22	0.57	-0.84	1.97
β10	1.65	-1.7	-1.89	-1.52	1.67	1.51	1.83	1.7	1.52	1.89	0.93	-0.47	2.34

Table 2. Table 18.1 simplified and corrected.

		MML: NLMIXED			CML: GENMOD			MML: GLIMMIX			JML: LOGISTIC		
Item	True value	Est.	Diff.	S.D.	Est.	Diff.	S.D.	Est.	Diff.	S.D.	Est.	Diff	S.D.
β1	-1.65	-1.60	0.05	0.09	-1.64	0.01	0.08	-1.60	0.05	0.09	-0.93	0.72	0.72
β2	-1.04	-0.99	0.05	0.08	-1.02	0.02	0.07	-0.99	0.05	0.08	-0.58	0.46	0.72
β3	-0.67	-0.75	-0.08	0.08	-0.78	-0.11	0.07	-0.75	-0.08	0.08	-0.45	0.22	0.72
β4	-0.39	-0.32	0.07	0.08	-0.35	0.04	0.07	-0.32	0.07	0.08	-0.21	0.18	0.72
β5	-0.13	-0.03	0.10	0.08	-0.07	0.06	0.07	-0.03	0.10	0.08	-0.05	0.08	0.72
β6	0.13	0.23	0.10	0.08	0.20	0.07	0.07	0.23	0.10	0.08	0.10	-0.03	0.71
β7	0.39	0.39	0.00	0.08	0.35	-0.04	0.07	0.39	0.00	0.08	0.19	-0.20	0.72
β8	0.67	0.65	-0.02	0.08	0.62	-0.05	0.07	0.65	-0.02	0.08	0.34	-0.33	0.72
β9	1.04	1.06	0.02	0.08	1.03	-0.01	0.07	1.06	0.02	0.08	0.57	-0.47	0.72
β10	1.65	1.70	0.05	0.09	1.67	0.02	0.08	1.70	0.05	0.09	0.93	-0.72	0.72

Table 3. Summary of simulation and analysis of 100 datasets equivalent to those used in Table 18.1

		JMLE: Winsteps				Wright-Douglas (1977)	
Item	True value	Est.	Diff.	S.D.	Theory S.D.	Bias-adjusted JMLE Est.	Diff.
1	-1.64	-1.86	-0.22	0.09	0.09	-1.68	-0.03
2	-1.04	-1.16	-0.13	0.08	0.08	-1.05	-0.01
3	-0.67	-0.75	-0.08	0.07	0.08	-0.68	0.00
4	-0.39	-0.44	-0.05	0.07	0.07	-0.39	-0.01
5	-0.13	-0.15	-0.02	0.07	0.07	-0.13	-0.01
6	0.13	0.14	0.01	0.07	0.07	0.13	0.00
7	0.39	0.44	0.05	0.08	0.07	0.39	0.01
8	0.67	0.76	0.09	0.07	0.08	0.69	0.01
9	1.04	1.17	0.14	0.09	0.08	1.05	0.02
10	1.64	1.85	0.21	0.10	0.09	1.67	0.02

Note from Rasch SIG Chair

Greetings Rasch SIG colleagues:

This year's AERA conference is fast approaching, and I would like to take this opportunity to invite you to the events that have been scheduled by our Program Co-Chairs Sara Hennings and Liru Zhang.

The 2016 program is once again full of great papers dealing with aspects of Rasch Theory. I hope that you will "rock up" to support your fellow Rasch researchers at this year's conference.

I am also very pleased to announce that our Keynote Speaker at this year's business meeting is Elena Kardanova. This is a landmark moment for the Rasch community because it is the first time that we have had a female keynote speaker at our meeting. Elena is the Director of the Center for Monitoring the Quality in Education at the National Research University Higher School of Economics (Moscow, Russia). Elena's topic is as follows: Applying the Rasch Model to Assess Cross-Cultural Comparability of Test Results.

In addition to Elena's address we will be announcing the winner of the Benjamin Drake Wright Senior Scholar Award. The Benjamin Drake Wright Senior Scholar Award is an AERA-sanctioned award. It is being presented for the first time in 2016 and is being awarded to an individual senior scholar for outstanding programmatic research and mentoring in Rasch measurement over the course of a career and who is still active in Rasch measurement research at the time the award is granted. This award represents the culmination of a lot of work by the current and past SIG committees to acknowledge the pioneering efforts of our leaders in the Rasch field. We received nominations for a number of truly outstanding candidates, so the inaugural winner will be announced at the meeting.

After the award presentation, there will be a brief summary of how our SIG is progressing. The meeting is scheduled for April 11 from 6:15pm to7:45pm in the Marriott Marquis Level Two Marquis Salon 1. Hors d'oeuvres and a cash bar will be provided.

I will send out more detailed information on all presentations and logistics prior to the AERA conference. Enclosed at the end of this issue will also be a list of known Rasch-related presentations appearing on the AERA program at the time of this writing.

This is an exciting time for the Rasch SIG. I look forward to the AERA conference in Washington DC and hope to see you there.

Jim Tognolini
Rasch Measurement SIG Chair

"Of course you still weigh 250 pounds! That's as high as the scale goes."

Rasch Measurement Transactions
www.rasch.org/rmt
Editor: Kenneth Royal
Email Submissions to: Editor \at/ Rasch.org

RMT Editor Emeritus: John M. Linacre
Rasch SIG Chair: Jim Tognolini
Secretary: Mikaela Raddatz
Treasurer: Leigh Williams
Program Chairs: Sara Hennings & Liru Zhang
Rasch SIG website: www.raschsig.org

Association of Rasch Person Reliability with Cronbach's Alpha, Ferguson's Delta and Gini Coefficients

The scale reliability is not an index of quality (i.e., a good measure), but of relative reproducibility (i.e., a repeatable measure) (Linacre, 1997). KR-20 (Cronbach's alpha) always exceeds the maximum reliability possible (and the Rasch person reliability) for the measures underlying in a study (Linacre,1997). However, we have not seen any simulation study exploring the association of both scale indices across different types of person distribution and item length and the relationship to the equality of the person scores.

