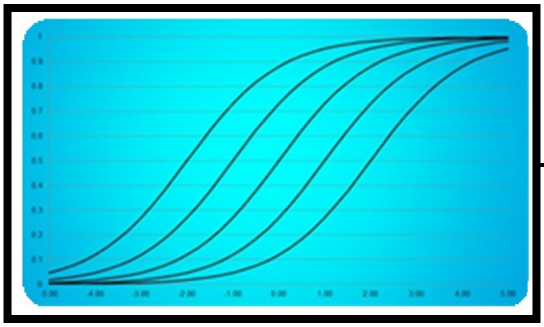


RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 28 No. 3 Winter 2014 ISSN 1051-0796

A Structure of Index and Causal Variables

Tesio (2014) responded with a very thoughtful article in RMT 28:1 to two articles by Stenner, Burdick and Stone that appeared in RMT, 22:1 (2008) and 22:4 (2009), which quite aptly, he referred to as "enlightening". The articles by these authors were concerned with distinguishing between what they referred to as *index* variables (also referred to as *formative* variables), and *causal* variables (also referred to as *reflective*). Essentially, as summarized by these authors, index variables are defined by their indices, whereas causal variables define or generate indices. Stenner, Burdick and Stone stressed that, although responses to a collection of items may fit the unidimensional Rasch model, their fit does not tell whether or not the variables are index or causal and that this distinction can only come from experimental evidence.

To contrast these two kinds of variables, Stenner, Burdick and Stone (2008) used Socio-economic status (SES) defined by *education*, *occupational prestige*, *income*, and *neighborhood* as an example of an index variable. They point out that these four indicators define SES, rather than that SES causes them, and that, for example, *If a person finishes four years of college, SES increases even if where the person lives, how much they earn, and their occupation stay the same.* (p.1153). In another terminology they employ, the indicators in an index variable are not *exchangeable*.

In contrast, they point out that reading, per se, is a causal variable, and that performance on a typical reading test is caused by a person's reading proficiency. As a result, items in a reading test are exchangeable. Of course, in assessing reading proficiency, items will not always be practically exchangeable. For example, very easy items and very difficult items may not be usefully exchanged in assessing very proficient readers or less proficient readers respectively. However, that is in part a practical matter and does not alter the definition of the variable of reading. It may also be a theoretical matter in the sense that it is necessary to understand why a reading task may be more difficult, perhaps qualitatively different in some sense from an easier one, but at the same time and in a relevant sense, be the same variable.

> Altogether these experiences – limited as they are to intelligence and attainment tests – suggest that once items have been constructed with an eye to uniformity of content, but variance in difficulty – which may even cover "complexity" – then there is a fair chance that they on the whole fit well into the model of simple conformity. (Rasch, 1960, p.125)

Tesio takes Stenner, Stone and Burdick's (2009) interpretation of the Functional Independence Measure (FIM$^{\text{TM}}$), which fits the Rasch model *reasonably well* and is discussed in Embretson (2006), as an index variable rather than a causal variable, and concludes: *I think that the FIM provides evidence of the fact that being an "index" rather than a "measure" is not necessarily an all-or-nothing concept* (p.1454).

He finally concludes that:

> *If my objections hold, an indicator that appears to be "formative" with respect to a high-order variable, can be "reflective" with respect to a lower-order one, closer to the biological extreme. Joint pain may be "formative" (hence, a poor item) with respect to "independence in daily life",*

Table of Contents

> *but "reflective" with respect to "perceived effectiveness of an anti-inflammatory drug".* (p.1455)

This note gives an example in educational assessment which seems to have the same features as articulated by Tesio. In particular, being index (formative) or causal (reflective) is not all-or-nothing, with the higher order level, which I call a "thick" variable, being index, and lower order level variables ("thin") being causal.

Consider the variable of temperature. Not ignoring the difficult road to constructing reliable thermometers and understanding what they were actually measuring, in physical terms temperature is an excellent example of a causal (reflective) variable and used by Stenner et al. illustratively. If an object is heated or cooled, it will cause a change in reading of temperature on the thermometer measuring it, and it will have the equivalent change on all thermometers. The thermometers are in principle exchangeable. There is again, in part, the practical matter of which kind of thermometer might be useful in any particular case. As with low and high proficiencies which require items of different difficulty, very high and very low temperatures may require different kinds of thermometers. Constructing such thermometers which are calibrated to the same scale requires also theoretical understanding of the materials and the way they react to changes in temperature. However, having and requiring the use of different kinds of thermometers at different temperatures calibrated on the same scale, reinforces that the definition of the variable in an important way is not changed.

Now consider the *assessment* of the *knowledge* about heat, the amount of which with respect to objects is measured by thermometers. Although it is a complex variable, for practical reasons of managing a curriculum in school and higher education, heat is often taught as a discrete topic. Assessment of the knowledge about heat can be made up of a collection of items and, in principle, many items can be constructed to assess the same knowledge of the variable of heat – the variable is causal (reflective) in that the knowledge of heat governs the probability of a correct response to each item. Many different items, dichotomous and polytomous, used to assess understanding of heat are constructed by teachers all over the world who teach this topic, and responses to a well-designed collection of items of different difficulty within a well-defined frame of reference may conform adequately to the unidimensional Rasch model.

Next consider the study of *heat* as a subset of the subject of *physics*, which is even more complex in its relationship to other fields of knowledge (e.g., chemistry), but which, for good practical reasons of managing a curriculum in school and higher education, is taught as a discrete subject at a higher order level than heat. The subject *physics* might be composed of not only *heat* but also *light, sound, electricity and magnetism* and *mechanics*. These topics can be thought of as relatively thin variables woven together to form a thick variable, much like a rope is made up of thinner strands. In this case it seems that the list of the five topics defines physics in some frame of reference and is, therefore, an index variable.

Finally, consider a test that is constructed to assess the understanding of physics conceived of as composed of the above five topics. We may imagine a 40 item test where each topic is assessed by eight items of varying difficulty within a topic, but more or less similar difficulties across topics. Within each topic, and as indicated above with respect to the topic of *heat*, the variables can be considered causal with many different collections of eight items exchangeable. The test is administered after the students have studied and revised all topics, a qualification which is part of the all-important frame of reference. We may find that in the sample of students assessed, the responses again conform adequately to the unidimensional Rasch model.

