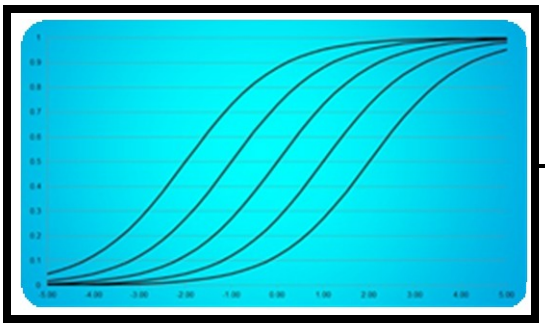

RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 27 No. 4 Spring 2014 ISSN 1051-0796

From Ordinality to Quantity

There are three essential attributes to measuring. Rasch (1968, 1977, 2001) addressed them in various papers, and they were explicated further by Stone & Stenner (2014).

1. Comparison is primary: To compare is to distinguish: a = b; a < b; a > b.
2. Order follows: If a < b, b < c, then a < c. Transitivity results, but a test for variable monotonicity is required to establish valid order.
3. Equal Differences require a Standard Unit. A Standard Unit is established either by a personage with power (king, pope), or by agreement arising from data and consensus (science).

The process is developmental. Nothing in measuring arises full-blown. Progress occurs by steps of continued understanding resulting from intuition, reason, and improved instrumentation. We designate this encompassing process *Measuring Mechanisms*.

The Mohs scale of hardness is based on a scratch test mechanism. Ten key values range from talc (#1) to diamond (#10). Comparison and order are satisfied. Equal differences are not satisfied inasmuch as the difference from 7 to 8 is not the same as the difference of 3 to 4. Unequal differences occur across all the scale values. Such a scale is termed ordinal by Stevens and others embracing a level of measurement schema. While the Mohs test is a well-recognized ordinal scale, the Vickers test is another matter.

The Vickers scale is used in engineering and metallurgy operating via two different mechanisms; indentation hardness and rebound hardness. The former is determined from a microscopic device equipped with a micrometer for measuring permanent deformation of the material tested. The indentation made from an experimental indenter is carefully measured resulting in a linear numerical value that is amenable to mathematical operations.

Rebound hardness is measured from the upward "bounce" of a carefully engineered hammer descending from a fixed height (see Stone and Stenner's 2014 explication of Rasch's ashtray dropping experiment. There is similarity, but Rasch employs a purely qualitative approach to make his case). The Vickers experiment uses a scleroscope to provide a precise linear measure of rebound height. Two other related scales for measuring rebound hardness are the Leeb rebound hardness test and the Bennett hardness scale.

It would be interesting to apply and especially to compare the "hardness" of the ten key minerals on Mohs scale to an exact measure of rebound from applying the Vickers measuring mechanism. One might expect this "predicted" scale to reflect an outcome similar to a Winsteps map of items. Talc (#1) and gypsum (#2) might be calibrated close to each other because both can be scratched with the

Table of Contents

From Ordinality to Quantity (Stone & Stenner)…… 1439

IMEKO Video Presentations Available Online (Fisher)..……..... 1440

3PL, Rasch, Quality-Control and Science (Linacre)……..... 1441

Message from Rasch SIG Chair (O'Neil).... 1444

Teaching Rasch: Students' Confusion with Levels of Agreement (Boone)…… 1445

IOMW and AERA Presentations …… 1446

fingernail. Topaz (#8), corundum (#9), and diamond (#10) might be somewhat close together at the other end of the scale inasmuch as these three represent gemstones. The remaining minerals might be scattered, or close to one another. The map would indicate "measured" differences in hardness for the key Mohs indicators whereby some of the ten might be greatly separated and others close or similar. Extensive lists of gemstone hardness indicate that most have Mohs values of 7 and higher.

CIDRA Precision Services LLC (2012) published data on mineral hardness showing the relationship of the Mohs scale to the Vickers. A plot of their data is given in Figure 1 modeled by a power curve with an $R^2 = 0.9857$. The first four values of the Mohs scale are found between zero and five hundred on the Vickers scale while the last three values on the Mohs scale show adjacent differences of about five hundred.

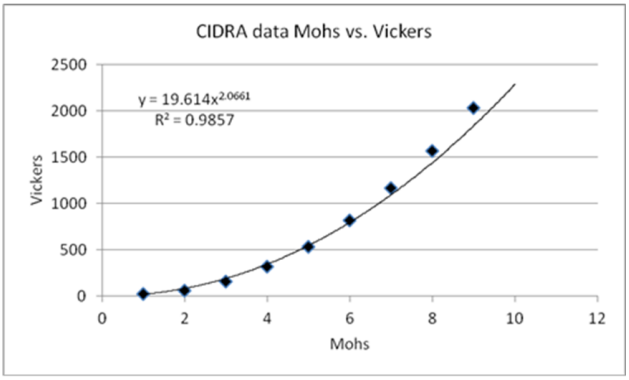

Figure 1: CIDRA data for Mohs vs. Vickers

Oppenheimer (1956) emphasized that "Analogy is an instrument in science". He identified analogies as vital and indispensable to conducting science. Hardness measured by the Vickers test for indentation or upward bounce progresses analogously beyond (1) comparison and (2) order by providing measures of hardness whereby (3) differences are expressed on an equal interval linear scale.

So, some mechanisms yield merely ordinal relations and others quantitative relations (i.e. homogeneous differences up and down the scale). One explanation for repeated failure to engineer a mechanism sensitive to variation in homogeneous differences for an attribute is that the attribute is ordinal, however, if the history of science is any guide it may take a century or two of effort before we confidently conclude that an attribute is in some fundamental sense merely ordinal.

References

CIDRA Precision Services LLC (2012). Wallingford, CT.

Oppenheimer, R. (1956). Analogy in science. *American Psychologist, 11*, 127-135.