Ferguson's Delta (Ferguson,1949) and Gini (1909) coefficients are both famous indices of evaluation the equality of data frequency. According to the KR-20 (Cronbach's alpha) calculation and the Rasch model person reliability, we defined the latter as: $R_p = SA_p^2 / SD_p^2 = 1-(MSE_p / SD_p^2)$, where $MSE_p = \sum_{n=1}^{n} SE_n^2 / N$, and the true variance in the person measures: $SA_p^2 = SD_p^2 = MSE_p$, the observed variance among persons, SD_p^2, the model squared standard error (SE) for each person measure: SE_n^2, the sample size: N (Schumacker & Smith, 2007). The real person reliability is adjusted for the SE replaced by Real SE (= Model SE * Maximum [1.0, sqrt (INFIT mean-square)] (Linacre, 1997). We can see that the denominator in the reliability equation is regarding the person distribution, the more tendency toward uniform, the higher equality indices will be.

We are thus interested in exploring the association of those indices when a sample with three types of distribution of dichotomous and polytomous 5-point rating scales across three number of item length (i.e., 10,20,30) fits a Rasch (1960) model using a simulation approach (Linacre, 2007). All the items, uniformly distributed from -2 to +2 logits, were simulated on hypothetical samples of 200 subjects with abilities distributed (Figure 1). This was repeated 100 times and their medians of the aforementioned indices reported.

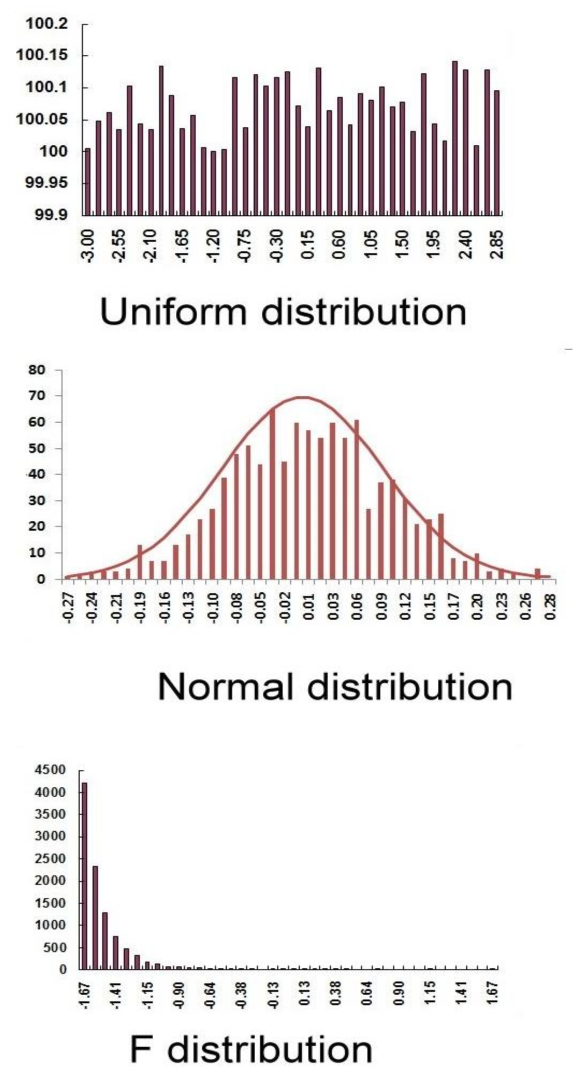


Figure 1. Three types of sample distribution in the study

We found that (1) the equality coefficients (i.e., Delta and 1-Gini limited to five bins of person measure frequency) (Chien & Djaja, 214) are stably and horizontally flattened across scenarios (e.g., DU, GU, DN, DF, GN, GF from high to low, see Figure 2). Others are increased from the left-bottom to the top-right (i.e., the shorter item lengths and less thresholds, the loser values). (2) Besides the scenario of Dicho-10, Rasch person real reliability can be similar to the the KR-20

(Cronbach's alpha) when data are normalized. With polytomous 5-point rating scales, Rasch reliability is almost equal to Cronbach Alpha. (3) The reliability is dependent of person distribution, the highest is the uniform, the normal following, and the lowest is the skewed like F distribution.

An implication is that we should further report the equality coefficient (Delta or 1-Gini) for a survey or a test to present the feature of the examinees' score distribution. It is because we can indirectly know the momentum of the test error when the reliability and the denominator in the reliability equation are known. For instance, the lower equality indices (Figure 3) impossibly cause a higher Rasch person real reliability due to many high person estimated standard errors in existence according to the formula: $R_p = SA_p^2 / SD_p^2 = 1 - (MSE_p / SD_p^2)$.