However, as indicated above, at the level of the variable of *physics*, the five topics listed are formative indices, and therefore the items between topics, even of the same relative difficulty, are not exchangeable. For example, if the topic of sound is excluded, the definition of physics is different from when it is included.

A further important qualification can be made. Simply listed as above, and treated as only a thick, index variable composed of *heat*, *light, sound, electricity and magnetism* and *mechanics,* does not highlight, let alone explain, why this set of thin variables is chosen to make up the thick variable of physics. The list can give the impression that it is merely a compendium of arbitrarily chosen thin variables. Although there might be some arbitrariness, including when and how they are taught, and so on, in any particular jurisdiction of education, the choice is not capricious. At deeper levels of understanding they become integrated. For students advanced in all topics, questions that show students' understanding of this complexity might be constructed. For example, waves appear in light and sound; energy is a governing principle but appears explicitly in *electricity and magnetism* and in *mechanics*, and so on. Thus in constructing items which assess such integrated understanding, the distinction between the five index variables that define physics above and the causal variables of which physics is composed again becomes blurred.

The blurring between index and causal variables should not paralyze us in thinking about constructing assessments. Instead, it is instructive to understand the part of the continuum between index and causal that one is operating at any given stage in the construction, analysis and interpretation of assessments.

The structure of the above illustration seems to be common in educational curricula and assessment, and seems analogous to the case made by Tesio in relation to the FIM in health outcomes assessment. Further considerations on how this structure of a combination of an index variable at one level, and causal variables at a lower level, might be usefully dealt with in some circumstances of assessment by applying the unidimensional Rasch model is beyond the scope of this note. However, I note that if the responses conform to the Rasch model to some level of precision, then to that level of precision the person profiles are relatively homogenous across the items. As in Tesio's example, a person whose profile does not conform to the model is not homogeneous, and this information can be used for diagnostic purposes in directing specific instruction.

David Andrich, University of Western Australia

References

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist, 61*(1), 50-55.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests* (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut. Available from www.rasch.org/books.htm.

Stenner, J., Burdick, D.S., & Stone M.H. (2008). Formative and reflective models: can a Rasch Analysis tell the difference? *Rasch Measurement Transactions, 22*(1), 1152-53.

Stenner, J., Stone, M.H., & Burdick D.S. (2009). Indexing vs. measuring. *Rasch Measurement Transactions, 22*(4),1176-77.

Tesio, L. (2014) Causing and Being Caused: Items in a Questionnaire May Play a Different Role, Depending on the Complexity of the Variable. *Rasch Measurement Transactions, 28*(1), 1454-56.

Items and Variables, Thinner and Thicker Variables: Gradients, not Dichotomies

The "formative vs. reflective variables" debate (Edwards, 2011) is an old and general epistemic and statistical issue, but for sure it is particularly interesting to all those working with latent variables under the item-response theory perspective. These variables by definition can only be hypothesized and thus we can decide whether they "exist" and must be discovered (a "realist" perspective, implying that observed behaviors "reflect" the causative entity) or whether they are pure mental creation (a "constructivist" perspective, implying that observed behaviors are "formative" with respect to the new entity). Highlighting the philosophical stance of the researcher and of the analyst is important because a satisfactory fit to statistical models (including Rasch models) can be achieved under either perspective. Thus, there is always the risk of circular, reassuring self-confirmation. Obtaining a measure "demonstrates" that the measured variable does indeed exist. Unfortunately, you can always measure an illusion: although, the realist perspective seems much less prone to such circularity (Borsboom, Mellenbergh and van Heerden, 2003). Personally, I follow a realist perspective with a pinch of constructivism: things are out there, they are real, but their meanings and goals are human constructions (Wright, 1973). That is why, in order to recognize an object, given their limited visual field, humans must look at the object from a sufficient distance. Otherwise, "the situation is reminiscent of the proverbial blind men touching an elephant; each describes the creature according to the part he can touch" (Shewmon, 2010). "Formative" variables are much easier to build, compared to "reflective" ones, because items come from your personal experience: you just need to assemble them; there is no need for you to "discover" them. Unfortunately, "formative" scales tend to be arbitrary checklists under the guise of measures. The empirical finding of a satisfactory fit with Rasch-expected parameters, attainable also with formative scales, does not contradict this position. An illusion, by definition, appears real to the deluded. Therefore, are "formative" variables the enemy?

Andrich's sharp commentary (Andrich, 2014) to a previous article of mine (Tesio, 2014) reinforces my warning against an overly dichotomic view of the problem. I gave the example of the FIM™-Functional Independence Measure disability scale. Andrich agrees with my opinion that *an indicator that appears to be "formative" with respect to a high-order variable, can be "reflective" with respect to a lower-order one, closer to the biological extreme. Joint pain may be "formative" (hence, a poor item) with respect to "independence in daily life", but "reflective" with respect to "perceived effectiveness of an anti-inflammatory drug".*

I whole-heartedly share the Andrich's suggestion to define the variables of higher or lower order of complexity as "thicker or thinner". He makes the clear example of "knowledge of physics" when a hierarchical component analysis of the curricula has to be conducted. I attempted to sketch a graphical representation of his thoughtful discussion in Figure 1 (contents slightly modified). On the left, Heat, Sound, Electricity & Magnetism and Mechanics can be seen as "formative" of (knowledge of) "Physics", and thus as an arbitrary choice of disconnected topics. If we delve deeper, however, we note that these can be seen as items "reflecting" one or more shared variables (Waves and Energy, for instance). Therefore, studying Heat (implying some understanding of Waves and Energy) may lead to a greater understanding of, say, Mechanics, and ultimately a greater knowledge of Physics as a whole. At the same time, the more disconnected the "items", and thus working as independent variables, the more likely it is that any gaps in teaching one topic will lead to a lower proficiency in "Physics". The sketch on the right emphasizes that being an "item" (a quantitative mark) or a "variable" (a qualitative entity) is also a location along a continuum. "Heat" can work as an item reflecting "Waves" or of "Energy". At the same time it can work as a variable inasmuch as it is disconnected from Sound, Electricity and Mechanics. In fact, it can "give rise to" its own reflective items such as Transfer efficiency, Entropy, Temperature.