Rasch, G. (1964). Objective comparisons. *Lectures at the UNESCO Seminar*, Voksenåsen, Oslo.

Rasch, G. (1977). On specific objectivity. *Danish Yearbook of Philosophy, 14*, 58-94.

Rasch, G. (2001). *Rasch lectures*. Lina Wøhik Olsen & Sven Kreiner (Eds.). Copenhagen: Copenhagen Business School.

Stone, M. & Stenner, A. J. (2014). Comparison is key. *Journal of Applied Measurement*, In press.

Mark Stone and Jack Stenner

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: Tim O'Neil
Secretary: Kirk Becker
Program Chairs: Kelly Bradley & Jessica Cunningham
Rasch SIG website: www.raschsig.org

IMEKO Video Presentations Available Online

The 2014 International Measurement Confederation (IMEKO) TC-1/TC-7/TC-13 Joint Symposium will be held in Funchal, Madeira Island, Portugal, on September 3-5, 2014. The symposium theme this year is "Measurement Science Behind Safety and Security." Topics of interest include fundamentals of measurement science, uncertainty evaluation, measurement education, and applications in physics, engineering, psychology, the social sciences, health care, the life sciences, and in everyday activities. For more information and abstract submission guidelines, see http://www.imekotc7-2014.pt/.

Videos of some of the presentations made at last September's 15th IMEKO Joint Symposium in Genoa, Italy, are available on the Spectronet web site at: http://spectronet.de/de/vortraege_bilder/vortraege_2013/15th-joint-international-imeko-tc1tc7tc13-symposiu_hlms3467.html.
Rasch-oriented presentations listed there are by Nikolaus Bezruczko, Fabio Camargo, William Fisher, David Torres Irribarra, Luca Mari, Bob Massof, Andy Maul, Jack Stenner, and Mark Wilson. Photos of participants are available at the bottom of the page.

Videos and slides from selected presentations made at the 2011 IMEKO Joint Symposium in Jena, Germany are available at: http://spectronet.de/de/vortraege_bilder/vortraege_2011/14th-joint-international-imeko-tc1tc7tc13-symposiu_gs4vctcp.html.
Unfortunately, only one of the several Rasch-oriented presentations is available online (Fisher's). RMT readers will, however, find the introductory talks by Linss and Ruhm, and Ruhm's tutorial on error models, of particular interest. Photos at the bottom of the page include shots of Mark Wilson, Stefan Cano, Thomas Salzberger, Jack Stenner, and William Fisher.

Audio recordings and slides from presentations made at the 2010 IMEKO joint Symposium in London are available at: http://spectronet.de/de/vortraege_bilder/vortraege_2010/13th-imeko-tc1%E2%80%93tc7-joint-symposium-london_gdpw31nj.html.
Presentations RMT readers may find of particular interest include those by Ludwik Finkelstein, Luca Mari, Karl Ruhm, Klaus-Dieter Sommer, Eric Benoit, Philip Thomas, and William Fisher. Photos of these and other speakers, including Nikolaus Bezruzcko, can be found at the bottom of that web page.

A roundtable on the International Vocabulary of Measurement (the VIM) was held at the 2009 IMEKO World Congress in Lisbon, Portugal: (http://spectronet.de/de/vortraege_bilder/vortraege_2009/imeko-xix-world-congress-lisbon_fzbh9xc2.html).
Presenters addressing the expanded scope of the recently released third edition of the VIM into psychology and the social sciences included Finkelstein, Mari, Pavese, Ehrlich, and Morawski. Other presentations of interest shown on that page include those by Rossi, Thomas, and others.

Slides from the 2008 IMEKO joint symposium in Annecy, France, are available at: http://spectronet.de/de/vortraege_bilder/vortraege_2008/12th-imeko-tc1-tc7-annecy_fknxgogl.html.
Of particular interest here will be presentations by Finkelstein, Mari, Pavese, Ruhm, Guerra, Rossi, Goodman, Eugene, and Fisher.

William P. Fisher, Jr.

3PL, Rasch, Quality-Control and Science

Bergan (2010) states "Science entails the development of theories involving hypotheses that can be tested through the examination of data." and also "Science proceeds in exactly the opposite fashion to the Rasch approach to model selection." Bergan describes an analysis by Christine Burnham of a 44-item math test administered to 3,098 Grade 5 students. Since "An important aspect of the IRT approach is the selection of an IRT model to represent the data", the data were analyzed using 1-PL [Rasch], 2-PL and 3-PL models. Bergan's conclusion "is that for this assessment the 3PL model is preferred over the 1PL and 2PL models because the 3PL model offers a significant improvement in the fit of the model to the data over the alternative models. In other words, the additional parameters estimated in the 3PL model are justified because they help provide a better fit to the data." From the standpoint of descriptive statistics, the discussion is over, but there is more to measurement than mere description.

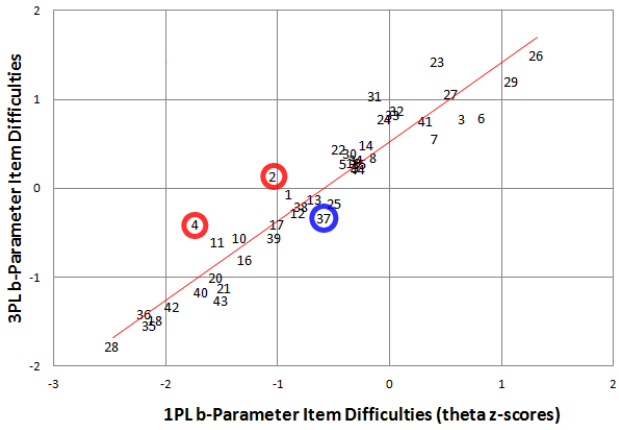

Figure 1. Plot of item difficulties from Bergan, Table 2. Person thetas are N(0,1).