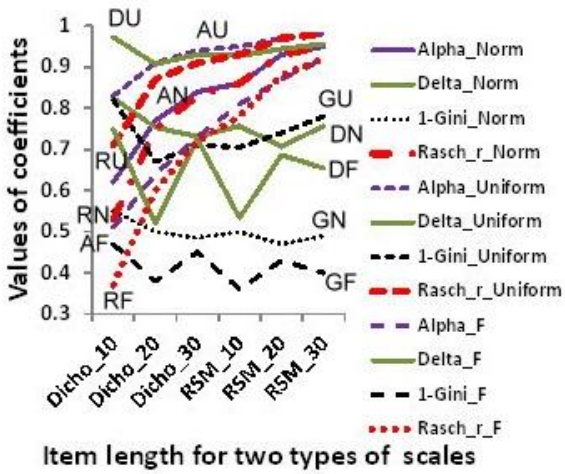


Figure 2. Rasch person reliabilities compared to the counterparts

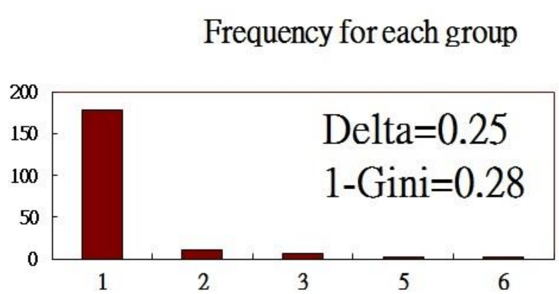


Figure 3. Calculation of Delta and 1-Gini coefficients

Tsair-Wei Chien
Chi Mei Medical Center, Taiwan

References

Chien,T.W., Djaja,N. (2014). Using Rasch Simulation Data to Verify Whether Ferguson's Delta Coefficient Can Report Students' Abilities Are Equal in a Class. *Rasch Measurement Transactions*, *28*(3),1484-5.

Ferguson, G. A. (1949). On the theory of test discrimination. *Psychometrika*, 4, 61-68.
Gini, C. (1909). Concentration and dependency ratios (in Italian). 1909 English translation in *Rivista di Politica Economica* 1997;87:769–789.

Linacre, J.M. (1997). KR-20 / Cronbach Alpha or Rasch Person Reliability: Which Tells the "Truth"? *Rasch Measurement Transactions*, *11*(3),580-1.

Linacre, J. M. (2007). How to Simulate Rasch Data. *Rasch Measurement Transactions, 21*(3), 1125.

Schumacker, R. E, & Smith,Jr. E. V. (2007). Reliability: A Rasch Perspective. *Educational and Psychological Measurement*, *67*(3),394-409.

Using Philosophy of Science to Expose Logical Fallacies in Measurement Research

Kristin Shrader-Frechette's 2014 book, *Tainted: How Philosophy of Science Can Expose Bad Science*, uses logic to expose fallacious arguments in scientific research. Most of the examples given involve the environment and pollution, but the lessons learned are easily extended to other areas, such as psychometrics (to which we will return). The book is structured to be an example of consistency and rationality, but in such a way that rigorous thinking sometimes becomes a bit rigid, and so, frustrating, in both its microscopic parsing of details (hair splitting) and in its inevitable oversights and interpretive challenges.

For instance, on page 18, Shrader-Frechette seems to present logical consistency as a methodological alternative to making value judgments, as though requiring logical consistency in arguments is not itself already involve value judgments. What she actually does is avoid *controversial* value judgments and focus on consistency as a provisional initial step in a larger argument aimed at parsing out a stronger basis for warranted inferences. Logical consistency is surely less controversial than determining costs relative to the benefits of environmental and safety regulations, but in a book that makes so many fine distinctions in its arguments, one wishes it would be possible to maintain that same consistency across, as well as within, each of them.

More broadly, Shrader-Frechette seems to ignore most of what's happened in philosophy of science in the last 50 years and more, given her strict emphases on straightforward logic, rationality, and truth. It should be possible to write a book like this, doing perhaps an even more effective job, without setting aside the philosophy of science's concerns with the inevitable influences of social, emotional, interpretative, historical, and cultural contexts.

That said, there is much to be gained from close study of the analyses Shrader-Frechette presents in *Tainted*. The opportunities provided in this book for identifying and analyzing the logical fallacies committed in scientific research should be of particular value in psychometrics, given the characterization of that field as "in the grip of some kind of thought disorder" (Michell, 1997, p. 355; 2003, 2004). As a start in that process, on pages 20-21, Shrader-Frechette offers a brief survey of several fairly common errors in reasoning:

- Deductive transitivity (A-B, B-C, so A-C) can lead to hasty generalization when reasoned inductively (all the apples I've seen are red, so apples are red).
- Appeal to authority involves taking something as true just because an authority says it is.
- Begging the question assumes the truth of a conclusion instead of presenting evidence and arguments pro and con.
- Equivocation uses a word repeatedly but with different meanings.
- Appeal to the people takes a conclusion as true on the basis of the fact that most people accept it as true.
- Appeal to ignorance takes it for granted that, if falsifying evidence has not been revealed, none will be.
- Affirming the consequent assumes that the production of a test result from a hypothesis predicting it means that the hypothesis is true and not simply not yet falsified, by only one form of evidence.

That Shrader-Frechette finds these kinds of logical confusions widespread in environmental quality and safety research, with huge associated human, social, and economic costs, is evidence of the difficulties encountered in avoiding them. The clarity of the analyses made possible by these criteria suggests how more effective critiques of psychometric methods might be conducted. Instead of asserting only the pathology of an underlying thought disorder without also offering a detailed etiology and course of preventative care, Shrader-Frechette's *Tainted* suggests something beyond the prescription of a restricted range of acceptable methods. In following through on Shrader-Frechette's start in this way, the hope emerges that reducing logical fallacies in psychometrics could be a significant factor in improving the quality of measurement research in psychology, education, and the social sciences.

One of Shrader-Frechette's main examples involves the U.S. Department of Energy's (DOE) Yucca Mountain nuclear waste repository, which, after decades of research study, with $15 billion invested, was not completed or used because of the logical fallacies underlying what turned out to be unwarranted conclusions. Shrader-Frechette (p. 24) quotes a DOE report, for instance, as saying:

No mechanisms have been identified whereby the expected tectonic processes or events could lead to unacceptable radionuclide releases. Therefore, the evidence does not support a finding that the site is not likely to meet the qualifying condition for post-closure tectonics.

