Equal-Interval Scaling

RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 24 No. 2 Autumn 2010 ISSN 1051-0796

Does the Rasch Model Convert an Ordinal Scale into an Interval Scale?

Promoting the Rasch Model

Empirical research papers advocating the use of the Rasch model (Rasch, 1960) typically emphasize the unique properties of Rasch measurement, for example, specific objectivity (Rasch, 1977), invariance, sample independence or raw score sufficiency, which are, in fact, closely related. While researchers using factor analytic approaches often downplay the role of specific objectivity or invariance, the scale level of the raw score remains a serious problem in factor analysis. Factor analysis is typically applied to the matrix of Pearson correlations, which require interval-scaled item scores. Therefore, the fact that the Rasch model does not depend on interval-scaled item scores is often put forward as an important property of the model.

Often it is explicitly stated that the Rasch model transforms ordinal raw scores into interval-scaled measures: Chien et al. (2009, p.418) say that "Rasch (1960) analysis transforms ordinal scores into the logit scale ...". Tennant and Conaghan (2007, p.1359) argue that Rasch analysis provides "a transformation of an ordinal score into a linear, interval-level variable, given fit of data to Rasch model expectations." Sometimes the argument is presented implicitly. Ewing et al. (2005, p.26) state that "Rasch measurement assumes responses on an ordinal level", Salzberger and Sinkovics (2006, p.412) point out that "[t]he manifest responses are assumed to be ordinal and need not be interval-scaled." Similarly, Pallant et al. (2006) as well as Pallant and Tennant (2007) speak of an ordinal raw score.

The scale level of the raw score

This claim deserves closer attention since, as a matter of principle, in statistics a lower scale level cannot be transformed to a higher level. Does the Rasch model travel faster than the speed of light? And if so, should we then not be allowed to transform the raw score in any way we want as long as the order is preserved? Actually the aforementioned claims are rather pragmatic and aimed at a non-Rasch audience. The statements simply express the fact that the item category scores merely have to be ordered with reference to the property to be measured. When applying the Rasch model, we actually do not have to be concerned with the scale level of the raw score. Based on a solid theoretical definition of a latent variable, fit of the data to the model assures us of having successfully measured a quantitative variable. However, Stevens' (1946, 1951) scheme of scale levels have been so influential that unavoidably the question arises as to which scale level we should actually ascribe to the raw score.

The raw score in the dichotomous Rasch Model

In the dichotomous case, the raw score is the observed number of items that are answered correctly (or agreed with) by the respondent. In other words, we count the number of correct items (Linacre and Wright, 1993). Counting, however, is distinctly different from measurement. The fact that it is often considered to be some sort of measurement is due to the misleading definition by Stevens (1946, 1951), who argued that measurement is achieved by assigning numerals to objects. Then the raw score would indeed succeed measurement, as the latter would be effected by coding the responses. Therefore, in the factor analytic world, manifest items are often referred to as measures of a latent variable.

In Rasch measurement, we define measurement as the successful discovery of the structure of quantity in the data (Michell, 1990, 1997), tantamount to data fitting the model. The raw score is actually the input to the analysis; it precedes rather than succeeds measurement. The raw score is the basis of an attempt to infer measures of a linear, interval-scaled latent variable. However, it is not some sort of "crude measurement" or "an approximation" per se. The scale level of the raw score is not an unconditional property of the score. It depends on what the scale level refers to. What we read off a measuring tape represents a ratio-scaled value of people's height, but as a raw score to be used in the measurement of

Table of Contents

intelligence, it would not even be ordinal-scaled. It is therefore improper to argue that the raw score a priori has a particular scale level pertaining to the quantitative property to be measured.

Given that the raw score is a count, its scale level is the highest possible, that is absolute. The point of origin is given by the extreme score of all items incorrect, while the unit is "one item". Obviously, it is permissible and meaningful to conclude that, for example, Mary has answered correctly twice as many items as John, if Mary mastered, say, six items, while John only got three items right. Statements of this sort are permitted regardless of whether the data fit the model or not, as we do not infer anything from the comparison of Mary and John beyond the number of correctly answered items. The absolute scale level of the raw score also implies that, and explains why, scale transformations of any sort are not allowed. It also justifies the fact that the raw score is calculated as the sum of individual item scores. If the individual item scores were ordinal, they could not be added up, since ordinal scale properties do not allow for addition.

Hence, the Rasch model does not "travel faster than light". Specifically, it does not transform an ordinal raw score into an interval-scaled measure. However, the Rasch model does not downgrade a higher scale level (absolute) to a lower one (interval), either. The fit of the data to the model tells us that an interval-scaled measure of a latent variable can be inferred from an a priori absolutely scaled observed raw score. If and only if data fit the model, we may ask what the scale level of the raw score a posteriori is with reference to the latent variable measured. The interval-scaled measures are derived from the raw score by a unique non-linear, s-shaped transformation. If the raw score were ordinal, such a transformation would not be possible. Consequently, the scale level of the raw score is higher than ordinal but lower than interval-scaled, as the unit is not preserved across the continuum. Thus, the Rasch model tests whether an a priori absolutely scaled raw score represents an a posteriori (that is after having demonstrated that a quantitative latent variable can be inferred from the data) non-linear raw score, which can be transformed into a linear interval-scaled measure of the latent variable (see table 1). Prior to the assessment of fit to the Rasch model, or in case of misfit, the scale level of the raw score with reference to the latent variable is undefined.

The raw score in generalized IRT

The term "generalized IRT" shall refer to all IRT models which are not Rasch models. In the Rasch model, the raw score does not depend on model estimates. By contrast, in the two-parameter logistic model (Birnbaum, 1968), the raw score is weighted by model parameters, which are a result of the model calibration. Thus, in the Rasch model, the input to and the output of the measurement analysis are strictly separated (which is just another way to express that the Rasch model features invariance). In generalized IRT the input and the output are entangled, unless the item discrimination parameters are known constants like in the one parameter logistic model (OPLM, Verhelst and Glas, 1995). Since the raw score in generalized IRT is not completely defined by the simple observation of items answered correctly, it is not a simple count. The distinction between an a priori raw score which is independent of the model estimates and an a posteriori raw score which has a scale level with reference to the latent variable is not possible, either. Since the fit of data to general IRT models cannot support the hypothesis of a quantitative variable, the scale level of the latent variable and of the weighted raw score remains questionable.

Raw score in the polytomous Rasch Model

Multicategorical responses have to be scored with successive integers starting at zero (Andersen, 1977; Andrich, 1978). This is compatible with the interpretation of the raw score as a count of all thresholds a respondent has passed. Consequently, the raw score is scaled absolutely in the polytomous case as well, provided the scoring of the categories adequately reflects the order of the thresholds (see Andrich, 1995a, 1995b). Strictly speaking, this qualification applies to the dichotomous model, too. If the response categories are wrongly scored, that is a score of one implies less of the property to be measured rather than more, the item will misfit. Rescoring the item will then resolve the problem, unless other reasons for misfit persist. In the polytomous model, the empirical thresholds may be reversed, signifying that the scoring is inappropriate. Then categories should be collapsed. However, rescoring the response categories alters the raw score. It is argued that both the original raw score as well as the revised raw score based on the amended scoring scheme are absolutely scaled, since both scores do not imply any meaning beyond the sheer count. Once the data have been shown to fit the polytomous Rasch model, we can ascribe meaning to the raw score with reference to the latent variable.

Fit of the data to the Rasch model	*Scale level a priori, with reference to the observed responses*	*Inference of measures of a quantitative latent variable*	*Scale level a posteriori, with reference to the quantitative variable*
not tested yet	absolute	not applicable	not applicable
misfit	absolute	impossible	not applicable
fit	absolute	possible	> ordinal, non-linear

Table 1: Scale level of the raw score

Conclusion

In summary, the fit of the data to the Rasch model implies that the raw score, which is scaled absolutely, conveys meaning regarding the quantitative property to be measured. With reference to the latent variable, the raw score is non-linear but clearly more than ordinal. In the case of misfit, though, the raw score has no such meaning at all. It is therefore recommended to better refrain from claims that the Rasch model transforms or converts ordinal scales into interval scales. Rather it should be pointed out that the Rasch model is capable of constructing linear measures from counts of qualitatively-ordered observations (Linacre and Wright, 1993), provided the structure of quantity is present in the data. The difference between ordered observations and an ordinal scale may seem subtle, but counts as such are certainly not merely ordinal, nor is the raw score merely ordinal with reference to the property to be measured once fit of the data to the model has been demonstrated. It goes without saying that those who apply the Rasch model are aware of this, at least implicitly. Alluding to ordinal scales of measurement may accommodate the traditional way of thinking, but it is misleading in the end.