Let's look more closely at these analyses. Bergan helpfully reports the item difficulties, *b,* according to 1PL and 3PL in his Table 2. These are plotted in our Fig. 1. The person ability *theta* distribution is stated to be constrained to N(0,1) in both 1PL and 3PL analyses. In the Figure, items 2 and 4 have the highest 3PL pseudo-guessing and item 37 has the lowest discrimination. Bergan attributes the average 0.5 z-score (unit-normal deviate) difference between the 1PL and 3PL estimates to the 3PL pseudo-guessing lower asymptote, *c,* which averages c=0.22 according to Bergan's Table 3. in particular, Bergan

identifies Item 2 as more accurately estimated by 3PL than by 1PL because its pseudo-guessing lower asymptote, *c,* corresponds to a probability of success of 45% (c=0.45). Similar reasoning would apply to item 4 which has the highest lower asymptote, (c=0.50). Surely we are surprised by the large amount of guessing associated with these items that are targeted near the average ability level of the students. On the other hand, item 26 is the most difficult item. We could expect this item to provoke guessing by the lowest third of the sample, but its pseudo-guessing is a relatively low c=0.17.

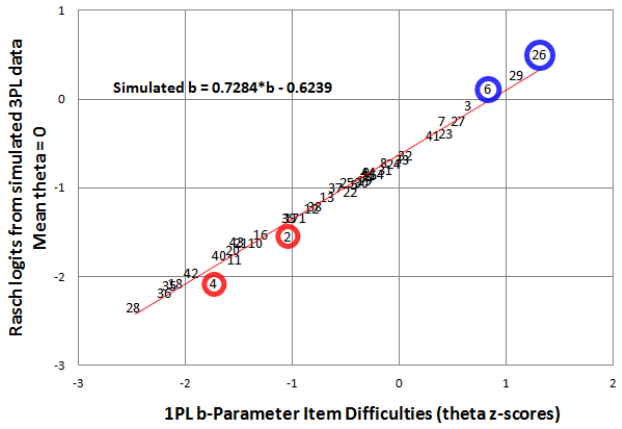

Figure 2. Plot of Rasch item difficulties estimated from data simulated with Bergan's 3PL estimates plotted against Bergan's 1PL estimates.

In order to verify that the 1-PL analysis does correspond to a standard Rasch analysis, I simulated data using Bergan's 3-PL parameter estimates and an N(0,1) theta distribution. Rasch *b*-parameters for these data were estimated with *Facets* (chosen because its weighting capabilities allow an exact match in the data to the 3PL ogives and theta distribution). The plot of item difficulties is shown in Fig.2. The noticeable outliers are items 2 and 4 (which have high 3PL pseudo-guessing values) and items 6 and 26 (which have high 3PL discrimination). Overall, this simulation confirms that the reported 1PL analysis reasonably matches a Rasch dichotomous analysis.

More interesting are the fit statistics for the simulated items from the Rasch analysis. All the items have acceptable fit statistics! The most under-fitting item is item 37 (lowest 3PL discrimination) with an outfit mean-square of 1.13. The most over-fitting item is item 9 (which has the highest 3PL discrimination) with an outfit mean-square of 0.89. The infit mean-squares are within the range of the outfit mean-squares. Surprisingly, item 2 (high 3PL pseudo-guessing) only slightly under-fits with an outfit mean-square of 1.09, and item 4 (high 3PL pseudo-guessing) slightly over-fits due to its high 3PL discrimination. Though many simulated responses are flagged by Facets as potential guesses, they are overwhelmed in the simulation by well-behaved data and so are have little influence on the Rasch fit statistics. Surprisingly, if the original data did accord with the estimated 3PL parameters, then those data would also accord with the Rasch dichotomous parameters. Bergan's comment that "In general, in science, the most parsimonious model (i.e. the model involving the least number of estimated parameters) is preferred to represent the data" would motivate the selection of Rasch over 3PL!

This advances us to the next step in any scientific investigation: quality control. A major flaw in 3PL analysis is its lack of quality-control of the data. What about item 2 with its high pseudo-guessing? Bergan admits that there can be bad items but does not describe any attempt to discover if item 2 or any other of the 44 items are bad items. Instead, he quotes Thissen and Orlando (2001) who say "The [Rasch] model is then used as a Procrustean bed that the item-response data must fit, or the item is discarded." The assumption is that item 2 fits the 3PL model and so is a good item (but no item-level fit statistics are reported to support this). The assumption is also that item 2 does not fit the Rasch model and so it would be discarded (again no item-level fit statistics are reported to support this). The simulated evidence suggests that Rasch would keep item 2, but, based on the empirical evidence, item 2 might be discarded by Rasch. Let's see why.

Bergan reports the 3PL parameter estimates in his Table 3. As we might expect, there is no correlation between 3PL item discrimination, *a,* and pseudo-guessing, *c,* and a small positive correlation, r=0.19 , between pseudo-guessing and item difficulty. There is a stronger positive correlation between item discrimination and item difficulty, r=0.33. As items become more difficult, they discriminate more strongly between high and lower performers. We might hypothesize that the more difficult items require technical knowledge of math, such as algebraic symbols, that is not taught to low performers. Thus the increase in item discrimination

with difficulty could be caused by classroom teaching practices.

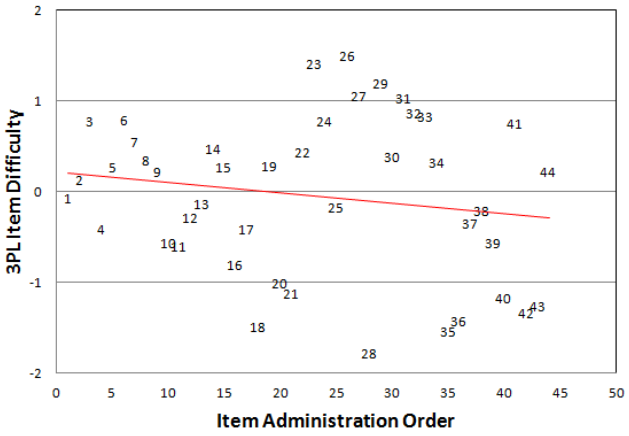

Figure 3. Plot of 3PL *b* item difficulty against item administration order.