In this and other parts of the report, Shrader-Frechette (p. 25) says,

DOE scientists repeatedly used appeals to ignorance instead of admitting their uncertainty and testing everything they were able to test. They could have avoided appeals to ignorance, as already noted, with weight-of-evidence or inference-to-the-best-explanation assessments. They could have used 'if..then' claims, such as: 'if our assumptions about Yucca are reliable for the centuries required, then the site would comply with regulations.'

Appeals to ignorance, affirming the consequent, and begging the question are all ways of construing Michell's primary evidence of a methodological thought disorder in psychometrics. He focuses on psychometricians' widespread failure to state and test the hypothesis that a construct is quantitative, while they simultaneously assume the truth of that hypothesis. Many psychometricians, in effect, make an appeal to ignorance, saying that no evidence contradicting the assumption of a quantitative construct has been identified, and so they feel justified in proceeding as though the construct is quantitative and the numbers in hand are measures of it. Begging the question also comes to bear here, in that quantitative status is assumed as a conclusion in the absence of evidence and argument. Affirming the consequent is also taking place in this context because results consistent with untested hypotheses about the quantitative status of the construct are interpreted as validation of those hypotheses.

Appeals to authority and to the people are also found in psychometrics, as they undoubtedly also are in other fields. One health care outcomes researcher, for instance, publicly claims to have obtained approval from a leading figure in psychometrics philosophically opposed to IRT to use it nonetheless as the methodological and conceptual basis for his work, and apparently takes that approval as a satisfactory justification for doing so. Similarly, in an overt appeal to the people, Hambleton, Swaminathan, and Rogers (1991, p. 87), say

The [IRT 2-PL, 3-PL] theta-scale, or any linear transformation of it, however, does not possess the properties of a ratio or interval scale, although it is popular and reasonable to assume that the theta-scale has equal-interval properties.

Is an appeal to the people, however "popular and reasonable" it might be, really a sufficient basis for the high stakes decisions that are based on educational admissions, graduation, licensure and certification tests? Should not assumptions concerning the ratio or interval properties of a scale instead be stated in terms of theorems, proofs, inferential requirements, hypotheses, experimental results, and warranted provisional conclusions, as is the case for a wide range of available psychometric models (Andersen, 1977, 1999; Andrich, 2010; Fischer, 1981, 1995; Newby, Connor, Grant, & Bunderson, 2009; Wright, 1984, 1997, 1999)? These are not posed as mere rhetorical questions, but as important issues that need to be resolved if psychometrics' potential for innovation and advancement is ever to be fulfilled.

Shrader-Frechette (pp. 26-27) raises a specific form of equivocation with special relevance to psychometrics. Her concern is with the ways in which the words "verify" and "validate" are used in the DOE Yucca Mountain report relative to algorithms and programs. Algorithms, being logical operations coded in a sequence, support claims that can be verified in terms of their results' production of the specified functions. Programs, however, being logical structures positing causal relationships in a model of the real world, are never verifiable. Shrader-Frechette is concerned with model testing as an end in itself, as a purely theoretical exercise with no empirical data at all involved. The situation is different in psychometrics, where the distinction between algorithmic verification of models via simulated data is clearly distinct from programmatic assessments of data from classrooms, clinics, and work places.

The distinction holds up when adapted to the psychometric context, however. Equivocation in general is rampant in psychometrics, where at least 122 different conceptions of validity and validation have emerged, with a great deal of conceptual overlap and confusion (Newton & Shaw, 2013, p. 312). But Shrader-Frechette's specific contrast of algorithmic vs. programmatic forms of verification is relevant to psychometrics in a particular way. For instance, psychometrics is often assumed to be algorithmic in the empirical, descriptive, and analytic senses of modeling data, converting ordinal scores into linear measures, and applying statistical cutoff criteria for model fit in item inclusion/exclusion determinations. Conversely, psychometrics can be considered programmatic in the senses of modeling causal relationships in the world, and assessing empirically and theoretically the practical utility of the measurement system and individual items via experimental methods (Salzberger, 2010; Stenner, Fisher, Stone, & Burdick, 2013; Wilson, 2005).

So whenever the focus of a study with real life consequences is on data modeling and on the analytic conversion of scores into measures, instead of on causal relationship modeling and the practical value obtained for managing outcomes, psychometricians equivocate and affirm the consequent. Demonstrating causal relationships and practical value cannot follow from model-fitting exercises alone, and are never conclusive, but must be demonstrated in live (in vivo) contexts, and couched in provisional, probabilistic language.

Rasch model applications thus also often equivocate and affirm the consequent in that the fit of data to a model is interpreted as supporting the conclusion that the construct is quantitative. As Rasch (1960, pp. 37-38; 1973/2011) took pains to explain, however, model fit is never perfect or conclusive, and even when multiple datasets provide provisional evidence in support of the construct's quantitative status, fallacious equivocation and affirming of the consequent continues as long as no explanatory theory in the form of a construct map (Wilson, 2005) or specification equation (Stenner, et al., 2013) is brought to bear.

Finally, in Chapter 7 of *Tainted*, Shrader-Frechette (p. 99) argues against "many US judges' and scientists' assumption that statistical-significance is necessary to hypothesize causal harm from agents like toxic chemicals." She observes (p. 107) that "Good scientific reasons, such as differences between statistical and scientific hypotheses, argue against requiring the statistical-significance rule for causal hypothesis-discovery and development." Her specific concern is that the exclusive focus on significance tests, which may be over- or under-powered, results in unjust determinations of harm or the lack of it. And quite apart from Shrader-Frechette's concerns, despite decades of commentaries and explanations seeking to clarify statistical methodology (Berkson, 1938; Coats, 1970; Cohen, 1994; Cowger, 1984), the mindless cult of researchers seeking $p < 0.05$ continues unabated, to the point that some approach it cynically, as a game to be played ruthlessly for career advancement (Bakker, van Dijk, Wicherts, 2012).