The essential difference between the Rasch model and models rooted in classical test theory lies in the definition of measurement. In the Rasch model, the assignment of numerals to response categories merely enables us to properly count the number of correct items, or passed thresholds, but it is not equivalent to measurement. Measurement is achieved by successfully demonstrating that the latent variable complies with the structure of quantity. In factor analysis, measurement is essentially still based on assignment in Stevens' tradition. Therefore, scale levels of codes assigned to response categories are so important, while in fact testing the correspondence of the data to the structure of quantity is the core problem of measurement.

Thomas Salzberger

References

Andersen, E.B. (1977). Sufficient Statistics and Latent Trait Models. Psychometrika, 42, 69−81.

Andrich, D. (1978). Application of a Psychometric Rating Model to Ordered Categories which are Scored with Successive Integers. Applied Psychological Measurement, 2 (4), 581−594.

Andrich, D. (1995a). Models for Measurement, Precision and the Non-Dichotomization of Graded Responses. Psychometrika, 60 (1), 7−26.

Andrich, D. (1995b). Further Remarks on Non-Dichotomization of Graded Responses. Psychometrika, 60 (1), 37−46.

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In F.M. Lord and M.R. Novick (eds), Statistical Theories of Mental Test Scores, Reading, MA: Addison-Wesley, Chapters 17−20.

Chien, T.-W., Hsu, S.-Y., Chein, T., Guo, H.-R., & Su, S.B. (2008). Using Rasch Analysis to Validate the Revised PSQI to Assess Sleep Disorders in Taiwan's Hi-tech Workers. Community Mental Health Journal, 44:417–425.

Ewing, M., Salzberger, T., & Sinkovics, R. (2005). An Alternate Approach to Assessing Cross-Cultural Measurement Equivalence in Advertising Research. Journal of Advertising, 34 (1), 17−36.

Linacre, M., & Wright, B. (1993). Constructing linear measures from counts of qualitative observations. Paper presented at the Fourth International Conference on Bibliometrics, Informetrics and Scientometrics, Berlin, Germany.

Michell, J. (1990). An Introduction to the Logic of Psychological Measurement. Hillsdale: Erlbaum.

Michell, J. (1997). Quantitative Science and the Definition of Measurement in Psychology. British Journal of Psychology, 88, 355−383.

Pallant, J.F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). British Journal of Clinical Psychology, 46 (1), 1-18.

Pallant, J.F., Miller, R.L., & Tennant, A. (2006). Evaluation of the Edinburgh Post Natal Depression Scale using Rasch analysis. BMC Psychiatry, 6:28.

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research, expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Rasch, G. (1977). On Specific Objectivity: an Attempt at Formalizing the Request for Generality and Validity of Scientific Statements. Danish Yearbook of Philosophy, 14, 58−93.

Salzberger, T., & Sinkovics, R. (2006). Reconsidering the Problem of Data Equivalence in International Marketing Research – Contrasting Approaches Based on CFA and the Rasch Model for Measurement. International Marketing Review, 23 (4), 390−417.

Stevens, S.S. (1946). On the Theory of Scales of Measurement. Science, 103, 667−680.

Stevens, S.S. (1951). Mathematics, Measurement, and Psychophysics. In S.S. Stevens (ed), Handbook of Experimental Psychology, New York, NY: Wiley, 1−49.

Tennant, A., & Conaghan, P.G. (2007). The Rasch Measurement Model in Rheumatology: What Is It and Why Use It? When Should It Be Applied, and What Should One Look for in a Rasch Paper? Arthritis & Rheumatism (Arthritis Care & Research), 57 (8), 1358–1362.

Verhelst, N.D., & Glas, C.A.W. (1995). The one parameter logistic model. In G.H. Fischer and I.W. Molenaar (eds), Rasch Models, Foundations Recent Developments, and Applications, New York: Springer, pp. 215-237.

An Investigation of Societal and Environmental Consciousness among Singaporean Early Teens

We sought to develop a measurement instrument that evaluated Singaporean early teens' societal and environmental consciousness; we call the instrument Singaporean Societal and Environmental Consciousness Inventory (SSECI).

The instrument comprises 12 items rated on a six-level rating scale, ranging from 1 (strongly disagree) to 6 (strongly agree). We reverse coded items 2, 9, and 12. We administered the inventory to 351 Singaporean early teens, aged 14. To test the psychometric properties of the measurement tool, we used a Rating Scale Model (RSM).

The first analysis showed that items 2 (outfit MNSQ = 1.49) and 12 outfit MNSQ = 1.65) were misfitting. The properties of Rasch models apply to the extent that the data fit the model. If the data do not fit, person trait level and item endorsability measures are inaccurate, and the data are unlikely to be unidimensional. We found that the average person measures of the responses in categories 3 and 5 in item 2, and in categories 3 and 4 in item 12, did not ascend with category scores; and a huge difference was observed between the observed and expected point measure correlation values for those items. Additionally, we investigated person performance patterns and fit. We identified 14 individuals with erratic response patterns and huge misfits: the scalograms showed that some

Table 1. *Psychometric Quality of the Refined Singaporean Early Teens' Societal & Environmental Consciousness Inventory*

Item	**Measure**	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Point Measure	Construct theory	Statement
								More demanding:
5	0.92	1.02	0.24	1.00	0.02	0.48	High	I often design or **make useful tools** (e.g., holding or standing aids for the needy).
10	0.50	0.98	-0.23	0.98	-0.27	0.53	High	I often design or make useful objects using **reused materials** (e.g., containers, papers and clothes for myself or others).
7	0.32	0.71	-4.36	0.71	-4.35	0.72	Medium	I often participate in environmental or energy **conservation activities**.
4	0.25	0.84	-2.18	0.84	-2.17	0.62	Medium	I often **participate in charity activities** or community services.
11	0.04	0.91	-1.18	0.91	-1.25	0.58	Medium	I am willing to **pay more for environmental friendly** products.
8	0.03	0.91	-1.29	0.91	-1.30	0.60	Medium	I like to read nature, wildlife or **environmental news** or magazines.
9REV	-0.04	1.30	3.82	1.35	4.36	0.26	High	I *seldom* use **environmentally friendly products** (e.g., recyclable bags, papers and non-polluting sprays). [Reversed]
6	-0.07	1.24	3.12	**1.41**	5.24	0.38	Low	I will not buy my favourite brand if I know the producer has been giving **unfair treatment to the workers.**
12REV	-0.25	1.28	3.61	1.30	3.77	0.35	Low	I will buy my favourite brand even if I know the material used or the producer was causing *harm* to **the environment**. [Reversed]
1	-0.39	0.82	-2.30	0.82	-2.38	0.62	Medium	I am willing to do **volunteer work**.
2REV	-0.64	1.13	1.56	1.27	3.01	0.39	Low	I feel I should *not* help to raise **funds for charity**. [Reversed]
3	-0.67	0.84	-1.98	0.83	-2.17	0.58	Medium	I am willing to **donate money** for charitable causes.
								Less demanding:

persons with higher trait levels had unexpectedly endorsed lower response categories and those with lower trait levels had endorsed higher response categories. We removed these 14 people for separate investigation and reanalyzed the remaining data. The results, as displayed in Table 1, were promising and closer to our expectations.