However, other correlations are more thought-provoking. Let's assume the usual situation that the 44 items are in the same order in the data as they were during the test administration. Then the correlation between item administration order and item difficulty is r = -0.17. Later items are easier overall than the earlier items. Indeed, Fig. 3 shows us that the easiest items are administered starting at item 18 of the 44 items. This is not disastrous but does contradict the folk wisdom that the easier items should be earlier in order to encourage the lower performers to do their best. We might want to point this out to the test constructors.

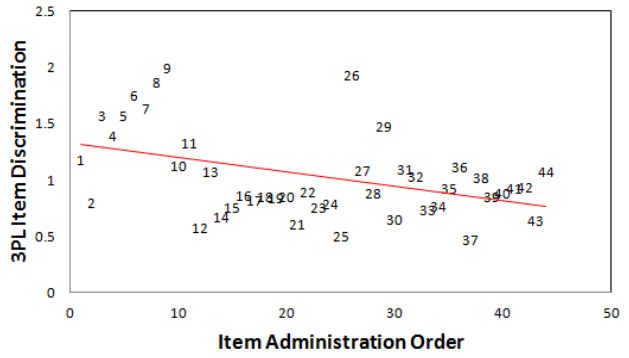

Figure 4. Plot of 3PL *a* item discrimination against item administration order.

The correlation between item administration order and item discrimination is r = -0.42. We would expect a slightly positive correlation. The pattern is shown in Fig. 4. Now we do need to put on our quality-control hats. The first 9 items show a sharp increase in item discrimination. Why? And there is the unusually high discrimination of item 26. 3PL estimation algorithms usually constrain the upper limit of item discrimination. In this estimation, the maximum item discrimination appears to have been constrained to 2.0, so both item 9 (a=2.0) and item 26 (a=1.94) may actually have higher discriminations. 3PL has blindly accepted this pattern of item discrimination. Rasch analysis would flag the items with higher discriminations as over-fitting and perhaps locally dependent.

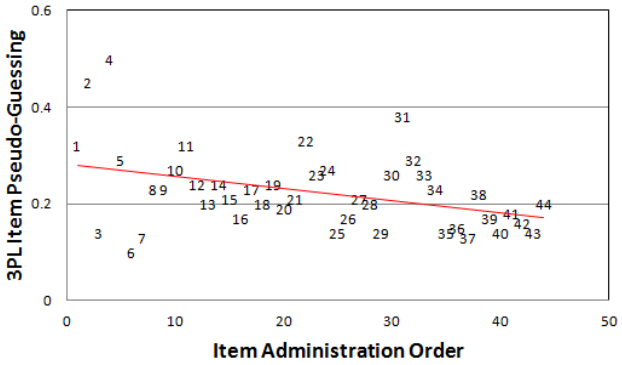

Figure 5. Plot of 3PL *c* item pseudo-guessing against item administration order.

The correlation between item administration order and pseudo-guessing is r = -0.39, when we would expect a small positive correlation. The pattern is shown in Fig. 5. Now the very high pseudo-guessing for items 2 and 4 stands out. There is a definite problem at the start of the test. On the other hand, at the end of the test, when we expect guessing to increase because of time constraints, student tiredness, student frustration, etc., pseudo-guessing is, in fact, decreasing, even though items 41 and 44 are among the more difficult items.

It appears that, if we are really interested in measuring student ability, as opposed to describing a dataset, then we should seriously consider jettisoning items 2-9, 26 and perhaps one or two other items.

Bergan writes, "In the Rasch approach, data that do not fit the theory expressed in the mathematical model are ignored or discarded. In the scientific [IRT] approach, theory is discarded or modified if it is not supported by data." This view of "science" allows problematic data to control our thinking. Rasch takes a pro-active view of science. Every observation is an experiment that requires careful scrutiny. Was the experiment a success or a failure?

Problematic data certainly should not be ignored, and if found to be fatally flawed must be discarded. Otherwise we risk making false inferences that could have severe repercussions throughout the academic careers of these students.

Bergan tells us, "it is expensive and risky to ignore objective data", but that is exactly what has happened in the 3PL analysis. The negative correlations and other potential aberrations in the objective data have been ignored, because the 3PL model has made no demands upon the quality of the data.

Bergan admits that "Adherence to a scientific [IRT] approach does not imply that there are no bad items. Indeed, measurement conducted in accordance with the traditional scientific approach facilitates effective item evaluation and selection." However, here it seems that 3PL does not accord with the traditional scientific approach. It fails to examine the data. It hides problems in the data, and so acts against an effective evaluation. 3PL fails as a tool of Science, but Rasch succeeds.

References

Bergan J.R. (2010) Assessing the Relative Fit of Alternative Item Response Theory Models to the Data. Tucson AZ: Assessment Technology Inc. http://ati-online.com/pdfs/researchK12/AlternativeIRTModels.pdf

Thissen, D., & Orlando, M. (2001) Test Scoring. Mahwah, NJ: Lawrence Erlbaum Associates.

John Michael Linacre

Notable Quote

"The connection between concepts and statements on the one hand and the sensory data on the other hand is established through acts of counting and measuring whose performance is sufficiently well determined."

Broadcast recording for Science Conference, London, September 28, 1941. Published in "Advancement of Science", London, 2, 5. Audible(!) at: http://www.openculture.com/2013/03/listen_as_albert_einstein_reads_the_common_language_of_science_1941.html.