Again, Shrader-Frechette's theme is echoed in psychometrics' longstanding concern for the differences between statistical and scientific hypothesis testing (Bolles, 1962; Fisher, 2010; Meehl, 1967; Michell, 1986; Wilson, 2013). Though it is commonly assumed that mathematical models in psychometrics are probabilistic because of the need to sample from populations too large to measure in their entirety, the stochastic structures of individual-level response processes and behaviors provide another, usually overlooked, significant motivation for probabilistic modeling (Duncan, 1992; Duncan and Stenbeck 1988; Molenaar, 2004; Rogosa, 1987). Too much of psychometric and statistical practice proceeds, at its own peril, as though the latter motivation does not exist and can be safely ignored.

In conclusion, perhaps in addition to other methodological enhancements (Fisher, 2013; Tesio, Simone, Grzeda, et al., 2015), peer review criteria for research proposals and articles submitted for publication in psychometrics and across the sciences ought to include items listing the logical fallacies described by Shrader-Frechette. The practical value of philosophy for informing critical readings of the scientific literature in psychometrics and elsewhere may be further borne out as an important factor in improving the quality of research by expanding the scope of Shrader-Frechette's focus on logical consistency to include interpretive criteria (Ricoeur, 1981) and considerations informed by Actor-Network Theory (Latour, 2005). Such an expansion may enrich the search for and implementation of meaningful results in ways we quite possibly cannot imagine from the vantage point of today's science.

William P. Fisher, Jr.

References

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*(1), 69-81.

Andersen, E. B. (1999). Sufficient statistics in educational measurement. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 122-125). New York: Pergamon.

Andrich, D. (2010). Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. *Psychometrika, 75*(2), 292-308.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543-554.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *American Statistical Association Journal, 33*(201-204), 526-536.

Bolles, R. D. (1962). The differences between statistical hypotheses and scientific hypotheses. *Psychological Reports, 11*, 639-645.

Coats, W. (1970). A case against the normal use of inferential statistical models in educational research. *Educational Researcher, 3*, 6-7.

Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist, 49*, 997-1003.

Cowger, C. D. (1984). Statistical significance tests: Scientific ritualism or scientific method? *Social Service Review, 58*, 358-372.

Duncan, O. D. (1992). What if? *Contemporary Sociology, 21*(5), 667-668.

Duncan, O. D., & Stenbeck, M. (1988). Panels and cohorts: Design and model in the study of voting turnout. In C. C. Clogg (Ed.), *Sociological Methodology 1988* (pp. 1-35). Washington, DC: American Sociological Association.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika, 46*(1), 59-77.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. Fischer & I. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 15-38). New York: Springer-Verlag.

Fisher, W. P., Jr. (2010). Statistics and measurement: Clarifying the differences. *Rasch Measurement Transactions, 23*(4), 1229-1230 [http://www.rasch.org/rmt/rmt234.pdf].

Fisher, W. P., Jr. (2013). Suggestions for improving AERA's peer review process and quality of symposia. *Rasch Measurement Transactions, 27*(1), 1408-1409 [http://www.rasch.org/rmt/rmt271.pdf].

Hambleton, R. K., Swaminathan, H., & Rogers, L. (1991). *Fundamentals of item response theory*. Newbury Park, California: Sage Publications.

Latour, B. (2005). *Reassembling the social: An introduction to Actor-Network-Theory*. Oxford, England: Oxford University Press.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*(2), 103-115.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin, 100*, 398-407.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*, 355-383.

Michell, J. (2003). The quantitative imperative: Positivism, naïve realism and the place of qualitative methods in psychology. *Theory & Psychology, 13*(1), 5-31.

Michell, J. (2004). Item response models, pathological science and the shape of error: Reply to Borsboom and Mellenbergh. *Theory & Psychology, 14*(1), 121-129.

Molenaar, P. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective, 2*(4), 201-218.

Newby, V. A., Conner, G. R., Grant, C. P., & Bunderson, C. V. (2009). The Rasch model and additive conjoint measurement. *Journal of Applied Measurement, 10*(4), 348-354.

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods, 18*(3), 301-319.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Rasch, G. (1973/2011). All statistical models are wrong! Comments on a paper presented by Per Martin-Löf, at the Conference on Foundational Questions in Statistical Inference, Aarhus, Denmark, May 7-12, 1973. *Rasch Measurement Transactions, 24*(4), 1309 [http://www.rasch.org/rmt/rmt244.pdf].

Ricoeur, P. (1981). The model of the text: Meaningful action considered as a text. In J. B. Thompson (Ed.), *Hermeneutics and the human sciences: Essays on language, action and interpretation* (pp. 197-221). Cambridge, England: Cambridge University Press.

Rogosa, D. (1987). Casual [sic] models do not support scientific conclusions: A comment in support of Freedman. *Journal of Educational Statistics, 12*(2), 185-95.

Salzberger, T. (2010). Does the Rasch Model convert an ordinal scale into an interval scale? *Rasch Measurement Transactions*, *24*(2), 1273-1275 [http://www.rasch.org/rmt/rmt242.pdf].

Shrader-Frechette, K. (2014). *Tainted: How philosophy of science can expose bad science.* New York: Oxford University Press.

Stenner, A. J., Fisher, W. P., Jr., Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. *Frontiers in Psychology: Quantitative Psychology and Measurement,* *4*(536), 1-14 [doi: 10.3389/fpsyg.2013.00536].

Tesio, L., Simone, A., Grzeda, M. T., Ponzio, M., Dati, G., Zaratin, P. et al. (2015). Funding medical research projects: Taking into account referees' severity and consistency through many-faceted Rasch modeling of projects' scores. *Journal of Applied Measurement, 16*(2), 129-152.