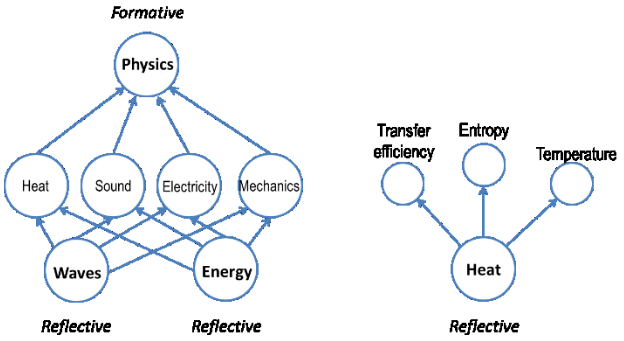


Figure 1.

Andrich must also be credited with the excellent, and very appropriate rope and strands metaphor he proposed in 2002 (Andrich, 2002). This metaphor is highly applicable to the concept of "complexity" (*cum plexus*, interwoven). Of course, the strands can be thought of as items of a thick rope, which is the variable. Alternatively, each item/strand can be seen as a thinner rope in itself.

Seeing the formative/reflective and, as a consequence, the item/variable dichotomies as continuous gradients allows us to better select our scale items and, most importantly, to focus our interventions, be they educational or therapeutic.

Ideally items should be homogeneous, which means they should be reflective of a unique shared real variable, not formative of an arbitrary, artificially constructed variable (a "realist" researcher needs to construct the scale, not the variable). But "pure" reflective items do not exist (see Figure 1): you have to select items "reflective and thus homogeneous enough" for your purpose. Novice analysts usually focus on items recalling personal observations, knowledge and beliefs, and thus start "constructing" their variable. This can be a deceptive process leading to a pseudo-variable, the measure of which is doomed to fail in terms of invariance. An effort of profound abstraction is required to "hypothesize" the latent, invisible trait: in my opinion this capacity implies long experience and deep reflections both in the specific field of application and in scale construction.

In 2002 I attempted to build a questionnaire measuring "severity of mental retardation" (Tesio, Valsecchi, Sala, Guzzon and Battaglia, 2002). Bladder continence seemed like a good item, because most of the subjects were incontinent. Unfortunately, Rasch modeling evidenced a severe misfit, as incontinence affected subjects with highly diverse levels of "retardation". After months of reasoning and software runs I understood that "continence" depended also on spasticity and epilepsy, common comorbidities in these subjects, yet it was quite unrelated to the trait of "retardation". And in fact, changing the item to "communicating voiding needs" (whatever the verbal or nonverbal code) removed the misfit, as a formative item was turned into a reflective one. Along the same line of reasoning, building "disease-specific" functional or disability scales is dangerous compared to building "generic" scales. Despite its derogatory connotation, "generic" means deeper, closer to the latent trait, and more explanatory. By contrast, "specific" means superficial and descriptive. Of course, you can easily find that "specific" scales are more precise: the closer your view the higher your resolution. The question remains, however, more precise about what? You know more about less. Clinical examples of specific scales are abound. "Manual ability" was proposed both as a "generic" scale (Simone, Rota, Tesio and Perucca, 2011) and as a family of scales adapted to the most various impairments (Arnould, Vandervelde, Batcho, Penta and Thonnard, 2012), stroke, neuromuscular diseases etc. "Pain" and "disability" also gave rise to countless disease-specific scales.

The effect of the "gradient" perspective is also relevant on the issue of treatment. In the figure it appears that teaching "physics" must probably imply the teaching of very generic and fundamental topics such as "Waves" and "Energy", intertwined with the direct teaching of "Heat", "Electricity", etc. How, when, how much and how long teaching should be provided to individuals and in distinct curricula (high-school vs. university, programs for physicists, engineers, physicians, etc.) is a difficult educational issue. The actual thickness and the length of the arrows in the figures are far from invariant.

In Medicine, behavioral interventions (such as exercise treatments) should be planned according to the same logic. Independence in daily life is "formed" by variables such as Dressing and Walking. Both, however, can be thought of as reflective of balance, visual acuity, spatial orientation etc. The time to "teach dressing" rather than walking, and/or "teach balance" rather than walking, is a matter of refined functional diagnostics on the individual patient. Ideally, a scale of balance, a scale of walking and a scale of independence should all be adopted, each provided with specific "reflective" items. The interactions across measures should suggest causal pathways towards "independence". In any case, Andrich's "thickness" gradient across distinct variables should be acknowledged and respected: a "one size fits all" policy leads to a very rough garment.

Luigi Tesio, Università degli Studi di Milano. Italy

References

Andrich, D. (2002). Implications and applications of modern test theory in the context of outcomes based education. Studies in educational evaluation, 28(2), 103-121.

Andrich, D. A. (2014). Structure of index and causal variables. Rasch Measurement Transactions, 28(3), 1475-1477.

Arnould, C., Vandervelde, L., Batcho, C. S., Penta, M., Thonnard, J. L. (2012). Can manual ability be measured with a generic ABILHAND scale? A cross-sectional study conducted on six diagnostic groups. BMJ Open, 2(6).

Borsboom, D., Mellenbergh, G. J., van Heerden, J. (2003). The theoretical status of latent variables. Psychological Review, 110(2), 203-219.

Edwards, J. R. (2011). The fallacy of formative measurement. Organizational Research Methods, 14(2), 370-388.

Shewmon, D. A. (2010). Constructing the death elephant: a synthetic paradigm shift for the definition, criteria, and tests for death. The Journal of medicine and philosophy, 35(3), 256-298.

Simone, A., Rota, V., Tesio, L., Perucca, L. (2011). The generic ABILHAND questionnaire can measure manual ability across a variety of motor impairments. International Journal of Rehabilitation Research, 34(2), 131-140.

Tesio, L. (2014). Causing and being caused: items in a questionnaire may play a different role, depending on the complexity of the variable. Rasch Measurement Transactions, 28(1), 1454-56.

Tesio, L., Valsecchi, M. R., Sala, M., Guzzon, P., Battaglia, M. A. (2002). Level of activity in profound/severe mental retardation (LAPMER): a Rasch-derived scale of disability. Journal of Applied Measurement, 3(1), 50-84.

Wright, L. (1973). Functions. Philosophical Review, 82(2), 139-168.