Message from Rasch SIG Chair

Greetings Rasch SIG colleagues,

I hope 2014 finds you well. In lieu of AERA fast approaching, I wanted to share a few brief comments. First, I wanted to acknowledge the efforts our Program Co-Chairs Kelly Bradley and Jessica Cunningham for their work in putting together this year's Rasch Measurement SIG sessions. This year's program includes several paper and poster sessions focusing on everything from instrument development to item and person fit and Rasch-based modeling approaches. I hope you'll make every effort to seek out these events if you'll be able to attend AERA.

I'm also quite excited to report that David Andrich has graciously agreed to speak at this year's business meeting. The title of his talk is "On Rasch Measurement Theory" in which he presents a timely discourse on the work of Georg Rasch:

"*This presentation argues that the contribution and significance of the work of Georg Rasch to the understanding and practice of social measurement is both understated and misunderstood when it is seen primarily as an exercise in response modeling. It is argued that a complex of understandings of his contribution, which entails (i) an a-priori criterion of invariance of comparisons within a specified frame of reference, (ii) rendering this criterion in the form of class of probabilistic models which have sufficient statistics for all parameters, (iii) articulating these models with a body of knowledge of statistical inference, and (iv) anchoring the models in an empirical paradigm of item and test construction, justifies it being referred to as the Rasch Measurement Theory. In making the case, the paper makes some comparisons and contrasts with the principles of Classical Test Theory and Item Response Theory and suggests that both these approaches to social measurement have less of a claim to being called a theory than the approach by Rasch.*"

Hopefully his abstract whets your appetite a bit and will have you marking your calendar accordingly. As with recent history, you can expect the business meeting to be in line with the past few in that I will be providing a brief State of the SIG address prior to introducing David. The meeting is scheduled for Thursday, April 3rd from 6:15pm to 7:45pm. 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.

I also wanted to acknowledge the fact that we have just wrapped up an election which will mark the first implementation of SIG bylaws. Specifically we'll be welcoming three newly elected SIG officers. Of course results are yet to be tallied and delivered, so I'm afraid I have nothing official to report as of this publication. But stay tuned, as we shall know within the month.

Thanks kindly. I look forward to seeing you in Philadelphia!

Tim O'Neil
Rasch SIG Chair

Teaching Rasch Measurement: Students' Confusion with Differing Levels of Agreement

As I have worked with colleagues and students just starting to learn how and why to measure with Rasch there has been a topic that has come up over and over. As I draw a vertical line for a test construct and identify one end of the line as "Easy" and the other end of the line as "Difficult", students/and colleagues have no problem thinking of test items as being possibly of different difficulty. But when I discuss a rating scale survey, there is often confusion how an item can be perhaps "Easier to Agree With" or "Harder to Agree With". Usually I try to draw an analogy to a test and ask my students and colleagues if it makes sense to them that some items of a survey can be easier to agree with as opposed to other items (e.g., a traditional Likert scale). Sometimes I think they get it, but honestly later on, it is clear that it is hard for them to understand this continuum for a survey.

What are some techniques I have used to help them?

Sometimes I will return to a plot of a vertical line for a test, and then instead of labeling the end points with "Easy" and "Difficult" I will label the end points "Less Difficult" and "More Difficult" as a reminder that we are talking about the "Difficulty" of items. Next, I will draw another line for the same test, and then I will label the ends of the line "Easier" and "Less Easy". Even though the phrase "Less Easy" is very awkward, these two newly labeled lines seem to help the learners understand that we are talking about a variable of difficulty. I point out to them that if we use the end terms "Easy" and "Difficult" we are also fine, but sometimes it is easier to grasp the issue, if the same sort of words are used to describe the ends of the variable.

I then move onto a rating scale. I ask students to consider a rating scale with just two possible ratings: "Agree" and "Not Agree". I ask my learners to imagine they are answering a 10 item survey and that they can think of "Agree" as a correct answer, and "Not Agree" as a wrong answer. This seems to help them see a link to a dichotomous test in which items are right/wrong, and that survey items can be of different levels of "difficulty" (but in this case items are of differing levels of "Ease to Agree With"!). This really seems to help them see that a survey with a rating scale can be along a construct.

I continue my work with the 10 item survey by drawing a vertical line and labeling the two endpoints with "Easier to Agree With" and "Harder to Agree With". Then I might ask my students what if the rating scale was "Easier to Disagree With" & "Harder to Disagree With" labeling the end points. Usually they are able to place the words in the right place and they see the same message using both labeling techniques. This activity seems to help them understand that not only is it possible to have a construct with a survey, but they begin to understand that a continuum can be defined with a survey, just as a right/wrong test can define a continuum.
The next step that I take involves a rating scale of "Strongly Agree", "Agree", "Disagree", and "Strongly Disagree". So now I move to surveys in which the rating scale is not dichotomous. Students now seem to "get it" that they could think of this 4 step scale as an "Agree" scale with a scale showing different levels of agreement. Often I will ask them to first re-label the scale with similar words...some will write something like this "Strongly Agree", "Agree", "Agree Less than Agree", "Hardly Agree at All". They understand that even though "Strongly Disagree" does not at first sound like a level of agreement that they can really just think of "Strongly Disagree" as a very low level of agreement.

The next step is for them to draw a line and label the end points with "More Strongly Agree" and "Really Less Strongly Agree". I point out that they could

think of "Really Less Strongly Agree" as "Strongly Disagree". Even though the words are awkward, this seems to work. My point is to help them understand the continuum and not to be tripped up on words that at first blush might seem to involve different issues (e.g., Agree, Disagree).

Now the grand finale is to talk about rating scales in which there is a wide mix of words to describe a rating scale step. My favorite is "Always", "Often", "Sometimes", "Seldom", and "Never". In this case none of the words look like they are linked in meaning based upon a similar word being present in a rating scale step (e.g., a scale of "Very Often", "Often".... or a scale of "Very Important", "Important"...).