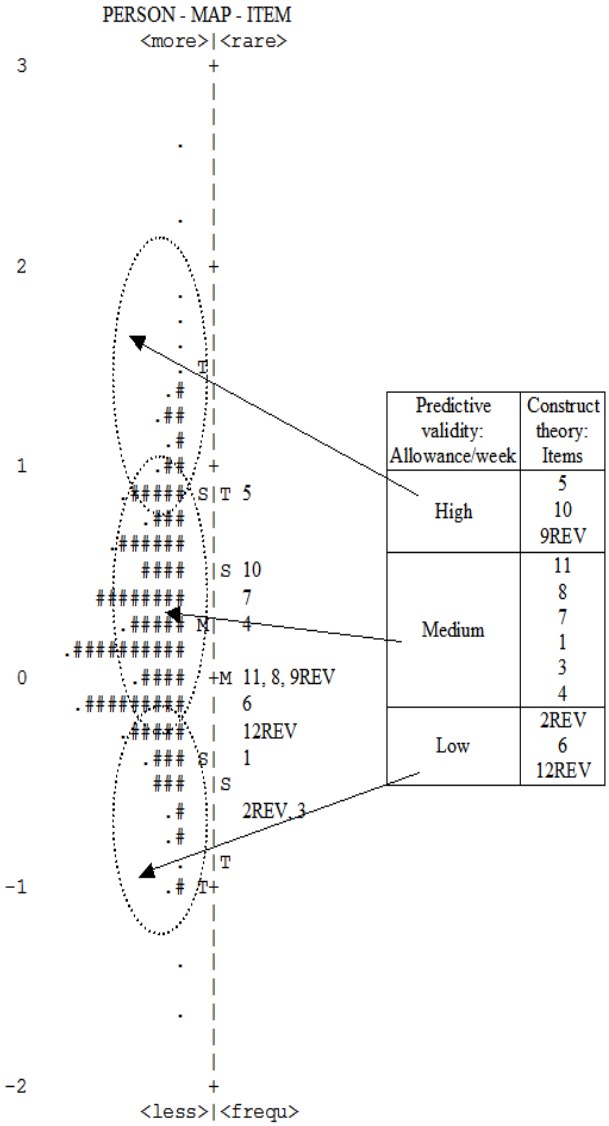

Figure 1. Item-person map (on the left side) and their hypothesized values (on the right side). Each "#" is 4 persons. Each "." is 1 to 3 persons.

Bond and Fox (2007) regard 0.6 to 1.4 as the acceptable range of fit indices. As Table 1 displays, fit MNSQ indices fell in the range of 0.71-1.41; item difficulty measure has a fairly large spread from a low of -0.67 to a high of 0.92; and point measure correlations are positive and considerably high.

Figure 1 graphically displays the location of items and persons in the second analysis and compares them with our hypothetical "construct map" (Wilson, 2005) on the right side. Reading the item contents, we expected that some items likely land on the top, some in the middle, and some at the bottom of the hierarchy. This expectation was met by several items, providing evidence for the construct validity of the measurement tool; the hypothesized latent trait is fairly well-targeted by items.

Predictive validity: persons were divided into high, medium, and low allowance-per-week subgroups. We expected that higher allowance subgroups would be positioned at the top of the map. The ellipses in Figure 1 serve to match the expected construct map against the actual Rasch measures. We observed that this expectation is met in many instances although several participants fell out of the expected areas. The overlapping areas of the ellipses represent the unexpected locations of people based on the allowance-per-week criterion.

Hui Ling Ng, S. Vahid Aryadoust, and Yau Che Ming
National Institute of Education,
Nanyang Technological University,
Singapore

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences. London: Lawrence Erlbaum Associates.

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

Rasch Analysis Symposium of Experiences in Journal Publication

Chi Mei Hospital, Tainan, Taiwan
Aug.7, 2010

Speakers included Professors ...
Tsair-Wei Chien, Chi Mei
Zhang Zhihong, Northwestern University,
Wang Wenzhong, the Hong Kong Institute of Education,
Xie Qinglin Rehabilitation, Department of National Taiwan University,
Liang Wenmin, Taichung,
Shih-Bin Su, Zhou Wei Ni, and Hung-Jung Lin, Chi Mei,
Fang Chen, Zhongshan, Ying-Yao Cheng, Shi Qinglin, and more.

The seminar discussed the applications of evidence based medicine, from survey and design, data analysis and evaluation, statistical analysis and methods of application,

quality management and strategic planning. Also ways in which the Rasch model can be used to improve the

measurement and assessment methods to distinguish between the item difficulty, discrimination, increased measurement reliability, validity and so.

"Clinical skills test" for future physicians, a necessary condition for national examinations, the implementation of "objective structured clinical examination (OSCE)" assessment all require objective assessment tools which Rasch provides.

Unifying the Language of Measurement

The frequent misplacement of Rasch measurement in the domain of Item Response Theory might be corrected by more persuasively and comprehensively situating it in the context of natural science, following the lead of L.L. Thurstone, G. Rasch, B.D. Wright, D. Andrich and others. The implicit and largely untested claim in this work is that the theory and practice of measurement across the sciences is unified insofar as it demands, models, and capitalizes on principles such as invariance, statistical sufficiency, parameter separation, unidimensionality, conjoint additivity, local independence, etc. Interest in such unification comes from both the social sciences and the natural sciences, as was indicated in a recent talk by Ludwik Finkelstein (2010; also see his 2004), a past editor (1982-2000) of *Measurement*, and a former Vice-President of the International Measurement Confederation:

> The development of measurement science as a discipline has not paid adequate attention to the wider use of measurement. It is increasingly recognized that the wide range and diverse applications of measurement are based on common logical and philosophical principles and share common problems. However the concepts, vocabularies and methodologies in the various fields of measurement in the literature tend to differ. The development of a unified science of measurement appropriate for all domains of application seems to be desirable. Such a unified measurement science would contribute to the meeting the needs of a better general education in measurement.

There are theoretical and practical reasons for pursuing a unified science of measurement. As Finkelstein says, common logical and philosophical principles share common problems, but differing concepts, vocabularies and methodologies obscure those commonalities and make the solution of their shared problems needlessly difficult. Some problems are likely to have been better addressed to date in the social sciences, and others, in the natural sciences.

For instance, the Rasch focus on invariance will likely be found of particular value within metrology. The obvious places to begin Rasch applications in the context of metrology are with regard to its educational and human resource needs for individualized tests informing differentiated instruction, computerized adaptive certification exams, admission and graduation standards, program improvement and comparison metrics, employee assessments and opinion surveys, etc. But Rasch theory and methods might also play a role in resolving some problems that social scientists assume to be completely under control in the natural sciences, as in the potential for Rasch models to inform genomic and proteomic metrologies (Markward & Fisher, 2004) or clinical laboratory disease severity measures (Fisher & Burton, 2010). Coming at it from the other direction, the metrological focus on the traceability of individual instruments and measures to universally uniform reference standards will likely result in significant advances within the social sciences. The role of these kinds of technical networks in reducing market frictions (Barzel, 1982) and in amplifying individual effects in a kind of choral collective cognition (Magnus, 2007) may have profound implications for the advancement of science (Latour, 1987, 2005), economics, government, and the work place. There is, then, a potential for the theory and practice of invariant measurement advanced in work following from Rasch to piggyback on the principle of networked metrological traceability, while those networks capitalize in new ways on the potentials brought to bear by Rasch's principles of invariance.

The International Measurement Confederation (IMEKO; http://www.imeko.org/) hosts annual and bi-annual meetings of its various technical committees (TCs) at locations globally. Of particular interest to Rasch measurement practitioners are the IMEKO TC1 on metrology education and TC7 on measurement science. The 13th IMEKO TC1-TC7 Joint Symposium took place September 1-3, 2010 at City University in London. For the first time, this Symposium included the IMEKO TC13 - Measurements in Biology and Medicine. The Symposium was organized by Sanowar Khan, Kenneth Grattan, Ludwik Finkelstein, and Panicos Kyriacou of the School of Engineering and Mathematical Sciences at City University London. Sponsors included the Institute of Physics (IOP), UK, City University, the Institute of Measurement and Control, and the Worshipful Company of Scientific Instrument Makers. The conference program can be accessed online at http://imeko.iopconfs.org/.

Three Rasch papers were presented at the conference (Bezruzcko & Fatani, 2010; Fisher, 2010; Fisher, Elbaum, & Coulter, 2010). These papers presented variations on the same rationale for presenting research employing Rasch measurement at such a conference, namely, that models requiring linear, invariant comparisons provide an equivalent basis for quantification, no matter the field in which they happen to be employed. There is a need for further elaborations and explorations of Rasch's appropriation of the structure of natural law embodied in the Standard Model used by Maxwell in his analysis of mass, force, and acceleration. In starting from Maxwell's work in this way, Rasch capitalized on the uniformity with which natural laws involve the equivalence of one parameter with the multiplication or division of two other parameters, such that "virtually all the laws of physics can be expressed numerically as multiplications or divisions of measurements" (Ramsay, Bloxom, & Cramer, 1975, p. 258). Models of this kind require well-defined homomorphisms between empirical and numerical relational structures. Referred to by Rasch as isomorphisms, these are usually absent in social science scaling models, which are typically presumed valid and fit to data whether or not the empirical and numerical

relational structures match (Krantz, Luce, Suppes, & Tversky, 1971).