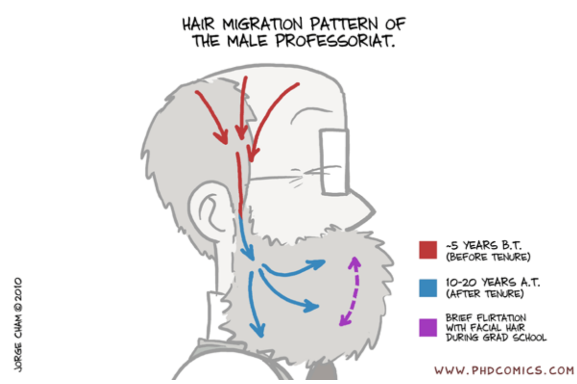
Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Wilson, M. R. (2013). Seeking a balance between the statistical and scientific elements in psychometrics. *Psychometrika, 78*(2), 211-236.

Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review, 3*(1), 281-288.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33-45, 52.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every educator and psychologist should know* (pp. 65-104). Hillsdale, New Jersey: Lawrence Erlbaum Associates.



Valuing One Another for All the Right Reasons

The previous issue of this publication honored Benjamin Wright, a leading scholar in the Rasch community and passionate advocate for objective measurement. The concept of objective measurement, in addition to its clear scientific value, carries ethical implications as well: as much as possible we want our evaluations of individuals to be *objective*, and thus invariant to construct-irrelevant factors, which often include socio-economic background, race and ethnicity, and biological sex and gender identity. In this way, the concept of objectivity pushes us to be more aware of how we valuing and treat one another, and our reasons for valuing and treating one another the way we do.

As I was reading the touching and heartfelt remarks about Ben in the previous issue, I came across this story recounted by Trevor Bond about a conversation with Ben:

Ben said: *You wrote this [Applying the Rasch Model] with Christine Fox?*

I replied: *Yes, from Toledo. Do you remember her?*

Ben said: *Aah, yes. She's beautiful*.

I then said: *And very smart.*

Then Ben said (smiling): *And very beautiful.*

In isolation, this story can be taken as a lightly humorous remembrance of a beloved colleague, and I have no doubt it was intended only in this way. But as I kept reading, it occurred to me that, in this issue of RMT, this was both the only story about a female scholar and also the only time in which physical appearance was mentioned. While this could easily be a fluke, it does resonate with some larger issues regarding the ways in which women are valued and treated throughout society, including in academia.

Undeniably, our community has historically been male-dominated. One need not look far for evidence: in the most recent issue of RMT, fourteen out of the sixteen remembrances of Ben are written by men. Ten out of the eleven chapters in the proposed festschrift for Ben are first-authored by men. All six of the photographed luminary figures on the last page of the issue are men. So far, all of the annual addresses at the Rasch SIG Business Meeting have been given by—you guessed it—men. (I believe similar comments could be made about race and ethnicity, and perhaps other factors as well.)

To be clear, I do not intend any of these comments to be taken as indictments against any individual person in our community, or even, necessarily, our community as a whole. Rather, I offer them only as observations, for which I do not have complete explanations, but which I nonetheless think deserve our collective awareness.

Some of my colleagues believe there are clear signs of progress in recent years, and I sincerely hope that this is true. I suspect that continued movement toward equity will depend on many factors, including the extent to which we can create a climate that feels welcoming to scholars from diverse backgrounds—a climate in which no one has any reason to doubt that they are valued and respected for all the right reasons.

As the next generation of scholars takes the stage, perhaps one way in which Ben's memory can be honored is via continued reflection on how, in both our professional and personal lives, we act in honor of the principles of objectivity, inclusiveness, and respect for the dignity of all persons. I welcome the thoughts of my colleagues on any of these issues.

Andrew Maul
University of California, Santa Barbara

***Editor's Note**: In full disclosure, a number of diverse individuals were invited to contribute to the Ben Wright memorial issue of RMT. My suspicion is that 14 of 16 remembrances written by males was largely coincidental. Additionally, the Rasch SIG leadership team for the past several years has consisted primarily of female leaders, and this year the SIG will welcome its first female speaker at the SIG Business Meeting. I tend to agree that signs of progress are evident, but certainly more can be done to foster inclusivity for all members of our community.

List of Rasch-Related Papers Scheduled for AERA 2016

A Rasch Analysis of Teachers' Use of School Discipline Techniques Survey: Perceptions of Teachers and Students - *Dandan Chen, University of Delaware; Ratna Nandakumar, University of Delaware*

A Rasch Analysis of the Current Opioid Misuse Measure for Chronic Pain Patients - *Lilian Linialy Chimuma, University of Denver; Kathy E. Green, University of Denver; Courtney R Morris, University of Denver*

A Rasch Measurement Analysis of an Adaptive Skills Survey for Individuals with Intellectual Disabilities - *Shelley C Sellwood-Davis, University of Kentucky*

A Validation Framework for Automated Essay Scoring Systems - *Lucy Lu, NSW Department of Education; James Tognolini, Pearson plc*

An Empirical Comparison of Classical Test Theory and Rasch Measurement: Scale Validation Application - *Mariya Yukhymenko, Ph.D., California State University - Fresno; Kimberly A. Lawless, University of Illinois at Chicago*

Application of Rasch Measurement Theory to Assess Career Aspiration Scale - *Hyojung Han, University of Georgia - Athens; Minho Kwak, University of Georgia - Athens; Jay W. Rojewski, University of Georgia*

Applying the Mixed Rasch Model to the Runco Ideational Behavior Scale - *Sedat Sen, Harran University*

Criteria for Formative and Summative Teacher Evaluation Using Peer Observations - *Rikkert van der Lans, University of Groningen; Wim van de Grift, University of Groningen; Klaas Van Veen, University of Groningen*