In this case we also draw a vertical line, and I make use of the reasoning that I have previously presented. I try to point out that the scale could have been "Often" or "Not Often", and that a line could be labeled with "More Often" and "Less Often", or (very awkward! "More Sometimes" and "Less Sometimes"). I think at the end of the activity I have helped them better understand that a rating scale can be expressed on a line, as one can do for item difficulty. Also the students better understand how to think about the meaning of going up or going down the line of the continuum. The understanding of going up or down the line is very important as they later learn how to interpret person measures and item measures.

William J. Boone
Miami University (Ohio)

"A teacher affects eternity: he can never tell where his influence stops." Henry Brooks Adams

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, 2014, so please make your submission by May 1, 2014 for full consideration. Please email Editor\at/Rasch.org with your submissions and/or ideas for future content.

IOMW 2014 Conference Program
Philadelphia, Pennsylvania
Mon., March 31 – Thur., April 3, 2014

*Program subject to change

A confirmatory Rasch Analysis of rubrics for an assessment of reflective judgment, **Theo Dawson*

A Family of Rater Accuracy Models, **Edward W. Wolfe*

A SOLO Approach to Validity, **Brent Duckor*

A Twenty Year Perspective Navigating Clinical Outcomes Measurement I: Routes, Roadblocks and Chasms, **Jeremy C. Hobart & *Stefan J. Cano*

Applying DIF detection methods to the Nonverbal Accuracy Assessment, **Beyza Aksu*

Applying NOUS to the National Medical Board Exam, **Gregory E. Stone, *Mark Moulton, *Toni A. Sondergeld, & *Kristin L.K. Koskey*

Applying the Rasch Model to Forced Choice Paradigms: How to Model Chance Performance (a.k.a., "Guessing"), **Robert W. Massof*

Are Subscales Compatible with Univariate Measures?, **Robert W. Massof*

Considerations on the Use of Statistical Evidence in the Assessment of the Threshold Order in Polytomous Items in the Rasch Model, **Thomas Salzberger*

Construct Maps as Boundary Objects in the Trading Zone, **Rich Lehrer & *Seth Jones*

Construct Maps as Mediating Objects in Trading Zones and in the Formation of Collective Intelligence, **William P. Fisher*

Constructing Data Modeling Assessments from Concepts and Practices Useful to Students, Teachers, and Assessment Developers, **Mark Wilson*

Determinants of artificial DIF – a study based on simulated polytomous data, **Curt Hagquist & *David Andrich*

Development of the Automated Scoring System for the Writing Section of the NEAT (II), **Hwanggyu Lim*

Diagnostic Opportunities with Distractor-Driven Multiple-Choice Items in the Context of a Physical Science Assessment, **Stefanie A. Wind & *Jessica D. Gale*

Differences in Optimal Response Scales for Measuring Positive and Negative Affect: An Application of the Partial Credit Model, *Monica Erbacher*

Differential Item Functioning on a Measure of Perceptions of Preparation for Teachers, Teacher Candidates, and Program Personnel, *Courtney Tobiassen*

Estimation of Subscales, *Edward W. Wolfe*

Everyone's Rasch Measurement Analyzer (Erma): An R Package for Rasch Measurement Models, *George Engelhard & *Jue Wang*

Examination of the Relationship between Item Difficulty and Item Attributes based on the LLTM, *Lin Ma & *Kathy E. Green*

Examining the TIMSS Items in Mathematics for Item Fit Under the Rasch Model, *Anatoly Maslak*

Explaining It Away: When Theory and Evidence Disagree, *Andrew Galpern*

Exploring Rating Scale Functioning using Rasch measurement theory and Mokken scaling, *Stefanie A. Wind*

Guessing in Rasch Modeling, *Hong Jiao, *Edward Wolfe, & *Tian Song*

Help Me Tell My Story, *Patrick Charles, *Michelle Belisle, *Kevin Tonita, & *Julie Smith*

How invariant are language versions of the same science test? A PISA 2006 case study, *Yasmine El Masri*

Improving the Science Behind Vertical Scaling, *Derek Briggs*

Instrument Development for Measurement of Critical Thinking Skills in Singapore Schools at the Primary 5 and Secondary 3 Levels, *Raymond Chang Chong Fong, *Flora Hoi Kwan Ning, *Laik Woon Teh, & *Helen Hong*

Is Psychological Measurement Possible? *Andrew Maul, *Mark Wilson, & *David Torres Irribarra*

Latent Transition in Geospatial Thinking and Reasoning For Tectonics Understanding, *Qiong Fu, *Alec M. Bodzin, & *Everett V. Smith, Jr.*

Many Regressions to Estimate Subscales, *John Michael Linacre with *Mark Moulton presenting*

Measurement of Analytic Rumination, *Skye Barbic, *Zac Durisko, & *Paul Andrews*

Measurement of Emotional and Behavioral Disorders in Adolescents by Using Latent Classification Models, *Sung Eun Kim & *Seul Ki Koo*

Measuring Academic Growth Contextualizes Text-Complexity Exposure, *Gary L. Williamson*

Metrological Principles Applied to Vision Psychophysics and Generalized to Psychometrics: Uncertainty as the Source of Stochastics in a Deterministic Dynamical System, *Robert Massof*

Modeling for directly mapping construct levels, *David Torres Irribarra & *Ronli Diakow*

Modeling Item Measures: Linear Logistic Test Model Elegance, *Karen Schmidt*

Modeling the Multidimensional Nature of the Nature of Science, *George M. Harrison*

Multilevel Models: Applications to Rating Designs, *Mihaela Ene*

On the Interpretation of Unidimensional Parameter Estimates for Data Assumed to Be Multidimensional, *Steffen Brandt*