Frames of Reference

> A young man walks into a London tailor shop to buy a handmade Italian silk suit. Four hours later he stands in front of the mirror observing that the left sleeve is a little long, the right pant leg is a little short and the shoulders slightly bowed. The tailor assures the patron that the suit is fine; all he needs to do to improve the fit is lean his head awkwardly forward, tuck his left arm in tight to his body and turn his right foot to point inward toward the left. Later two sisters of the cloth meet the man walking toward them and the first speculates that a skiing accident may have been responsible for his crippled state and the second replies, "Ah yes! But Sister is that not the most beautiful suit you have ever seen?" And indeed it was! (old tale, retold)

In a recent paper (Stone & Stenner, 2014), we made the following points amplified below:

1. We construct science by making comparisons. These comparisons must be made by following a procedure leading to specific objectivity. Theory guides this process, but experimentation determines the outcome of theorizing and hypothesizing.

2. The two-way inter-individual frame of reference specifies the agent (e.g., text), object (e.g., readers) and resultant outcomes (e.g. comprehension rate or counts correct). The two-way intra-individual frame of reference specifies the agent (e.g., texts), the object (e.g., a single reader observed over time) and resultant outcomes (e.g., comprehension rate or counts correct on each measurement occasion along the individual's trajectory).

3. The two-way inter-individual frame of reference arranges and subsequently summarizes the comparisons. These comparisons are fundamental to what Rasch designated as *specific objectivity*. The two-way intra-individual frame of reference summarizes comparisons over time within person: one attribute and one person varying over time.

4. An inter-individual frame of reference may or may not be homologous with an intra-individual frame of reference, i.e. the attribute on which I differ from myself over time may not be the same attribute on which I differ from my brother (Borsboom, Kievit, Cervone and Hood, 2009).

5. Measurement follows from the results of qualitative comparisons that have been constructed in a systematic way using order as the fundamental characteristic.

What exactly is the *frame of reference*? Borowski & Borwein (1991) define the frame of reference as,

> "Any set of lines, directions, planes, etc., such as the coordinate axis relative to the position of a point in a space to be described; and in mechanics as, a particular choice of origin and basis vectors in three-dimensional space, and of a fixed initial point of the real line indexing time, to which the observations of a given observer may be referenced."(p. 230)

Compare this definition for "frame of reference" to what Einstein (1921) wrote,

> "If instead of "body of reference" (railway carriage or embankment) we insert "system of co-ordinates," which is a useful idea for mathematical description, we are in a position to say: The [dropped] stone traverses a straight line relative to a system of co-ordinates rigidly attached to the carriage, but relative to a system of co-ordinates rigidly attached to the ground (embankment) it describes a parabola. With the aid of this example it is clearly seen that there is no such thing, as an independently existing trajectory (lit. path-curve), but only a trajectory relative to a particular *body of reference*. . . . We must specify how the body alters its position with time; i.e. for every point on the trajectory it must be stated at what time the body is there. These data must be supplemented by such a definition of time that, in virtue of this definition, these time-values can be regarded essentially as magnitudes (results of measurements) capable of observation." (p. 60, original italics)

Compare it to another quote by Einstein (1923):

> "In all mechanical experiments, no matter what type, we have to determine positions of material points at some definite time, just as in the above experiment with a falling body. But the position must always be described with respect to something, as in the previous case to the tower and the scale. We must have what we call some *frame of reference,* a mechanical scaffold, to be able to determine the positions of bodies. In describing the positions of objects and men in a city, the streets and avenues form the frame to which we refer. So far we have not bothered to describe the frame when quoting the laws of mechanics, because we happen to live on the earth and there is no difficulty in any particular case in fixing a frame of reference, rigidly connected with the earth. This frame to which we refer all our observations, constructed of rigid unchangeable bodies, is called *the co-ordinate system*." (p. 156, original italics)

Note the change of phrase from "body of reference" to "frame of reference" in the space of two years, 1921 to 1923. This may suggest a transition in thinking, although it may be that two different translations of the same phrase from German to English account for this difference (we have not checked the German editions to date). These quotes are associated with Einstein's famous illustration of dropping a ball from a railway carriage and observing its trajectory (linear or parabolic) relative to the carriage or to the embankment which he used for introducing the concept of relativity. The frame of reference is a co-ordinate system.

Is Einstein the source for Rasch's use of this phrase? Rasch alludes to statements concerning physics applications throughout his work, but they are usually fleeting and tangential to the topic he was explicating (Rasch, 1960; 1967). Nevertheless, it seems clear that there is a decided affinity between Einstein and Rasch both expressing "the frame of reference" in a very similar context. Rasch (1977) explains:

> "… if this globality within A holds for any two objects O1 and O2 in O… the pair wise comparison is defined as specifically objective within the frame of reference F. The term 'objectivity' refers to the fact that the result of any comparison of two objects within O is independent of the choice of the agent A within A and also of the other elements in the collection of objects O; in other words: independent of everything else within the frame of reference, than the two objects which are to be compared and their observed reactions." (p. 76)
>
> … the qualification "specific" is added because the objectivity of these comparisons is restricted to the frame of reference F... denoted as the *frame of reference* for the specifically objective comparisons in question." (pp. 75-77, our italics)

Rasch makes very clear,

> "specific objectivity is not an absolute concept, it is related to the specific frame of reference... this

Rasch Measurement Transactions
www.rasch.org/rmt
Editor: Kenneth Royal
Email Submissions to: Editor \at/ Rasch.org
Copyright © 2014 Rasch Measurement SIG, AERA
Permission to copy is granted.
RMT Editor Emeritus: John M. Linacre
Rasch SIG Chair: Jim Tognolini
Secretary: Mikaela Raddatz
Treasurer: Leigh Williams
Program Chairs: Jessica Cunningham & Sara Hennings
Rasch SIG website: www.raschsig.org

> definition concerns only comparisons of objects, but within the same *frame of reference* it can be applied to comparisons of agents as well." (p. 77, our italics)

This is maximally important because Rasch specifies:

> "In order to distinguish this type of objectivity from other use of the same word I shall call it 'specific objectivity,' and in passing I beg you notice the relativity of this concept: it *refers only to the framework specified by the class of objects, the class of agents and the kind of observations which define the comparison*." (pp. 2-3, our italics)

Rasch also appears to have circumspectly avoided the philosophic issues of objectivity in his book (1960), and in his papers. Rasch (1977) later writes,

> "The concept [specific objectivity] has therefore not been carved out in a conceptual analysis, but on the contrary its necessity has appeared in my practical [statistical] activity." (p. 58)