Thus, Rasch models are properly situated in the tradition of measurement in the natural sciences because of their formal properties. But these properties are insufficient in themselves to the task of unifying measurement across the sciences. Ramsay, et al. (1975, p. 262) recognize that "Progress in physics would have been impossibly difficult without fundamental measurement," and that "we may have to await fundamental measurement before we will see any real progress in quantitative laws of behavior." But fundamental measurement and rigorously validated quantitative laws of behavior have been available now for decades, with little recognition or acceptance of their value in mainstream social science. Plainly something else besides mathematical proofs, experimental evidence, predictive theories, and persistently invariant instrumentation is needed for social scientists to adopt fundamental measurement and build out the theory and practice of psychosocial laws integrating qualitative and quantitative data and methods.

Latour (1987, 2005) provides convincing arguments and evidence to the effect that social networks are essential to the spread of new ideas and methods in science. There seems to be a popular notion that fundamental measures incorporated in lawfully regular patterns of inter-related phenomena are what one might call "naturally natural," and that these somehow propagate themselves spontaneously into existence as universally uniform and available things or effects. Latour and others reveal the huge resources invested in, and material practices associated with, making constructs that are incredibly rare in nature seem quintessentially natural. Steel, for instance, may well exist in nature, but certainly not in the quantities in which it has been manufactured over the last century and more. Rather than discovering pre-existing phenomena in nature, science and technology combine together to isolate useful and meaningful phenomena that are then exported from laboratories. The key to the process resides in the way technical media encapsulate and package an effect so as to keep it always connected with the networks of energy, communications, tools, and technicians that make it seem naturally universal. It may be then that to make psychosocial constructs seem "naturally natural," social scientists need to find a way to deploy those constructs via networks of measures metrologically traceable to reference standards.

A question that arose in the course of making the Rasch presentations at the IMEKO meeting led to some insight into the kinds of challenges likely to arise in the course of addressing the need for a unified language and practice of measurement. Though the term "calibration" is commonly employed in Rasch measurement to refer to the process of evaluating the invariance properties and estimating the parameters of an instrument, this process is almost always undertaken in an exploratory fashion, with no reference to a previously existing uniform standard metric (or even to previously calibrated instruments measuring the same construct). In the natural sciences and engineering, however, calibration does not mean anything except confirming traceability to a reference standard. Calibration is always relative to a standard.

This difference in the use of the same term represents a significant way in which barriers to understanding might arise. Because universal uniform metric standards, such as degrees Celsius, kilograms, meters, the second, etc., are nearly nonexistent in the social sciences, calibration is not yet a matter of establishing that kind of correspondence. Conversely, such standards are the norm in the natural sciences. New constructs are either rare or not considered candidates for calibrated instrumentation until standards are developed. There are, accordingly, few, if any, of the theoretical or practical guidelines for evaluating invariance in new constructs, estimating initial parameters, equating instruments, etc. that are taken for granted in Rasch-oriented psychometrics.

In light of this contrast between measurement in the natural and social sciences, other seeming similarities took on new significance. For instance, multiple papers presented at the London conference took up issues involving ordinal and nominal scales, referring to Stevens' fourfold measurement classification system in positive terms. When the question was raised as to why any interest would be invested in ordinal scales in the context of the natural sciences, the reply repeated the previous emphasis on the fact that all measurement, ordinal and nominal as well as interval and ratio, is performed relative to existing standards.

The Mohs Hardness Scale, for instance, provides an ordinal standard of measurement that works because it definitively encompasses virtually the entire range and every instance of possible variation in the construct in a

IMEKO 2011

The next Joint International IMEKO TC1-TC7-TC13 Symposium will be in Jena, Germany, August 31 - September 02, 2011. The symposium web site at http://www.tu-ilmenau.de/fakmb/Home.2382.0.html will begin accepting presentation proposal abstracts on January 1, 2011. The submission process will close on March 31, 2011, with notification of acceptances by April 30. Full papers will be due by June 1, and presenters will be expected to register for the conference by July 1. Papers are presented and published in English.

There is considerable interest in learning more about Rasch measurement among the IMEKO TC1, TC7, and TC13 membership. Plans are to include an overview of Rasch theory and methods in a major plenary session. Members of the Rasch Measurement SIG and readers of Rasch Measurement Transactions are strongly encouraged to submit their best work for presentation in Jena at the IMEKO meeting.

metric that informs almost all applications involving it. Similarly, nominal standards define the shape of geometrical figures in the same way dictionaries define the meaning of words.

Reservations concerning the value and utility of ordinal measures in the social sciences, in turn, were seen in a new light by natural scientists at the IMEKO conference when the incomparability of scores from two different mathematics tests was raised and was amplified by then proposing to add a new item to both tests, which would make subsequently gathered scores incomparable with each of the already incomparable original tests. The chaos and confusion in that scenario was briefly contrasted with the simplicity and elegance of the comparisons that could be made of measures from exactly the same groups of items if those items had been drawn from a bank of items calibrated to measure in the same unit. Once again, all the difference was made by the presence or absence of a calibrated standard.

Two themes running through virtually all of the papers presented at the conference concern the most exciting and challenging areas for measurement-focused collaboration between social and natural scientists: metrological traceability and measurement uncertainty. The former is the domain of the International Vocabulary of Measurement, or VIM, now in its third edition (BIPM, et al., 2008). The latter is covered in the Guide to the expression of Uncertainty in Measurement, or GUM (BIPM, et al., 1995). These works are significant in being created and adopted as standards by an authoritative international group known as the Joint Committee for Guides in Metrology (JCGM), which includes the International Bureau of Weights and Measures (BIPM), the International Electrotechnical Commission (IEC), the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), the International Organization for Standardization (ISO), the International Union of Pure and Applied Chemistry (IUPAC), the International Union of Pure and Applied Physics (IUPAP), the International Organization of Legal Metrology (OIML), and the International Laboratory Accreditation Cooperation (ILAC).

Several presentations at the 2010 IMEKO meeting in London, including the closing plenary, updated the IMEKO membership on the GUM and the VIM, especially as regards work towards the next, fourth, edition of the latter, expected to be published in 2018. These presentations were given primarily by Luca Mari (Università Cattaneo, Italy) and Paul de Bièvre (Editor, *Accreditation & Quality Assurance*), both of whom were involved the production of the VIM3 and are continuing work toward the VIM4. One of the goals for the VIM4 is to include, so far as possible, concepts and terminology that unify the language of measurement across the sciences. For some background on where work in this direction is starting from, see Mari (2010a, 2010b), Mari and Ugazio (2010), and Mari and Giordani (2010). Paul de Bièvre's (2006, 2010; Price & de Bièvre, 2009) articles and editorials in *Accreditation and Quality Assurance* are also illuminating. Other IMEKO presentations on probabilistic inferences (Rossi, 2010), uncertainty (Pavese, 2010; Pertile & Debei, 2010; Weiβensee, Kühn, & Linβ, 2010), multiscale models (Abdulla, Imms, Schleich, & Summers, 2010) and ordinal scales (Benoit, 2010) are also illustrative of current perspectives on problems related to "soft" or "wider" measurement in physics and engineering.

The VIM3 defines the concepts and associated terms employed in identifying units of measurement that are comparable across samples, instruments, operators, labs, time, and space. The realization of comparability requires a prior positive outcome of an experimental test of the hypothesis that an invariant, additive unit exists. The hows and whys of producing this outcome for new, previously unmeasured variables are not obvious or self-evident, but the culture of measurement in the natural sciences is oriented to the implementation of existing standards. The VIM3 says little or nothing concerning the form or content of hypotheses of invariant constructs, the observational frameworks, experimental designs, and estimation methods used in evaluating it, or the criteria for determining if and when that hypothesis is falsified.

Though these aspects of measurement theory and practice have developed to mature and widely applied forms over the last 80 years, they have not been proposed, debated, or consolidated as standard procedures. It may not, in fact, be appropriate to present them as standards. Instead, perhaps it would be better to provide methodological recommendations, and to focus on (a) specifying the properties of a variable, such as reading or cognitive development, already measured in a unit functioning as a de facto standard, and (b) defining the concepts needed for establishing traceability to it as a recognized *de jure* standard. It should be expected that these concepts will likely differ significantly from those associated with calibration to existing standards in the natural sciences and engineering, given the intangible, social nature of the constructs and their basis in ordinal observations.

There is a great need for the involvement of Rasch measurement theoreticians and practitioners in the formulation of a unified language of scientific measurement. The challenges are huge, but the returns on the investments, when measured in terms of human value, social cohesion, and environmental quality, stand to be even huger.

William P. Fisher, Jr.