Peer assessment of a posters session: an application of the Many-Facet Rasch measurement model, **Christophe Chénier & *Nadine Talbot*

Predictors of Teacher Performance, **Maria Veronica Santelices*

Psychological Attributes, Statistical Models, and the Representational Fallacy, **Joshua McGrane & *Andrew Maul*

Psychometric Wormholes: A General Method for Trading Information between Subspaces in Within-Item and Between-Item Multidimensional Datasets, **Mark Moulton*

Rasch applications to quality-assured measurement with ordinal data, **Leslie R Pendrill*

Reliability and Validity in the National English Ability Test, **Seul Ki Koo, & *Yongsang Lee*

Social Justice and Educational Measurement, **Zak Stein*

Students perception of classroom assessment practices scales: a comparative analysis of the Rasch, rating scale and partial credit models, **Nadine Talbot, *Gilles Raiche, & *Nathalie Michaud*

Testing the multidimensionality of the Inventory of School Motivation in a Dutch student sample, **Hanke Korpershoek*

The Berkeley Assessment System Software: Facilitating Teacher Management of Assessment, **David Torres Irribarra & *Rebecca Freund*

The Roles of Measurement Theory and Substantive Theory in Assessment, **Michael Kane*

Toward A New Measurement Paradigm in Program Evaluation: Adapting Published Instrument Calibrations to Evaluating Workshop Outcomes, **Paula Petry*

Towards a Caring Science: Applying Developmental Theory to Measuring the Moral Construct of Caring in Nursing, **Jane Sumner*

Using Measures and Equated Scales to Build Medicare G-code Modifiers, **Robert W. Massof*

Validation of a Brief Version of the Recovery Self Assessment (RSA-B) for Assertive Community Treatment, **Skye Barbic, *Maria O'Connell, *Larry Davidson, *Kwame McKenzie, & *Sean Kidd*

Validity Revisited, **Jack Stenner*

What is a Valid Writing Assessment? **Nadia Behizadeh & *George Engelhard, Jr.*

What you don't know can hurt you: Missingness and Partial Credit Model estimates, **Sarah Thomas*

Within-Item Multidimensional Modeling of the Heteroscedastic Interactions Between Multiple Constructs, **Nathaniel J. S. Brown*

Yes, the Partial Credit Model is a Rasch Model, **David Andrich*

AERA 2014 Rasch-related Papers
Philadelphia, Pennsylvania
Thur., April 3 – Mon., April 7, 2014

Applying the Rasch Measurement Model to Measure Changes in Colleges Students' Mathematics and Statistics Perceptions, **Letao Sun, University of Kentucky; *Kelly D. Bradley, University of Kentucky; *Michelle L. Smith, University of Kentucky*

Assessing Students' Understanding of the Energy Concept Across Science Disciplines, **Mihwa Park, SUNY; *Xiufeng Liu, University at Buffalo-SUNY*

Comparing Three Estimation Approaches for the Rasch Testlet Model, **Tian Song, Pearson Assessment & Information; *Yi-Hung Lin, University of California-Berkeley*

Computerized Adaptive Testing for Forced-Choice Ipsative Items, **Xue-Lan Qiu, The Hong Kong Institute of Education; *Wen-Chung Wang, The Hong Kong Institute of Education*

Creating a Physical Activity Self-Report Form for Youth Using Rasch Methods, **Christine DiStefano, University of South Carolina; *Russell Pate, University of South Carolina; *Kerry McIver, University of South Carolina-Columbia; *Marsha Dowda, University of South Carolina-Columbia; *Michael Beets, University of South Carolina-Columbia; *Dale Murrie, University of South Carolina-Columbia*

Developing a Short form of the Personal Style Inventory-II with the Rasch Model, **So Young Kim, Korea University; *Sehee Hong, Korea University*

Developing a Universal Metric for Measuring Chinese Language Learning Motivation Among

Heritage Learners, **Mingyang Liu, University of Toledo*

Developing and Validating Measures of Noncognitive Factors in Middle and High School Students, **Rachel Levenstein, University of Chicago; *Courtney M. Thompson, Consortium on Chicago School Research at the University of Chicago; *Camille A. Farrington, University of Chicago*

Development and Validation of a Multidimenisional Measure of Reading Strategy Use, **Diana J. Arya, University of Colorado-Boulder; *Susan Ebbers, University of California-Berkeley; *Andrew Maul, University of Colorado-Boulder; * Alison Gould Boardman, University of Colorado-Boulder; * Janette K. Klingner, University of Colorado-Boulder; *Amy Lynn Boele, University of Colorado-Boulder*

Equating Surveys with Variable Rating Scales, **Zongmin Kang, DePaul University; *Gregory E. Stone, University of Toledo*

Evaluation of the Quality of Nine Item-Fit Statistics of Rasch Model and Statistics Criteria Used in the Northwest Evaluation Association Item Calibration Procedure, **Shudong Wang, NWEA; *Gregg Harris, NWEA*

Examining Rater Effects in Charter School Fund Applications with a Many-Facet Rasch Model, **Wei Xu, University of Florida; *M. David Miller, University of Florida; *Nancy thornqvist, University of Florida*

Expeditionary Learning Implementation Review: Instrument Development, **Sue Leibowitz, University of Massachusetts; *Larry H. Ludlow, Boston College; *Thomas S. Van Winkle, University of Wisconsin-Madison*

Exploring Aberrant Responses Using Person Fit and Person Response Functions, **Angela Adrienne Walker, Emory University; *George Engelhard, University of Georgia; *Kenneth Royal, University of North Carolina-Chapel Hill; *Mari-Wells Hedgpeth, University of North Carolina-Chapel Hill*

Exploring Differential Facet Functioning Models, **Luke Stanke, University of Minnesota; *Mark L. Davison, University of Minnesota*