It is important to observe a distinction made by Rasch between "indicators" and "specific objectivity". The essential point regarding this difference rests upon the comparison of two objects (or agents) independent of agent (or object) in the collection of objects (or agents), and their observed reactions within a specified frame of reference. As indicated earlier in the quotes above, not all comparisons meet the conditions of specific objectivity. A key issue is distinguishing "… those statements dependent on the agent (object)," specified by Rasch to be "local comparisons" from those produced as "specific objectivity." In Rasch's words (1977),

> "Objectivity is achieved when a comparison of any two objects is independent of everything else within the frame of reference other than the two objects which are to be compared and their observed reactions." (p. 77)

Independent comparisons, i.e. specific objectivity, result from a demonstrable frame of reference. Data cannot be "rasched" and cleansed of impurities simply from using software because it is not the values of the data matrix that define the frame of reference (think exploratory factor analysis), but the predictive matrix specified and confirmed by substantive theory (think LISRL for example). Our quotes and summary show *frame of reference* is a critical concept regarding the row by column, or agent by object matrix of data. It is especially critical to the conceptual frame of reference that theory be imposed prior via *the expected values of the frame of reference*. This is an essential point implied in Einstein's quotes.

Humphry and Andrich (2008, pp. 249, 261) define a frame of reference as:

> "A class of persons responding to a class of items in a well-defined response context... Frames of reference may be defined in terms of any empirical factor such as a characteristic shared by a class of items, an empirical condition for assessment, or a characteristic shared by persons."

Implied in the above quote is that frames of reference are not only specified by the dual carriage/embankment orientation illustrated in Einstein's example; they can be conceptualized as multiple (e.g. many-faceted).

While making the characteristics of their construction paramount, we must especially distinguish between a prescriptive causal model and a descriptive one (Stenner, Burdick, & Stone, 2008). In a descriptive Rasch model, data defines the frame of reference. In a two way frame of reference A and O, when "persons" or "items" are deleted, to improve data fit to the model, then the frame of reference changes (David Andrich, Personal Communication). The consequences are that items may assume different positions in the newly constituted frame of reference and similarly for persons. Though these differences in scale locations for persons and items may be small, this fact should not disguise the reality that the frame of reference has shifted. We might argue that data editing "cleans" up the data and the newly constituted frame of reference is closer to the "true" frame of reference were there no aberrations in the data. However, without a data independent statement of our intention, how do we know that data editing is not moving us further away from our intended frame of reference? Doubly prescriptive models force us to recognize that substantive theory and Rasch model fit must be jointly satisfied. In a doubly prescriptive Rasch model the substantive theory is sacrosanct. The frame of reference is theory dependent and misfit is unambiguously seen to be a problem with the data, not with the frame of reference. A stronger caveat can be found, 80 years ago, in the words of the late astronomer Arthur Eddington (1935):

> "It is also a good rule not to put too much confidence in the observational results that are put forward until they are confirmed by theory." (p. 211)

As it turns out, fortunately for this paper, the well-dressed young man is a working psychometrician wrestling with a misbehaving data set in a two way frame of functional independence items and a sample of mixed gender octogenarians. He fits the data to a dichotomous Rasch model and finds he has a few misfitting persons and misfitting items. He removes the misfits and reruns the analysis. To his surprise, a new set of persons and items misfit. After a couple of more 'rasch and repeats', he grows uncomfortable and asks himself, "What violence have I done to my original intended frame of reference?" A prescient question because in a data dependent frame of reference, any and all data edits will modify the frame of reference. A consequence of such edits may be improved fit to the Rasch model, but what about fit to the original conception of the attribute? Note, that whether the shift to the frame of reference is small, medium or large is irrelevant. Our question is a logical one, "Do we want differences between persons or differences between items to depend upon which items we include and which persons we include in the calibration study?" If the sole criterion is fit to a Rasch model then it would seem that 'nips and tucks' to the data matrix are defensible. However, if a doubly prescriptive model (prescriptive as to substantive theory *and* Rasch data structure) is employed, then 'nips and tucks' do not disturb the frame of reference precisely because that framework is built from theory and not data.

Mark Stone and Jack Stenner

References

Borsboom, D. Kievit, R.A., Cervone, D. and Hood, B.S. (2009). The two disciplines of scientific psychology, or: The disunity of psychology as a working hypothesis. In J. Valsiner, P. C.M Molenaar, M.C.D.P. Lyra and N. Chaudhary (eds.) Dynamic Process Methodology in the Social and Developmental Sciences. London: Springer Science & Business Media.

Eddington, A. (1935). *New pathways in science*.
Cambridge: Cambridge University Press.

Einstein, A. *Relativity, the special and general theory.* (1921). New York: Henry Holt. Reprinted in Relativity Theory (1968), L. P. Williams (Ed.) New York: Wiley.

Einstein, A. & Infeld, L. (1923). T*he evolution of physics*. Reprinted 1966, Simon & Schuster, New York.

Humphry, S. M. & Andrich, D. (2008). Understanding the unit in the Rasch Model. *Journal of Applied Measurement, 9 (*3), 249-264.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Rasch, G. (1967). *An informal report of objectivity in comparisons*. Psychological measurement theory. Proceedings of the NUFFIC International summer session in science at "Het Oude Hof" at Den Haag July 14-28.

Rasch, G. (1977). *On Specific Objectivity: An Attempt at Formalizing the Request for Generality and Validity of Scientific Statements*. Danish Yearbook of Philosophy, 14, 58-94.

Stenner, A. J. Burdick, D. & Stone, M. (2008). Formative and reflective models: Can a Rasch analysis tell the difference? *Rasch Measurement Transactions, 22* (1), 1059-1060.

Stone, M. H. & Stenner, A. J. (2014). Comparison is key. *Journal of Individual Measurement, 15* (1), 26-39.

Model-Data Fit and Adjustments for Rater Effects

Although complete rating designs are interesting from a theoretical perspective, most operational rater-mediated assessment systems involve various forms of incomplete assessment designs. When each rater does not score every student, then there is the potential for unfair ratings based on the good or bad luck of the rater draw. It is essential to establish a common scale for describing student achievement that minimizes the effects of individual rater characteristics when incomplete assessment networks are utilized. If rating data are collected with sufficient links between raters, then the Facets model (Linacre, 1989) can be viewed as a type of equating model with parameter estimates describing rater severity, student achievement, and other facets of interest on a common scale (Lunz & Suanthong, 2011). In the context of rater-mediated assessments with student, rater, and item facets, the Facets model serves as type of equating model in which each student encounters a subset of common "items" (in this case, raters), and achievement estimates are adjusted for variations in rater severity and item difficulty.