References

Abdulla, T., Imms, R., Schleich, J. M., & Summers, R. (2010). Multiscale information modelling for heart morphogenesis. *Journal of Physics: Conference Series, 238*: http://iopscience.iop.org/1742-6596/238/1/012062.

Barzel, Y. (1982). Measurement costs and the organization of markets. *Journal of Law and Economics 25*: 27-48.

Benoit, E. (2010). An ordinal metrical scale built on a fuzzy nominal scale. *Journal of Physics: Conference Series, 238*: http://iopscience.iop.org/1742-6596/238/1/012034.

Bezruczko, N. & Fatani, S. S. (2010). Probabilistic measurement of non-physical constructs during early childhood: epistemological implications for advancing psychosocial science. *Journal of Physics: Conference Series, 238*: http://iopscience.iop.org/1742-6596/238/1/012053.

BIPM, IEC, IFCC, ILAC, IUPAC, IUPAP, ISO, OIML (2008) International Vocabulary of Metrology: basic and general concepts and associated terms (VIM) 3rd edn. Online available as JCGM 200:2008 at: http://www.bipm.org/en/publications/guides/vim.

BIPM, IEC, IFCC, ILAC, IUPAC, IUPAP, ISO, OIML (1995) Guide to the expression of Uncertainty in Measurement (GUM). Online available as JCGM 100:2008 at: http://www.bipm.org/en/publications/guides/gum.

de Bièvre, P. (2006). Counting is measuring: Learning from the banks? *Accreditation & Quality Assurance, 11*: 1-2.

de Bièvre, P. (2010). A metrological traceability chain prevents circular reasoning in measurement design. *Accreditation & Quality Assurance, 15*: 491-492.

Finkelstein, L. (2010). Measurement and instrumentation science and technology—the educational challenges. *Journal of Physics: Conference Series, 238*: http://iopscience.iop.org/1742-6596/238/1/012001.

Fisher, W. P., Jr. (2010, September 1-3). The standard model in the history of the natural sciences, econometrics, and the social sciences. *Journal of Physics: Conference Series, 238*(1), http://iopscience.iop.org/1742-6596/238/1/012016.

Fisher, W. P., Jr., & Burton, E. (2010). Embedding measurement within existing computerized data systems: Scaling clinical laboratory and medical records heart failure data to predict ICU admission. *Journal of Applied Measurement,11*, 271-287.

Fisher, W. P., Jr., Elbaum, B., & Coulter, A. (2010). Reliability, precision, and measurement in the context of data from ability tests, surveys, and assessments. *Journal of Physics, Conference Series, 238*(1), http://iopscience.iop.org/1742-6596/238/1/012036.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement. Volume 1: Additive and polynomial representations.* New York: Academic Press.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society.* New York: Cambridge University Press.

Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory.* Oxford, England: Oxford University Press.

Magnus, P. D. (2007). Distributed cognition and the task of science. *Social Studies of Science, 37*(2), 297-310.

Mari, L. (2010a). Properties as measurands: an overview and some critical issues. *Journal of Physics: Conference Series, 238*: http://iopscience.iop.org/1742-6596/238/1/012002.

Mari, L. (2010b). A VIM3-compliant measurement model and some related issues. *Journal of Physics: Conference Series, 238*: http://iopscience.iop.org/1742-6596/238/1/012025.

Mari, L & Ugazio, E. (2010). Preliminary analysis of validation of measurement in soft systems. *Journal of Physics: Conference Series, 238*: http://iopscience.iop.org/1742-6596/238/1/012026.

Mari, L & Giordani, A. (2010). Towards a concept of property evaluation type. *Journal of Physics: Conference Series, 238*: http://iopscience.iop.org/1742-6596/238/1/012027.

Markward, N. J., & Fisher, W. P., Jr. (2004). Calibrating the genome. *Journal of Applied Measurement 5*(2), 129-41.

Pavese, F. (2010). Comparing statistical methods for the correction of the systematic effects and for the related uncertainty assessment. *Journal of Physics: Conference Series, 238*: http://iopscience.iop.org/1742-6596/238/1/012041.

Pertile, M. & Debei, S. (2010). Comparison between two modern uncertainty expression and propagation approaches. *Journal of Physics: Conference Series, 238*: http://iopscience.iop.org/1742-6596/238/1/012033.

Price, Gary & de Bièvre, Paul. (2009). Simple principles for metrology in chemistry: Identifying and counting. *Accreditation & Quality Assurance, 14*: 295-305.

Ramsay, J. O., Bloxom, B., & Cramer, E. M. (1975, June). Review of *Foundations of Measurement, Vol. 1*, by D. H. Krantz et al. *Psychometrika, 40*(2), pp. 257-262.

Rossi, G. B. (2010). Probabilistic inferences related to the measurement process. *Journal of Physics: Conference Series, 238*: http://iopscience.iop.org/1742-6596/238/1/012015.

Weiβensee, K., Kühn, O. & Linβ, G. (2010). Knowledge-based uncertainty estimation of dimensional measurements using visual sensors. *Journal of Physics: Conference Series, 238*: http://iopscience.iop.org/1742-6596/238/1/012023.

ICOM 2010: International Conference on Outcomes Measurement

1-3 September, 2010
NIH Natcher Conference Center - Bethesda, MD

Wednesday, September 1, 2010

Opening Plenary: Introduction to ICOM
1. Wim van der Linden (CTB/McGraw-Hill, Monterey, CA): State of the Art of Psychometrics Today.
2. Robert Massof (Johns Hopkins University, Baltimore, MD): Intro/Overview for Rasch Measurement.
3. Karon Cook (University of Washington, Seattle, WA): "Bewildered, Befuddled, Be-Tooled: A Blue-Collar Psychometricians Defense of Measurement Models Tools. . . Just Tools".

Chair: Kendon J. Conrad, Discussant: Ron Hambleton (University of Massachusetts, Amherst, MA)

1A.1: Causality
1. A. Jackson Stenner*, Mark H. Stone, Donald S. Burdick, Stedman Stevens: Causal Rasch Models
2. Adam C. Carle*: Using multiple group (MG) multiple indicator multiple cause (MIMIC) models to evaluate multiple sources of measurement bias simultaneously
3. Andre A. Rupp*: Modern Diagnostic Measurement using Latent-variable Methods: Methods, Theory, and Applications

Chair: Karen M. Conrad, Discussant: Michael Dennis

1A.2: Item Development
1. I-Chan Huang*, Pey-Shan Wen, Elizabeth Shenkman, Patricia Shearer, Gwendolyn Quinn: Developing initial item pools to measure quality of life for young adult survivors of childhood and adolescent cancer
2. Richard D. Lennox*: Use of Cognitive Interviewing Techniques for Improving Outcome Measures

Chair: Jessica Mislevy, Discussant: Richard Sawatzky

1A.3: Considering Cronbach's Alpha
1. Agustin Tristan*: Theoretical Alpha values for objective tests
2. David Andrich*: Cronbach Alpha in the presence of subscales

Chair: Ken Conrad, Discussant: Steven Reise

Post-Lunch Plenary Session 1
1. Steven Reise (University of California, Los Angeles, CA): The Impact of Multidimensionality on Unidimensional Item Response Theory Model Parameters
2. Stephen Humphry (University of Western Australia): Developing Systems of Units in Psychology

1B.1: Applications in Mental Health
1. Pey-Shan Wen*, Kay Waid-Ebbs, Neila J. Donovan, Shelley C. Heaton, Craig A. Velozo: The Rater Effects of the Behavior Rating Inventory of the Executive Function-Adult in Individuals with Traumatic Brain Injury and Their Caregivers
2. Jason E. Chapman*, Michael R. McCart, Ashli J. Sheidow, Elizabeth J. Letourneau: The Use of Rasch and Many-Facet Rasch Models to Compare Untrained and Partially-Trained Raters in the Measurement of Therapist Adherence
3. Ann M. Doucette*: Implications of Unexamined Measurement Properties in Modeling Psychotherapy Intervention Outcomes

Chair: Richard Lennox, Discussant: Paul Pilkonis

1B.2: Cut Scores and Anchors
1. Brian J. Hess*, Weifeng Weng, Rebecca S. Lipner: Setting Cutscores on Composite Measures of Clinical Performance
2. Sergio Romero*: Using the Rasch Model to investigate suggested cut-off score of the Berg Balance Scale (BBS)
3. Monica K. Erbacher*, Karen M. Schmidt, Steven M. Boker, Cindy S. Bergeman: Apples to Apples: A Comparison of Four Methods for Anchoring Partial Credit Model Trait Level Scores Across Time