Does Item Sequence Order Impact Local Item Dependence in Surveys? - *Kent Hecker, University of Calgary; Kenneth Royal, North Carolina State University*

Equating International Performance Indicators in Primary Schools (iPIPS) Measures across Different Countries and Cultures - *Elena Kardanova, National Research University High School of Economics; Alina Ivanova, Higher School of Economics, Russia; Peter B Tymms, Durham University*

Evaluating a Revised Developmental Progression for Volume Measurement—Kindergarten through Grade 2 - *Douglas W. Van Dine, University of Denver*

Evaluating Rater Accuracy and Cognition for Document-Based Literacy Assessments Using a Mixed Methods Approach - *Jue Wang, University of Georgia - Athens; George Engelhard, University of Georgia; Kevin Raczynski, University of Georgia; Tian Song, Pearson; Edward W. Wolfe, Pearson*

Evaluating the Reliability and Validity of the English as an Additional Language/Dialect Learning Progression - *Joshua McGrane, NSW Department of Education and Communities; Lucy Lu, NSW Department of Education; Margaret Turnbull, NSW Department of Education and Communities*

Exploring Rater Errors and Systematic Biases in Language Assessment Using Mokken Scale Analysis - *Stefanie Anne Wind, The University of Alabama - Tuscaloosa; George Engelhard, University of Georgia*

Exploring Technology-Enhanced Item Format as Common Stimulus Using Multidimensional Rasch Latent Regression Models - *Daeryong Seo, Pearson Assessment & Information; Husein M. Taherbhai, Pearson; Anna M. Topczewski, Pearson*

Extending the Additive Factors Model to Assess Student Learning Rates - *Ran Liu, University of Pennsylvania; Kenneth R. Koedinger, Carnegie Mellon University*

Finding Optimal Number of Rating Scale Categories in Multiple Mini-Interviews Using the Many-Facet Rasch Measurement Model - *Vernon Mogol, University of Auckland; Warwick Bagg, University of Auckland; John Shaw, University of*

Auckland; Phillippa Poole, University of Auckland; John Monigatti

Graphical Aids in Middle School Mathematics: Impacting the Achievement Gap for English Learners - *Albert Manuel Jimenez, Kennesaw State University; Casey B. Nixon, Piedmont College*

Investigating the Influence of Student Religious/Social Conservatism on Perceptions of Intellectual Diversity on the University Campus: A Case for the Polytomous Rasch Model - *Mark Vincent Brow, University of Illinois at Chicago*

Item Format, Cognitive Domain, and Gender Interaction in TIMSS (Trends in International Mathematics and Science Study) 2011 Science Results - *Pey-Yan Liou, National Central University - Graduate Institute of Learning and Instruction; Okan Bulut, University of Alberta*

Many-Faceted Rasch Measurement: Assessing Rater Errors in Performance Assessment - *Priyalatha Govindasamy; Kathy E. Green, University of Denver; Maria del Carmen Salazar, University of Denver*

Measuring the Effectiveness of the Woodrow Wilson–Rockefeller Brothers Fund Fellowship for Aspiring Teachers of Color - *Tolani Britton, Harvard University*

Messy Middle or Messy Model: Challenges in Learning Progression Assessments - *Lokman Akbay, Rutgers University; Nathan D Minchen, Rutgers University - New Brunswick/Piscataway; Jimmy de la Torre, Rutgers University*

Motivational and Classroom Predictors of Academic and Career Choice in the Geosciences - *Kevin J. Pugh, University of Northern Colorado; Michael M. Phillips, University of Northern Colorado; Julie Sexton, University of Northern Colorado; Cassendra M. Bergstrom, University of Northern Colorado; Eric M Riggs, Texas A&M University; Selani D. Flores, University of Northern Colorado*

Peer Evaluation in the Austin Independent School District - *Lisa Schmitt, Austin Independent School District*

Reusing Mixed-Format Tests: Can We Adjust for Scoring Shift without Rescoring Previous Constructed Responses? - *Chi-Wen Liao, Educational Testing Service; Wei Wang, Educational Testing Service; Yi Cao, Educational Testing Service*

Revision and Validation of the Rasch-Based Scenario Scales for Measuring Activity Engagement in Older Adults - *Kelsey Klein, Boston College; Larry H. Ludlow, Boston College; Christina Matz-Costa, Boston College*

School Success Attitudes and Classroom Racial Diversity - *Odelia Simon, University of California - Santa Barbara*

South Carolina Educators for the Practical Use of Research: Examination of Item Quality in a Statewide Music Assessment Program Using Rasch Methodology - *Yin Burgess, University of South Carolina; Mihaela Ene, University of South Carolina; Elizabeth Leighton, University of South Carolina*

Subject Matter Knowledge of Geometry Needed in Tasks of Teaching and Experience Teaching Geometry - *Inah Ko, University of Michigan - Ann Arbor; Patricio G. Herbst, University of Michigan - Ann Arbor; Yung-Chi Lin, National Changhua University of Education*

Teacher Observations of Preschoolers' Social-Emotional Behavior: A Formative Evaluation - *Claire Christensen, SRI Internatinal; Katherine M. Zinsser, University of Illinois at Chicago*

Teaching for Equity Enactment Scenarios: An Application of Rasch Measurement Principles - *Larry H. Ludlow, Boston College; Wen-Chia Claire Chang, Boston College*

Testlet Effects on Pass/Fail Decisions under Competing Rasch Models - *Kari Hodge, NACE International Institute; Grant B. Morgan, Baylor University*

The Development of an Online Reading Ability Assessment for Junior High School Students - *Hsiu-Shuang Huang, National University of Tainan; Monlong Gan, National University of Tainan; Ya-Ying Tseng; Li Yun Hsu, National University of Tainan*