The importance of establishing sufficient connectivity between facets in rating designs has been emphasized in previous research (Eckes, 2011; Engelhard, 1997). However, it has not been widely recognized that model-data fit also plays a crucial role in linking student achievement estimates across raters. Even a well-designed study with adequate links between raters does not guarantee the advantages of invariant measurement unless model-data fit is systematically examined and supported.

In this note, variation in rater discrimination is used to illustrate the impact of one type of rater misfit on the interpretation of student achievement estimates on a common scale. Rater misfit to the Rasch model can occur in several different ways (Wright & Linacre, 1994), and a variety of Rasch-based statistics are frequently used to identify rater effects, including Infit and Outfit statistics (Engelhard, 2013). A rater discrimination parameter is used in this illustration because it shows the consequences of one type of model-data misfit on adjustments for rater severity in the context of rater-mediated assessments.

Three raters are included in the illustration: Rater A is a lenient rater with adequate fit to the Rasch model (slope = 1.00); Rater B is a severe rater with adequate fit to the Rasch model (slope = 1.00); and Rater C is a severe rater who does not fit the Rasch model (slope = 0.60). The top panel of Figure 1 displays conditional expected ratings for two raters who meet the expectations of the Rasch model: Rater A (lenient) and Rater B (severe). When raters display adequate model-data fit, it is possible to make a constant adjustment for differences in rater severity using the expected ratings. In other words, parallel response functions imply that student achievement can be interpreted on a common scale across severe and lenient raters. The bottom panel of Figure 1 displays conditional expected ratings for a lenient rater who meets the expectations of the Rasch model (Rater A) and a severe rater who displays misfit to the Rasch model (Rater C) with a slope parameter less than 1.00. When crossing rater response functions are observed, it is not possible to make a constant adjustment to account for differences in rater severity. In other words, there is no simple adjustment can equate achievement estimates for students who were scored by Rater A and students who were scored by Rater C.

The purpose of this note is to remind researchers using Rasch measurement theory that the desirable properties related to invariant measurement and invariant rater-mediated assessment are only available when adequate model-data fit is present. Simply using Rasch measurement theory and adjusting for rater differences in severity without checking for model-data fit can lead to misleading adjustments. Model-data misfit as captured in the widely used Infit and Outfit statistics does not provide sufficient information to diagnose the sources of misfit (e.g., variation in rater slopes) that may impact the quality of the adjustments for variation in rater severity.

Stefanie A. Wind, Georgia Institute of Technology
George Engelhard, Jr., University of Georgia

References

Eckes, T. (2011). Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments. Frankfurt am Main: Peter Lang.

Engelhard, G., Jr. (1997). Constructing rater and task banks for performance assessments. Journal of Outcome Measurement, 1(1). 19-33.

Linacre, Lunz, M., & Suanthong, S. (2011). Equating of multi-facet tests across administrations. Journal of Applied Measurement, 12(2), 124-134.

Linacre, J. M. (1989). Many-facet Rasch measurement. Chicago: MESA Press.

Wright, B.D. & Linacre, J. M. (1994). Reasonable mean square fit values. Rasch Measurement Transactions, 8(3), 370.

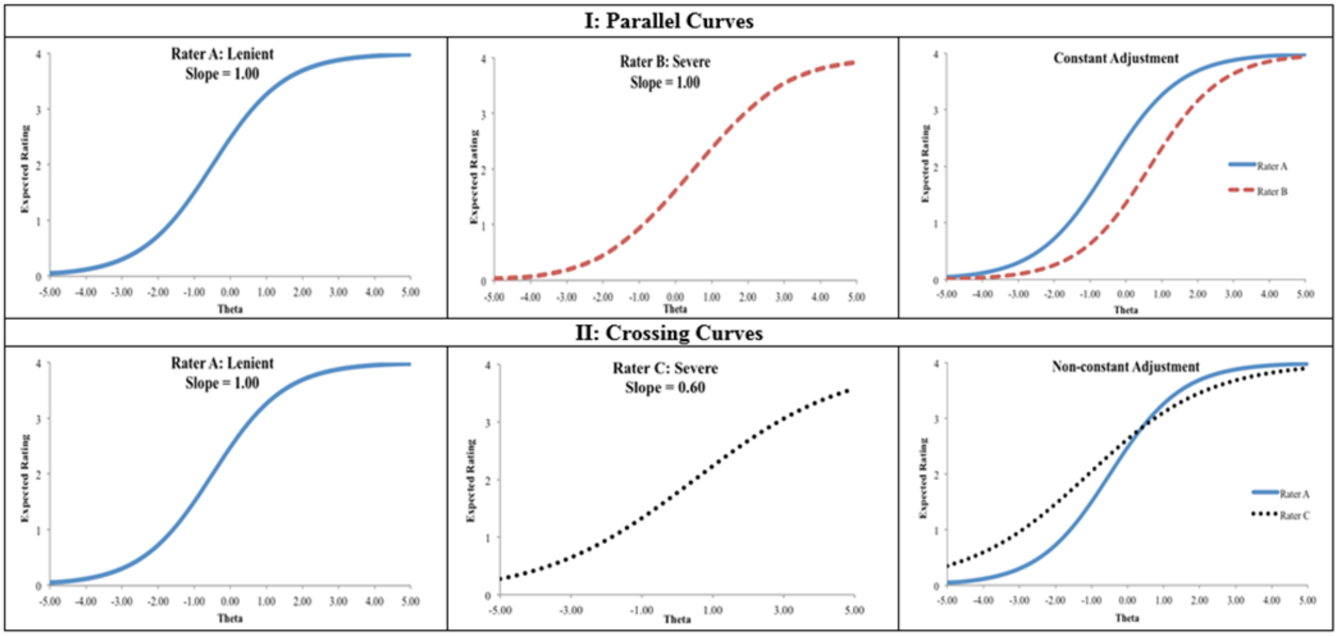


Figure 1. *Rater Adjustment and Model-Data Fit*

Using Rasch Simulation Data to Verify Whether Ferguson's Delta Coefficient Can Report Students' Abilities Are Equal in a Class

Many teachers are concerned about whether students' abilities are equal. The more equal students' abilities are, the more willing many teachers are to teach the class. A coefficient is required to compare the degree of equality between students' academic abilities. Ferguson's Delta (1949) an index of discrimination measured by the proportion of discriminations (i.e., the degree to a uniform distribution), reported that a normal distribution would be expected to have a discrimination of Delta > 0.90.