Chair: Agustin Tristan, Discussant: Ronald Hambleton

1B.3: CAT Topics
1. Craig Velozo*, Leigh Lehman, Ying-Chih Wang, Pey-Shan Wen, Sergio Romero: Developing a Hierarchically-Based Physical Function CAT Battery
2. Richard Sawatzky*, Pamela A. Ratner, Bruno D. Zumbo, Jacek A. Kopec, Amery Wu: Examining the Implications of Sample Heterogeneity with Respect to the Measurement Validity of Computerized Adaptive Tests
3. Jessica Mislevy*, André Rupp*, Jeffrey Harring: Identifying Local Item Dependence in Computer Adaptive Health-Outcomes Assessments

Chair: Carl Granger, Discussant: Richard Gershon

1B.4 Measurement Error and Demographics

1. Adam C. Carle*: Systematic measurement errors influence on disparities in national rates of children with special health care needs across Spanish and English speaking households.
2. Adrienne Garro*: Measuring Quality of Life and Coping Strategies in Families of Children with Asthma: A Comparison between Latinos and Caucasians

Chair: Junius Gonzalez, Discussant: Barth Riley

Thursday, September 2, 2010

Plenary Session 2

1. Comparing and Integrating Classical and Modern Measurement Models
2. David Andrich (University of Western Australia): Cronbach Alpha and the Rasch model indices of reliability in the presence of skewed distributions of subscales
3. Carl Granger and Paulette M. Niewczyk (University at Buffalo, Amherst, NY): Combining Classical and Rasch Techniques in Developing a Pediatric Measure

Discussant: Michael Dennis (Chestnut Health Systems, Normal, IL)

2A.1: Interpreting DIF

1. I-Chan Huang*, Pey-Shan Wen, John Nackashi, Dennis Revicki, Elizabeth Shenkman: Using different techniques to detect differential item functioning in pediatric quality of life between children with and without special health care needs: do they make differences?
2. Curt Hagquist*: Resolving Differential Item Functioning using principles of equating
3. Adam C. Carle*: A new method for evaluating and translating measurement bias practical impact: A description and example using the Children with Special Health Care Needs Screener.

Chair: Robert Massof, Discussant: Steven Humphry

2A.2: CAT Applications

1. Stephen F. Butler*, Ryan A. Black*: Developing a Computerized Adaptive Testing Version of the ASI-MV®
2. Barth B. Riley*, Michael L. Dennis, Kendon J. Conrad: CAT Item Selection, Person Fit and Detection of Atypical Suicide

Chair: Thomas F. Hilton, Discussant: Wim van der Linden

2A.3: Substance Abuse Measurement

1. Zhiqun Tang*, Ronald E. Claus, Robert G. Orwin, Carlos Arieira, Wendy Kissin: Combining CTT and Rasch Modeling to Create a Scale for Gender Sensitive Programming in Substance Abuse Treatment: Dealing with Small Sample Size and Potential Multi-Dimensional Issues
2. Jason E. Chapman*, Ashli J. Sheidow, Scott W. Henggeler: Rasch-Based Evaluation of a Test for Measuring Longitudinal Changes in Therapist Knowledge of an Evidence-Based Treatment for Adolescent Substance Use
3. Gopika Chandra*, Thomas Lyons, Edward Mensah, Jacek L. Ubaka: Does Stages of Change Readiness and Treatment Eagerness Scale (SOCRATES) Measure Transtheoretical Model of Change for methamphetamine users?

Chair: Jessica Mazza, Discussant: Brian Rush

Rasch-related Coming Events

Oct. 22 - Nov. 19, 2010, Fri.-Fri. Online course: Rasch - Further Topics (intermediate) (M. Linacre, Winsteps), www.statistics.com

Oct. 27-30, 2010, Wed.-Sat. ISOQOL 17 International Society for Quality of Life Research, London, England www.isoqol.org

Nov. 26, 2010, Fri. In-person workshop: "Modelos de Rasch en Administración de Empresas" IUDE-University of La Laguna. Tenerife. Canary Islands (Spanish), www.institutos.ull.es/view/institutos/iude/Inicio/es

Dec. 1-3, 2010, Wed.-Fri. In-person workshop: Introduction to Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Dec. 6-8, 2010, Mon.-Wed. In-person workshop: Intermediate Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Jan. 7 - Feb. 4, 2011, Fri.-Fri. Online course: Rasch - Core Topics (Winsteps, introductory) (M. Linacre, Winsteps), www.statistics.com

Jan. 26, 2011, Wed. 5th UK Rasch User Group meeting, Warwick, UK www.rasch.org.uk

Feb. 28 - June 24, 2011, Mon.-Fri. Online course: Advanced course in Rasch Measurement of Modern Test Theory (Andrich, Marais, RUMM2030), www.education.uwa.edu.au

March 4 - April 1, 2011, Fri.-Fri. Online course: Rasch - Test Equating and Linking (Winsteps, intermediate) (M. Linacre, Winsteps), www.statistics.com

Apr. 8-12, 2011, Fri.-Tues. AERA Annual Meeting, New Orleans, LA, www.aera.net

April 29 - May 27, 2011, Fri.-Fri. Online course: Rasch - Core Topics (Winsteps, introductory) online course (M. Linacre, Winsteps), www.statistics.com

June 24 - July 22, 2011, Fri.-Fri. Online course: Many-Facets Rasch Measurement (Facets, intermediate) (M. Linacre, Facets), www.statistics.com

July 4-5, 2011, Mon.-Tues. International Workshop on Patient Reported Outcomes and Quality of Life, Paris, France, www.lsta.upmc.fr/mesbah/PROQOL/

Jan. 9-15, 2012, Mon.-Wed. In-person workshop: Introductory Rasch course (Andrich, RUMM2030), Perth, Australia, www.education.uwa.edu.au

Jan. 16-20, 2012, Mon.-Wed. In-person workshop: Advanced Rasch course (Andrich, RUMM2030), Perth, Australia, www.education.uwa.edu.au

Jan. 23-25, 2012, Mon.-Wed. Fifth International Conference on Probabilistic Models for Measurement in Education, Psychology, Social Science and Health, Perth, Australia, www.education.uwa.edu.au

Post-Lunch Plenary Session 2
1. Mark Wilson (University of California, Berkeley, CA): Latent Growth Item Response Models
2. Bruno Zumbo (University of British Columbia, Vancouver, BC): Measurement Validity and Validation: A Meditation on Where We Have Come From and the State of the Art Today

2B.1: Applications of Rasch Measurement (Map, DIF, Person Fit) Using the Global Appraisal of Individual Needs (GAIN)
1. Kendon J. Conrad, Karen M. Conrad, Barth B. Riley, Rod Funk, Jessica Mazza*, Michael L. Dennis: Dimensionality, Hierarchical Structure, Validity, and Age Generalizability of an Externalizing Disorders Measure
2. Kendon J. Conrad, Barth B. Riley, Karen M. Conrad*, Ya-Fen Chan, Michael L. Dennis: Validation of the Crime and Violence Scale (CVS) against the Rasch Model including Differences by Gender, Race and Age In Assessing Criminality
3. Kendon J. Conrad*, Karen M. Conrad, Barth B. Riley, Rod Funk, Michael L. Dennis: Validation of Rasch Person-Fit Statistics in Biopsychosocial Screening

Chair: Michael L. Dennis, Discussant: Brian Rush

2B.2: Measurements in Health Care
1. Mary Slavin*, Alan M Jette*, David Tulsky, Pamela Kisala, Pengsheng Ni: A Contemporary Spinal Cord Injury Physical Function Outcome Instrument
2. J. Kay Waid-Ebbs*, Pey-Shan Wen, Shelley Heaton, Craig Velozo, Neila J. Donovan: Using Rasch Analysis to examine the psychometric properties of the BRIEF-A on individuals with TBI
3. Adam C. Carle*, Stephen J. Blumberg, Charlie Poblenz: Internal Psychometric properties of the Children with Special Health Care Needs Screener

Chair: Dave Cella, Discussant: Allen Heinemann

2B.3: Cautionary Tales
1. Michael Fendrich*, Ozgur Avci, Laura Otto-Salaj: Refining an HIV Knowledge Measure: What Don't We Know about Don't Know Responses?
2. Karen M. Schmidt*: More is NOT Better: Obtaining Ideal Rescoring Combinations for Lengthy Rating Scales
3. Joan E. Broderick*, Arthur A. Stone, Kevin Weinfurt; The Ecological Validity of Recall PROs: One Size Does Not Fit All