The Interrater Reliability of an Adapted Version of the Association of American Colleges and Universities VALUE Rubric: Evidence From Multifaceted Rasch Model - *Huda Sarraj, The University of Texas – Arlington; Araya Maurice, The University of Texas - Arlington*

The Validity and Reliability of the Newark Public School's Framework for Effective Teaching - *Andrew P. Swanlund, American Institutes for Research; Ryan Eisner, American Institutes for Research*

Use of Fine-Grained Learning Maps to Develop a Large-Scale Alternate Assessment - *Neal M. Kingston, University of Kansas*

Using a Modified Bookmarking Procedure on Survey Data to Inform Academic Misconduct Policy – *Mari-Wells Hedgpeth, North Carolina State University; Kenneth Royal, North Carolina State University; Keven Flammer, North Carolina State University*

Using Person Response Functions to Investigate Patterns of Person Misfit Related to Item Characteristics - *Angela Adrienne Walker, Emory University; Jeremy Kyle Jennings, University of Georgia - Athens; George Engelhard, University of Georgia*

Using Qualitative and Quantitative Research to Inform Survey Development - *Yan Wang, American Institutes for Research; Cong Ye, American Institutes for Research*

Using Rasch Modeling and Option Probability Curves to Diagnose Students' Misconceptions - *Cari F. Herrmann-Abell, American Association for the Advancement of Science; George E. DeBoer, American Association for the Advancement of Science*

Validating a Montessori High School Teacher Evaluation Survey - *Anthony P Setari, University of Kentucky; Kelly D. Bradley, University of Kentucky*

Validity Evidence for a Measure of Preventive Coping Resources - *Molly Allender, The University of Texas – Austin; Susan Murphy, The University of Texas – Austin; Richard G. Lambert, University of North Carolina – Charlotte; Christopher J. Mccarthy, The University of Texas – Austin; Maytal Eyal, The University of Texas - Austin*

What PISA (Programme for International Student Assessment) 2012 Reveals about Parental Involvement and Mathematical Achievement - *Luis Lizasoain*

2016 IMEKO Conference

The 2016 IMEKO TC1-TC7-TC13 Joint Symposium will take place at the University of California-Berkeley's Clark Kerr Campus, 3 - 5 August 2016. A call for papers, a registration and accommodation portal, and further information is available at http://imeko-tc7-berkeley-2016.org/.

The Symposium is organized by the Berkeley Evaluation and Assessment Research (BEAR) Center, in the UC Berkeley Graduate School of Education. The scope of the Symposium includes the main topics covered by three IMEKO Technical Committees:

- TC1 Education and Training in Measurement and Instrumentation
- TC7 Measurement Science
- TC13 Measurements in Biology and Medicine

The Symposium follows the tradition of the previous events of this well-established series. Three invited keynote lectures from renowned scientists will focus on key engineering, psychometric, and philosophical areas in measurement science.

The theme of this year's Joint Symposium follows from efforts initiated in the third edition of the International Vocabulary of Measurement (JCGM, 2008) to strive for broader cross-disciplinary inclusiveness in the definitions of measurement terms and concepts. In that context, engineers and psychometricians have initiated productive new exchanges of information and perspectives.

Looking forward to seeing everyone in Berkeley this summer!

William Fisher and Mark Wilson

Journal of Applied Measurement
Vol. 17, No. 1, 2016

Richard Smith, Editor, www.jampress.org

Call for Submissions

Research notes, news, commentaries, tutorials and other submissions in line with *RMT*'s mission are welcome for publication consideration. All submissions need to be short and concise (approximately 400 words with a table, or 500 words without a table or graphic). The next issue of *RMT* is targeted for June 1, 2016, so please make your submission by May 1, 2016 for full consideration. Please email Editor\at/Rasch.org with your submissions and/or ideas for future content.

Rasch-related Coming Events

Mar. 18, 2016, Fri. UK Rasch User Group Meeting, Durham, UK, www.rasch.org/uk

Mar. 23-24, 2016, Thur.-Fri. In-person workshop: Introduction to Rasch using Winsteps (W. Boone), Cincinnati, OH, www.raschmeasurementanalysis.com

Apr. 4-6, 2016, Wed.-Thur. IOMW2016 Conference, International Objective Measurement Workshop, Washington, DC, www.iomw.org

Apr. 8-12, 2016, Fri.-Tues. AERA Annual Meeting, Washington, DC, www.aera.net

Apr. 27-29, 2016, Wed.-Fri. In-person workshop: IRT/Rasch and CAT using Concerto (R), Cambridge, UK, www.psychometrics.cam.ac.uk/trainingworkshops

May 11-13, 2016, Wed.-Fri. In-person workshop: Introductory Rasch (M. Horton, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometrics

May 16-18, 2016, Mon.-Wed. In-person workshop: Intermediate Rasch (M. Horton, RUMM), Leeds, UK,

May 27-June 24, 2016, Fri.-Fri. Online workshop: Practical Rasch Measurement – Core Topics (E. Smith, Winsteps), www.statistics.com

June 16-19, 2016, Thur.-Sat. In-person workshop: Introduction to Rasch measurement analysis in the healthcare sciences and education (in English), Barcelona, Spain (L. Gonzalez de Paz, W. Boone, Winsteps)

July 1-29, 2016, Fri.-Fri. Online workshop: Practical Rasch Measurement – Further Topics (E. Smith, Winsteps), www.statistics.com

July 30-31, 2016, Sat.-Sun. PROMS 2016 Pre-Conference Workshop, Xi'an, China

Aug. 1-3, 2016, Mon.-Wed. PROMS 2016 Conference, Xi'an, China

Aug. 1-Nov. 25, 2016, Mon.-Fri. Online course: Introduction to Rasch Measurement Theory EDU5638 (D. Andrich, RUMM2030), www.education.uwa.edu.au