We are thus interested in verifying whether Delta is > 0.90 when a sample with a normal distribution fits a Rasch (1960) model. Rasch simulation data (Linacre, 2007) were used when sample sizes were 10, 50, 100, 200, 500, and 1000, item lengths were 5, 10, 20, 40, and 60, and 4 kinds of categories from 2 to 5 were manipulated in the study.

Sample data from normal and uniform distributions were yielded using both dichotomous Rasch and its Rasch Rating Scale model, respectively on two kinds of studied data. Ferguson's dichotomous Delta and Hankins' polytomous Delta_ g (2007/2008) were respectively produced. Another Delta setting a fixed number to 5 bins (Delta_5 shown in Figure 1) was also computed for comparison with the former two. Here, Delta_5 =g/(g-1)*(n^2-(SUMSQ(fg)))/n^2, where g=the number of bins, n=sample size, SUMSQ(fg) is the summation of all the bin's squared frequency (i.e., fg from 1 to 5).

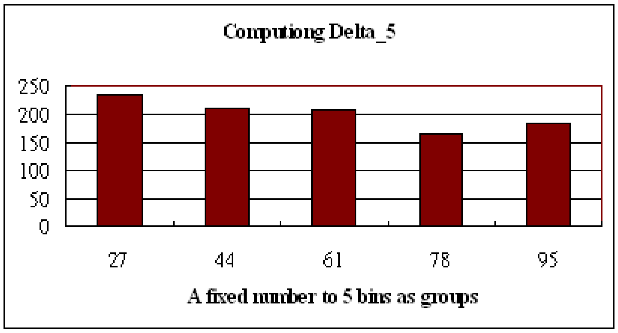


Figure 1. Example of computing Delta_5

We simulated 100 times for those 24,000 (2 distributions × 6 samples × 5 item lengths × 4 categories) possible combinations and calculated 95% confidence intervals (CIs) for the three aforementioned Delta values to verify whether Delta is > 0.90 when a sample comes from a normal or a uniform distribution when the data fit a Rasch model.

We found that (1) when samples are uniformly distributed and respond to a 2-point Rasch model test, Ferguson's dichotomous Delta = 0.96 (95% CI = 0.86, 0.99), and Delta_5 = 0.94 (95% CI = 0.80, 0.99). When responding to a polytomous Rasch model test, Delta_g = 0.97 (95% CI = 0.88, 0.99), and Delta_5 = 0.96 (95% CI = 0.89, 0.99).

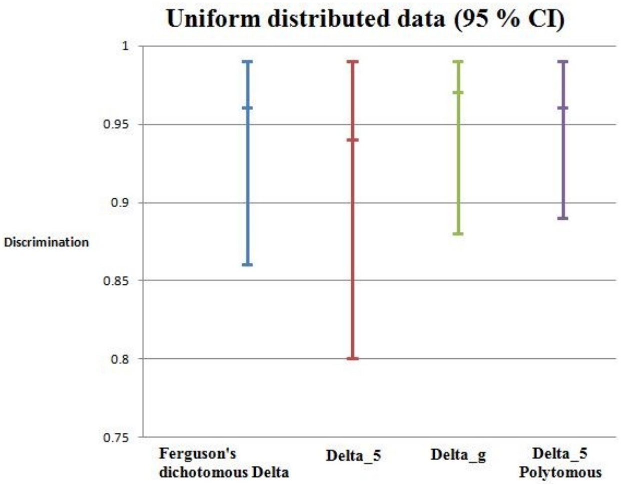


Figure 2. Result from uniform distributed data

(2) When samples are normally distributed and respond to a 2-point scale, Delta = 0.91 (95% CI = 0.82, 0.97), and Delta_5 = 0.89 (95% CI = 0.79, 0.96). When responding to a polytomous test, Delta_g = 0.94 (95% CI = 0.86, 0.98), and Delta_5 = 0.89 (95% CI = 0.80, 0.97).

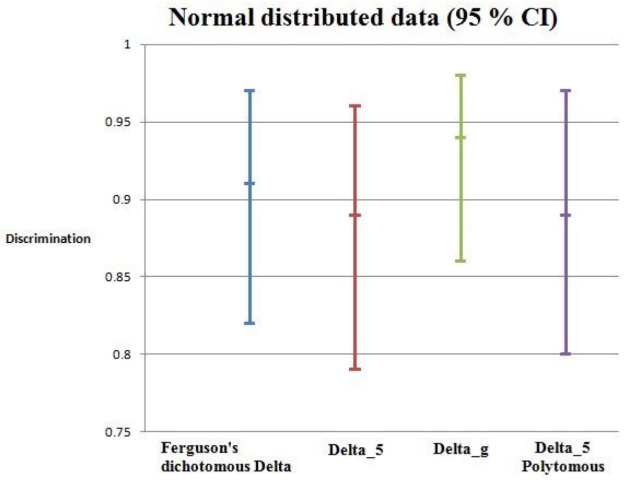


Figure 3. Result from normal distributed data

There is insufficient evidence to expect that a normal distribution or a uniform distribution has a Ferguson’s Delta > 0.90 when considering its 95% CI. For simple and easy use in the education field, we suggest that Delta_5 be used to describe the degree of equality of students’ abilities within a class or between classes in a school.

Tsair-Wei Chien,' Chi Mei Medical Center, Taiwan
Ngadiman Djaja, School of Public Health and Social Work, Queensland University of Technology, Australia

References

Ferguson, G.A. (1949). On the theory of test discrimination. Psychometrika, 4, 61-68.