Chair: Craig Velozo, Discussant: Robert Massof

Workshops
1. David Andrich (University of Western Australia): RUMM Workshop
2. Stacie Hudgens (PsyMes Consulting, LLC., Chicago, IL): Winsteps Workshop
3. Mark Wilson (University of California, Berkeley, CA): Fitting Latent Growth Item Response Models with ConQuest

2C.1: Improving and Reducing Measures
1. P. Birjandi, Parisa Daftarifard, Agustin Tristan*: Scaling of the IETLS. The Case of Iran
2. Glenn A. Phillips*, Lei Chen, Joseph Johnston, Virginia Stauffer, Bruce J. Kinon, Haya Ascher-Svanum, Sara Kollack-Walker, Dieter Naber: Factor analysis and item response theory analysis of the Subjective Wellbeing Under Neuroleptic Treatment scale: Suggestion for item reduction
3. Allen Heinemann*, Susan Magasi, Rita Bode, Joy Hammel, Dustin Williams: Measuring Community Participation

Chair: Bryce Reeve, Discussant: Bruno Zumbo

2C.2: Generalizing Measurement
1. Ryan A. Black*, Stephen F. Butler*: Examining Dimensionality in the SOAPP-R
2. Richard P. Moser*, Brad Hesse, Abdul Shaikh, Paul Courtney, Gordon Willis: The Grid-Enabled Measures (GEM) Database: A research tool that uses Web 2.0 capabilities to facilitate the use of standardized measures and sharing harmonized data
3. Ning Yan Gu*, Trevor G. Bond, Benjamin M. Craig: Evaluating the Measurement Properties of an Augmented EQ-5D Using the US National Representative Sample

Chair: Karon Cook, Discussant: Carl Granger

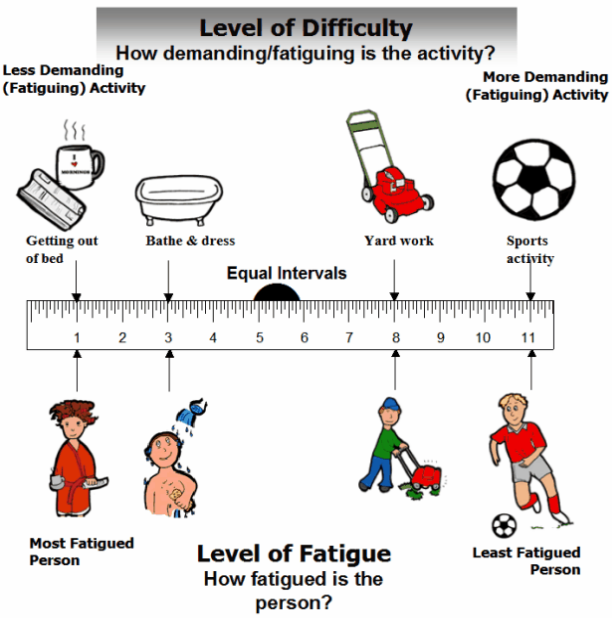

Figure: Measurement map of Fatigue. (Mallinson, 2001)

Mallinson T. (2001) Measuring the Impact of Fatigue on Everyday Activities during Chemotherapy. Paper presented at COMET, Chicago.

2C.3: Status of evaluation in education and health in Latin America
1. Margarita Pena*: National Exams in Colombia
2. Andres Sanchez-Moguel*, Hector Robles*: National Exams in Mexico
3. Aura Nidia Herrera*, Agustin Tristan*: Psychometrics in Education and Health in Latin America
4. Agustin Tristan*: Scale development of a nursing program of cultural home care for elderly adults, using the Rasch model

Chair: Jessica Mazza, Discussant: Ken Conrad

Poster Session
1. Isam Atroshi*, Per-Erik Lyrén, Christina Gumesson: Responsiveness of the 6-item CTS symptoms scale in carpal tunnel syndrome
2. Amaia Bilbao, Jose M. Quintana*, Antonio Escobar, Carlota Las Hayas, Miren Orive: Validation of a proposed WOMAC short form for patients undergoing total hip replacement
3. Shu-Ren Chang*, Gene A. Kramer*, Chien-Lin Yang*, Tsung-Hsun Tsai*, Shu-Mei Lien*: The impact of scored pretest items on proficiency/below proficiency decisions
4. Lei Chen*, Glenn Phillips, Joseph Johnston, Virginia Stauffer, Bruce Kinon, Haya Ascher-Svanum, Sara Kollack-Walker, Paul Succop, Dieter Naber: Relationships among Multiple Outcome Measures in the Study of Schizophrenia
5. Desiree A. Crevecoeur-MacPhail*, Richard A. Rawson*, Ana Ceci Myers*, Loretta Ransom*, Nancy Diep*: Inside the Black Box of Treatment: Using Encounters to Assess Outcomes and Program Performance
6. Doris Howell*, Margaret Fitch, Debra Bakker, Esther Green, Jonathan Sussman, Diane Doran, Tala Chulak, Samantha Mayo, Shan Mohammed, Charlotte Lee: Patient-Centered Outcomes in Cancer: Reaching Consensus on a Core Set for Monitoring Clinical Care Quality
7. Jose M. Quintana*, Amaia Bilbao, Nerea Fernandez de Larrea, Marisa Bare, Eduardo Briones: Identification of variables of adverse evolution at the hospital ER in patients with COPD exacerbation
8. Jose M. Quintana*, Amaia Bilbao, Nerea Gonzalez, Iratxe Lafuente, Cristobal Esteban: Single questions for the evaluation of patients with acute COPD exacerbations
9. Corrie L. Vilsaint*, Melvin N. Wilson, Thomas J. Dishion: Measure of Drug and Alcohol Use Consequences: Differential Item Functioning by Ethnicity

Friday, September 3, 2010

PROMIS Panel
1. Richard Gershon (Northwestern University Feinberg School of Medicine, Chicago, IL)
2. Seung Choi (Northwestern University Feinberg School of Medicine, Chicago, IL)
3. Dave Cella (Northwestern University Feinberg School of Medicine, Chicago, IL)

Discussant: Bryce Reeve (Applied Research Program National Cancer Institute, Bethesda, MD)
Moderator: Barth Riley (Lighthouse Institute of Chestnut Health Systems, Oak Park, IL)

Plenary Session 4. Topic: The Future of Measurement
1. Richard Gershon (Northwestern University Feinberg School of Medicine, Chicago, IL)
2. Wim van der Linden (CTB/McGraw-Hill, Monterey, CA)
3. Mark Wilson (University of California, Berkeley, CA)
4. Bruno Zumbo (University of British Columbia, Vancouver, BC)
5. Karon Cook (University of Washington, Seattle, WA)

Moderator: Ann Doucette (George Washington University, Washington, DC)

Measurement, test construction and data analysis resources

Associate Professor Margaret Wu and Professor Ray Adams have kindly made the following publication available for download from www.edmeasurement.com.au - *Learning Corner* - to people interested in learning more about measurement.

Wu, M. & Adams, R. (2007). ***Applying the Rasch model to psycho-social measurement: A practical approach.*** *Educational Measurement Solutions, Melbourne.*

Introduction

Chapter One: What Is Measurement?

Chapter Two: An Ideal Measurement

Chapter Three: Developing Tests From IRT Perspectives – Construct And Framework

Chapter Four: The Rasch Model (The Dichotomous Case)

Chapter Five: The Rasch Model (The Polytomous Case)

Chapter Six: Preparing Data For Rasch Analysis

Chapter Seven: Item Analysis Steps

Chapter Eight: How Well Do The Data Fit The Model?

They recommend that you download the Introduction chapter first to see which of the other chapters may be of interest. As a general guide, the first two chapters are useful as an introduction to measurement. Chapter three considers the development of tests from a measurement perspective. Chapters four and five introduces the Rasch model, one particular item response theory model, that is used to analyze measurement data. Chapters six to eight provide guidelines on how to prepare and conduct your own Rasch analysis of test data.