Improving Item Bank Deficits by Modifying Existing Items: A Nudge Versus a Shove, **Karen A. Sutherland, Pearson VUE; *John A. Stahl, Pearson VUE; * Ada Woo, National Council of State Boards of Nursing*

Investigating the Performance of Person-Fit Measures Under Rasch Multidimensional Models, **Yan Xia, Florida State University; *Insu Paek, Florida State University*

Item Response Model Approaches to Evaluating the Item Format Effects, **In-Yong Park, Yonsei University; * Yongsang Lee, Korea Institute for Curriculum and Evaluation; *Hwang gyu Lim, Korea Institute for Curriculum and Evaluation*

Massachusetts School Classroom Environment Survey: Development and Validation of a Qualitatively Enriched Rasch-Based Instrument to Measure Teacher Practices Within Massachusetts Schools, **Shelagh M. Peoples, Massachusetts Department of Elementary and Secondary Education (MDESE); *Claire Abbott, MDESE; *Elizabeth Davis, MDESE; *Kathleen Marie Flanagan, MDESE; *Jennifer Malonson, MDESE*

Measurement of Teachers' Professional Performance, **Anatoly Andreyevich Maslak, Branch of Kuban State University at Slavyansk-on-Kuban*

Monitoring Rater Facet in a Highland Dance Championship, **Nicole Makas Colwell; * Beyza Aksu Dunya, University of Illinois at Chicago*

Multidimensional Random Coefficients Multinomial Logit Differential Item Functioning (DIF) Decomposition Modeling for a Testlet Item DIF Investigation, **Insu Paek, Florida State University; *Hirotaka Fukuhara, Pearson*

Posterior Predictive Checks and Discrepancy Measures for Polytomous Item Response Theory Models, **Allison Jennifer Ames, University of North Carolina-Greensboro*

Psychometric Evaluation of the Revised Current Statistics Self-Efficacy (CSSE-30) in a Graduate Student Population Using Rasch Analysis, **Pei-Chin Lu, University of Northern Colorado; *Samantha Estrada, University of Northern Colorado; *Steven Pulos, University of Northern Colorado*

Rasch Analysis of Conference Proposal Ratings, **Kelly D. Bradley, University of Kentucky; *Richard Mensah, University of Kentucky*

Rasch Analysis of the Rosenberg Self-Esteem Scale for African American Students, **Courtney Tobiassen, University of Denver; *Kathy E. Green, University of Denver; *Ruth C. L. Chao, University of Denver*

Rasch-Derived Teachers' Emotions Questionnaire, **Kristin L. K. Koskey, The University of Akron; *Renee R. Mudrey-Camino, The University of Akron*

Scoring and Aggregating Data from Scenario-Based Assessments to Recover Learning Progressions, **Peter Van Rijn, ETS Global; *Edith Aurora Graf, ETS; *Paul Deane, ETS*

SimScientists: Interactive Simulation-Based Science Learning Environments, **Matt Silberglitt, WestEd; *Mark Loveland, WestEd; *Edys S. Quellmalz, WestEd*

Students' Perceptions of Preservice Teachers' Behavior: Development and Evaluation of a Questionnaire Using Rasch and Multilevel Modeling, **Ridwan Maulana, University of Groningen; *Michelle Helms-Lorenz, University of Groningen; *Wim van de Grift, University of Groningen*

Testing for Differential Functioning and Group Differences on Cognitive Attributes: An Approach Based on the Least Squares Distance Method of Cognitive Diagnosis, **Dimiter M. Dimitrov, George Mason University; *Dimitar V. Atanasov, New Bulgarian University , Bulgaria*

Use of Rasch Rating Scale Modeling to Develop a Measure of District-Level Practices Identified to Increase Student Achievement, **Paul Soska, III, Eastwood Local School District; *Toni A. Sondergeld, Bowling Green State University; *Paul Andrew Johnson, Bowling Green State University*

Using a Rasch Analysis to Refine a Musicians' Self-Efficacy to Maintain Practice Schedules Scale, **D. Gregory Springer, Boise State University; *Joanne P. Rojas, University of Kentucky; *Kelly D. Bradley, University of Kentucky*

Using the Mixture Rasch Model to Explore Knowledge Resources Students Invoke in Mathematics and Science Assessments, **Danhui Zhang, Beijing Normal University; *Chandra H. Orrill, University of Massachusetts-Dartmouth; *Todd Campbell, University of Connecticut*

Validation of a Motivational Regulation Scale for Korean Elementary, Middle, and High School Students, **Hye-Sook Park, Honam University*

Ohio River Valley Objective Measurement Seminar (ORVOMS)

The fourth annual Ohio River Valley Objective Measurement Seminar (ORVOMS) will be held on May 2, 2014 at the Cincinnati Children's Hospital Medical Center in Cincinnati, Ohio.

This year's program will include presentations on topics such as the dichotomous model, facets applications, scale construction, paired comparisons, and logistic regression with Rasch models.

There is no fee to attend!

For information or to be placed on our mailing list please contact Melanie Lybarger (mlybarger \at/ theabfm.org).

Rasch-related Coming Events

Mar. 12-14, 2014, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), Leeds, UK,

Mar. 21, 2014, Fri. UK Rasch User Group Annual Day, York, UK, www.rasch.org/uk

Mar. 31-Apr. 3, 2014, Mon.-Thur. IOMW Biennial Meeting. Philadelphia, PA, www.iomw.org

Apr. 3-7, 2014, Thurs.-Mon. AERA Annual Meeting, Philadelphia, PA, www.aera.net

May 2, 2014, Fri. ORVOMS: Ohio River Valley Objective Measurement Seminar, Cincinnati, OH,

May 14-16, 2014, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), Leeds, UK,

May 19-21, 2014, Wed.-Fri. In-person workshop: Intermediate Rasch (A. Tennant, RUMM), Leeds, UK