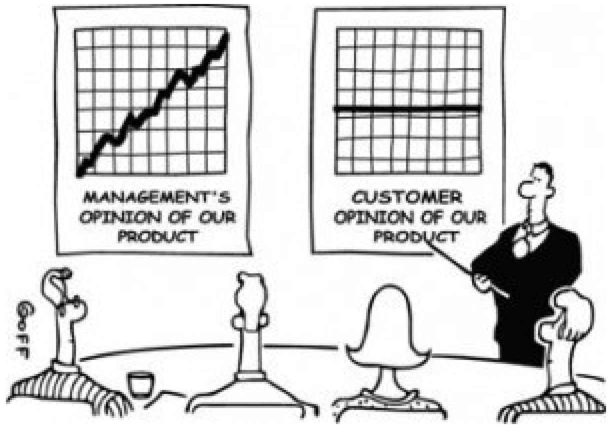
Linacre, J.M. (2007) How to Simulate Rasch Data. Rasch Measurement Transactions, 21(3), 1125.

Hankins, M. (2007). Questionnaire discrimination: (re)-introducing coefficient Delta. BMC Medical Research Methodology, 7, 19.

Hankins, M. (2008). How discriminating are discriminative instruments? Health Qual Life Outcomes, 6, 36.

Hankins, M. (2008).Discrimination and reliability: equal partners? Understanding the role of discriminative instruments in HRQoL research: can Ferguson's Delta help? A response. Health and Quality of Life Outcomes, 6, 83.

Differential Person Functioning?


MANAGEMENT'S OPINION OF OUR PRODUCT
CUSTOMER OPINION OF OUR PRODUCT
GoFF

Georg Rasch is Still Making the News

Like great composers, artists and writers, psychometricians can achieve a measure of immortality via the display, production and reproduction of their work outputs. Psychometricians can “live on” in this way, through the useful application of the standardized tests and questionnaires they helped create and develop. Think Binet, Wechsler, Eysenck, and Raven... to name but a few psychometricians. Georg Rasch “lives on” in this way too - through the test, the BPP (Børge Prien’s Prøve) (see Wright, 1991, RMT 5:3). The BPP is a test Georg Rasch helped create and develop with his son-in-law, Børge Prien, in Denmark. This test is still being used today, in its original form, and was recently referred to in the science news magazine the New Scientist and was then picked up by the international news wires.

See the New Scientist article (subscription required):

http://www.newscientist.com/article/mg22329830.400-brain-drain-are-we-evolving-stupidity.htmlutm_source=NSNS&utm_medium=SOC&u

tm_campaign=facebookgoogletwitter&cmpid=SOC%7CNSNS%7C2012-GLOBAL-facebookgoogletwitter

Or the newspaper article:

http://www.dailymail.co.uk/sciencetech/article-2730791/Are-STUPID-Britons-people-IQ-decline.html

The quote "*An IQ test used to determine whether Danish men are fit to serve in the military has revealed scores have fallen by 1.5 points since 1998*." from the newspaper article caught my eye, and reminded me of David Andrich's recent paper (Andrich, 2013) which describes Rasch's psychometric approach with "a Danish military intelligence test" (see Marosszeky, 2014, RMT 28:2). Could this be Georg Rasch's test? I asked myself.

Upon accessing the original article, I found the key source and commentator, Professor Thomas W. Teasdale, (see http://psychology.ku.dk/Academic_staff/?pure=en%2Fpersons%2Fthomas-william-teasdale(dbc867bb-0954-4525-ade1-cd610aa8c757)%2Fpublications.html)). His publications list then led me to a number of recent papers about the BPP (Teasdale and Owen, 2008; Teasdale, 2009; Teasdale, et al. 2011).

Interested RMT readers may like to see how Georg Rasch is still making the news, and contributing to scientific discourse and advancement.

Nick Marosszeky, Macquarie University (Australia)

References

Andrich, D. (2013). The legacies of R.A. Fisher and K. Pearson in the application of the Polytomous Rasch Model for assessing the empirical ordering of categories. *Educational and Psychological Measurement, 20*, 1-28.

Marosszeky, N. (2014). Games Psychometricians Play. *Rasch Measurement Transactions, 28*(2).

Teasdale, T. W. (2009). The Danish Draft Board's intelligence test, Børge Prien's Prøve: Psychometric properties and research applications through 50 years. *Scandinavian Journal of Psychology, 50*, 633-638.

Teasdale, T. W., Hartmann, P. V. W., Pedersen, C. H., & Bertelsen, M. (2011). The reliability and validity of the Danish Draft Board Cognitive Ability Test: Børge Prien's Prøve. *Scandinavian Journal of Psychology, 52*, 126-130.

Teasdale, T. W., & Owen, D. R. (2008). Secular declines in cognitive test scores: A reversal of the Flynn Effect. *Intelligence, 36*, 121-126.

Wright, B. D. (1991). Georg Rasch's BPP. *Rasch Measurement Transactions, 5*(3).

Rasch-related Coming Events

Jan. 2-30, 2015, Mon.-Fri. Online workshop: Practical Rasch Measurement – Core Topics (E. Smith, Winsteps), www.statistics.com

Jan. 12-14, 2015, Mon.-Wed. 6th Rasch Conference: Sixth International Conference on Probabilistic Models for Measurement in Education, Psychology, Social Science and Health, Cape Town, South Africa www.rasch.co.za/conference.php

Mar. 11-13, 2015, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Mar. 20, 2015, Fri. UK Rasch User Group Meeting, London, United Kingdom, www.rasch.org.uk

Mar. 26-27, 2015, Thur.-Fri. In-person workshop: Introduction to Rasch Measurement with Winsteps (W. Boone), Cincinnati, www.raschmeasurementanalysis.com

April 16-20, 2015, Thurs.-Mon. AERA Annual Meeting, Chicago, IL, www.aera.net

April 21-22, 2015, Tues.-Wed. IOMC 2015: International Outcomes Measurement Conference, Chicago, IL, www.jampress.org

May 13-15, 2015, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), Leeds, UK,

May 18-20, 2015, Mon.-Wed. In-person workshop: Intermediate Rasch (A. Tennant, RUMM), Leeds, UK,

May 29-June 26, 2015, Fri.-Fri. Online workshop: Practical Rasch Measurement – Core Topics (E. Smith, Winsteps), www.statistics.com

July 3-31, 2015, Fri.-Fri. Online workshop: Practical Rasch Measurement – Further Topics (E. Smith, Winsteps), www.statistics.com

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 March 1, 2015, so please make your submission by February 1, 2015 for full consideration. Please email Editor\at/Rasch.org with your submissions and/or ideas for future content.