A Survey of "Advances in Rasch Measurement, Volume One"

edited by Mary L. Garner, George Engelhard, Jr., William P. Fisher, Jr., and Mark Wilson
JAM Press, Minnesota, 2010 - www.jampress.org
$69 Hard Cover (ISBN 978-1-934116-06-7), $57 Soft Cover (ISBN 978-1-934116-07-4)

Ch.	Title	Authors	P.	T.	C.	G.	N.	Analytical model	Rasch software
1	The Rasch Model and Additive Conjoint Measurement	Van A. Newby, Gregory R. Conner, Christopher P. Grant, and C. Victor Bunderson	11	S	T	No	No	Dichotomous	x
2	Reducible or Irreducible? Mathematical Reasoning and the Ontological Method	William P. Fisher, Jr.	33	F	T	Y	No	x	x
3	Using Paired Comparison Matrices to Estimate Parameters of the Partial Credit Rasch Measurement Model for Rater-Mediated Assessments	Mary L. Garner and George Engelhard, Jr.	19	F	P	No	Y	PCM - MFRM	SP
4	A Family Approach to Assessing Fit in Rasch Measurement	Richard M. Smith and Christie Plackner	22	N	P	Y	Y	Dichotomous	IPARM
5	Plausible Values: How to Deal with Their Limitations	Christian Monseur and Raymond Adams	24	S	P	No	Y	D-MD	ConQuest
6	The Practical Application of Optimal Appropriateness Measurement on Empirical Data Using Rasch Models	Iasonas Lamprianou	23	F	P	No	Y	Dichotomous	Analysis
7	Considerations About A Posteriori Estimation in Adaptive Testing: Adaptive A Priori, Adaptive Correction for Bias, and Adaptive Integration Interval	Giles Raiche and Jean-Guy Blais	22	S	P	Y	Y	Dichotomous	SP
8	Features of the Sampling Distribution of the Ability Estimate in Computerized Adaptive Testing According to Two Stopping Rules	Jean-Guy Blais and Giles Raiche	12	F	P	No	Y	Dichotomous	SP
9	Local Independence and Residual Covariance: A Study of Olympic Figure Skating Ratings	John M. Linacre	19	N	A	Y	Y	RSM	Winsteps
10	Thinking About Thinking - Thinking About Measurement: A Rasch Analysis of Recursive Thinking	Ulrich Muller and Willis F. Overton	19	N	A	Y	Y	Dichotomous	Bigsteps
11	Using Adjusted GPA and Adjusted Course Difficulty Measures to Evaluate Differential Grading Practices in College	Dina Bassiri and E. Matthew Schulz	16	F	A	Y	Y	RSM	Winsteps
12	Constructing One Scale to Describe Two Statewide Exams	Insu Paek, Deborah G. Peres, and Mark Wilson	21	F	P	Y	Y	D-MD (MRCMLM)	ConQuest
13	Development of Scales Relating to Professional Development of Community College Administrators	Edward W. Wolfe and Kim E. Van Der Linden	24	N	A	Y	Y	D-MD (MRCMLM)	ConQuest

14	An Application of the Multidimensional Random Coefficients Multinomial Logit Model to Evaluating Cognitive Models of Reasoning in Genetics	Edward W. Wolfe, Daniel Hickey, and Ann C. H. Kindfield	16	F	A	Y	Y	D-MD (MRCMLM)	ConQuest
15	Mapping Multiple Dimensions of Student Learning: The GradeMap Program	Cathleen A. Kennedy and Karen Draney	22	N	P	Y	No	RSM-MD (MRCML)	ConstructMap
16	A Comparative Analysis of the Ratings in Performance Assessment using Generalizability Theory and Many-Facet Rasch Measurement	Sungsook C. Kim and Mark Wilson	24	F	P	Y	Y	MFRM	ConQuest
17	Reliability of Performance Examinations: Revisited	Mary E. Lunz and John M. Linacre	14	N	P	Y	Y	MFRM	Facets
18	Comparison of Single- and Double-Assessor Scoring Designs for the Assessment of Accomplished Teaching	George Engelhard, Jr. and Carol Myford	27	N	A	Y	Y	MFRM	Facets
19	Exploring Differential Item Functioning (DIF) with the Rasch Model: A Comparison of Gender Differences on Eighth Grade Science Items in the United States and Canada	Tasha Calvert Babiar	29	N	A	Y	Y	MFRM	Facets
20	Using Classical and Modern Measurement Theories to Explore Rater, Domain, and Gender Influences on Student Writing Ability	Ismail S. Gyagenda and George Engelhard, Jr.	32	N	A	Y	Y	MFRM	Facets
21	Multidimensional Models in a Developmental Context	Yiyu Xie and Theo L. Dawson	15	F	P	Y	Y	D-MD (MRCMLM)	ConQuest
22	Developing a Domain Theory	C. Victor Bunderson	38	N	T	Y	No	MFRM	Facets
23	Towards a Domain Theory in English as a Second Language	Diane Strong-Krause	13	F	A	Y	Y	MFRM	Facets
24	The Role of Design Experiments and Invariant Measurement Scales in the Development of Domain Theories	C. Victor Bunderson and Van Newby	33	N	T	Y	No	x	x
25	Children's Understanding of Area Concepts: Development, Curriculum and Educational Achievement	Trevor G. Bond and Kellie Parkinson	24	F	A	Y	Y	Dichotomous	Quest
26	Comparing Decalage and Development with Cognitive Development Tests	Trevor G. Bond	19	N	A	Y	No	Dichotomous	Quest
27	Concrete, Abstract, Formal and Systematic Operations as Observed in a Piagetian Balance-beam Task Series	Theo L. Dawson, Eric A. Goodheart, Karen Draney, Mark Wilson, and Michael L. Commons	19	F	P	Y	Y	Dichotomous - Saltus	Quest + Saltus

P. = Page count
T. = Technical level: Novice - Familiar - Specialist
C. = Content: Theory - Practice - Application
G. = Graphs and/or pictures: Yes - No
N. = Numerical tables: Yes - No

IRT and Confusion about Rasch Measurement

The International Journal of Educational and Psychological Assessment contains papers relating to Classical Test Theory (CTT), Item Response Theory (IRT) and Rasch measurement. Unfortunately these papers present a confusing perspective on Rasch measurement. For instance,

1. Cutoff Scores: The Basic Angoff Method and the Item Response Theory Method. *Niclie Tiratira.*
tijepa.books.officelive.com/Documents/article5v1.pdf

This articles employs Bigsteps and Winsteps, without citing the sources, in an IRT context, and reports results with no sense of the fingernails-on-the-blackboard dissonance infit and outfit have relative to IRT.

2. The Measurement of Change in Groups and Individuals With Particular Reference to the Value of Change Scores: A New IRT-Based Methodology for the Assessment of Treatment Effects. *Jörg A. Prieler and John Raven.*
tijepa.books.officelive.com/Documents/A3V3_TIJEPA.pdf

This article asks and answers the following questions:

Question 1. Does the portrayal of parallel item characteristic curves (ICCs) in Rasch computer analysis output mislead many into thinking the item difficulty order is independent of ability?

Their answer: Yes. *But the Rasch answer:* No. Rasch analysis deliberately constructs item difficulties which are as independent as statistically possible of ability.

Question 2. Do the crossing ICCs in a 3-parameter IRT analysis indicate that the appropriate model for these data is one that describes the interactions making item and person estimates dependent on one another?

Their answer: Yes. *But the Rasch answer:* No. Crossing ICCs are never an appropriate model (except for some polytomous models.) Construct validity demands that the item difficulty hierarchy is invariant across person abilities.

Question 3: Are the "most popular versions of Item Response Theory" often loosely referred to as "the Rasch model"?

Their answer: Yes. *But the Rasch answer:* No. IRT is a descriptive statistical methodology originated by Frederic Lord. Rasch analysis is a prescriptive measurement methodology originated by Georg Rasch. One of Lord's IRT models resembles a Rasch model.

Comment: What chaos! The force of the academically-dominant IRT paradigm's influence is truly impressive: it is able to make people see things that don't exist, and to let them ignore existing things that don't fit with their preconceptions. How will these kinds of misconceptions ever be corrected? Will it all come out in the wash at some point down the road when invariance, sufficiency and additivity are demanded as basic elements of credible psychometric measurement?

William P. Fisher, Jr.

Rasch Measurement Transactions
www.rasch.org/rmt
Editor: John Michael Linacre

SIG Chair: Michael Young
Secretary: Kenneth Royal
Program Chairs: Leigh Harrell & Stephen Jirka
SIG website: www.raschsig.org

The ***Albertina Rasch Dancers*** demonstrate equal-interval scaling.
"The photo above shows the Albertina Rasch Dancers in costume for the *Florenz Ziegfeld* produced musical *Rio Rita* in 1927. They are credited to photographer *Florence Vandamm."*
As displayed on http://songbook1.wordpress.com/pp/fx/features-2-older-2/albertina-rasch-dancers/