PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2017, 54(2), 1-7

Pontificia Universidad Católica de Chile
Facultad de Educación

Special section introduction

Applied Educational Measurement

Mark Moulton[1], Hong Jiao[2] & María Verónica Santelices[3]

[1]Educational Data Systems
[2]Department of Human Development and Quantitative Methodology,
University of Maryland, College Park
[3]Pontificia Universidad Católica de Chile

The International Objective Measurement Workshop (IOMW) is the biennial conference from which the four articles in this special issue of PEL were drawn. They were based on presentations given in Washington, DC, in April 2016. The IOMW has long nurtured an interest in the philosophy and possibilities of what is called "objective measurement", or "invariance", specifically as implemented by the Rasch model. That interest is as keen today as it was in the 1980's when the conference began. As an informal introduction, it may be useful to review what "objectivity" means, how it is rooted in the physical sciences, and why the authors of these papers consider it an important thing to have.

What does it mean to do a quantitative analysis of a dataset?

The answer varies quite a bit across fields. In statistics, quantitative analysis is intended to yield a concise mathematical description of a dataset with emphasis on deciding whether observed numerical differences are "significant", meaning not likely to have occurred by chance. This involves calculating means, standard deviations, standard errors, and related statistics, more or less the approach taken by "classical test theory".

In the 1950's, thinking about a language aptitude dataset, Danish mathematician Georg Rasch realized that a statistical description of how students perform on a particular test is *not* what he wanted. Suppose students are given different test forms with different sets of items. Suppose the test forms change from year to year. Suppose we want to compare students from different grades. Suppose data is missing, and not at random. Under these circumstances, statistical description of performance on a particular test is not sufficient to compare students across tests in any kind of generalizable way. In addition, Rasch realized that he wanted to speak clearly about *individual* students, not the population as a whole, and he did not want to use statistics that depend on the performance of other students ("grading on the curve") or on the items they happened to take, which seems manifestly unfair. In short, Rasch wanted a way to measure student ability that would be as simple, reproducible and fair as measuring a student's height with a yard stick or measuring physical quantities, like force and mass with a force gauge or pan balance.

ISSN:0719-0409 DDI:203.262, Santiago, Chile
doi: 10.7764/PEL.54.2.2017.1

How *do* physicists measure things? Rasch considered the example of force and mass as defined by Newton (Rasch, 1960):

$$f = ma$$

The force exerted on a mass (in a given direction) is defined to equal the magnitude of the mass multiplied by its acceleration. Although scientists measure mass using a pan balance and various other methods, Rasch realized that Newton's formulation can be interpreted as a circular definition. Force is defined in terms of mass; mass is defined in terms of force. Suppose, then, that the only data we have are the accelerations that occur when various forces act on various masses. The result is a matrix:

Table 1
Matrix of Accelerations as the Product of Force and Mass

Force / Mass	M1	M2	M3
F1	A11	A12	A13
F2	A21	A22	A23
F3	A31	A32	A33

Given just the values of the accelerations and Newton's force equation, it is *almost* possible to calculate values for the individual forces and masses. What we get are their *ratios*. For example, if A11 is half as large as A21, we can calculate that F1 is also half as large as F2:

A11 = F1/M1
A21 = F2/M1
A11/A21 = (F1/M1) / (F2/M1) = F1/F2 = ½

In other words, because the two accelerations apply to the same mass (M1), the M1 factors out and we are able to get the ratio of the two forces from just the accelerations *without knowing M1*. We get that same force ratio whether the mass is M1, M2, M3 or any other mass. In other words, the force ratio is "invariant" across masses, a property Rasch called "specific objectivity". By the same reasoning, it is clear that each mass ratio is also invariant across forces.

What we don't have are absolute values for F1 and F2 and the various masses. This is addressed by establishing a convention—defining a "reference mass". For instance, we can define M1 to be the unit mass, relating it to something like the weight of a volume of water. Because we already have a measurement for A11, that means:

F1 = M1 * A11 = 1 * A11 = A11
which we can then use to calculate values for other masses like M2:
M2 = F1 / A12

Therefore, so long as we relate all forces to the same reference mass, values for the force measurements can be calculated, and from these the values of the remaining masses, and they will be comparable in all frames of reference that are based on the reference mass. A definition of a reference mass was proposed, in fact, on April 7, 1795, when the gram was decreed in France to be "the absolute weight of a volume of pure water equal to the cube of the hundredth part of the metre, and at the temperature of melting ice" ("Décret…", 1795). The standard has been tweaked a few times but is still pretty much the same.

(Although defining physical mass in terms of force was championed by Ernst Mach (1919) and was theoretically sufficient for Rasch's purposes, it runs into practical difficulties related to the impossibility of measuring acceleration at an instant and disentangling forces at the atomic level (Belkind, 2012)).

What does this have to do with testing children? Rasch realized that just as physics can set up a system of

measures for force and mass using just accelerations, it should be possible to set up a system of measures for student ability and item difficulty using just the student performances on each item. To get individual values, we set up something analogous to a reference mass by anchoring the mean of the item difficulties in a test at zero. To compare students across tests, we define "reference items" to which all students can be connected via items in common across test forms.

Rasch also realized that measuring students in this way introduces a built-in system of quality control. We noted that the ratio of any two forces, and of the accelerations associated with those forces, should be the same for each mass. Therefore, to the degree we do *not* see this equivalence (the acceleration ratios for one of the masses, for instance, do not match those of the other masses), we know that objective (invariant) measurement has not occurred. The data for that mass needs to be scrutinized and probably omitted from the dataset. What a scientist *cannot* do is pretend that those accelerations are valid and use them uninhibitedly in his calculations. In other words, Rasch's model imposes a requirement that the data must fit the expectations generated by applying the measurement model; otherwise, they cannot be used. This is the opposite of the approach usually taken in statistics where, in order to protect against selection bias, data are treated as sacrosanct and new model parameters are introduced and tweaked if necessary to improve fit to the data.

In retrospect, the analogy to force and mass is irresistible. Students are indeed like forces trying to push a test question (a mass) from a state of unsolved to solved. The main difference is that student test data do not consist of smooth, precise, metric numbers like accelerations but of jerky, imprecise, non-metric numbers like 0 and 1 ("incorrect" and "correct"), making it necessary to estimate and work with probabilities. So instead of saying:

$$a = \frac{f}{m}$$

we say

$$\frac{p[success]}{p[failure]} = \frac{ability[person\ n]}{difficulty[item\ i]}$$

, where p means "probability". The two formulations are mathematically very similar.

Because humans find it easier to picture quantities on an interval rather than a ratio scale, we take the (natural) log of both sides of the equation to make it additive rather than multiplicative:

$$log\left(\frac{p[success]}{p[failure]}\right) = log\left(\frac{ability[person\ n]}{difficulty[item\ i]}\right)$$
$$log\left(\frac{p[success]}{p[failure]}\right) = log(ability[person\ n]) - log(difficulty[item\ i]$$

To use a common Rasch nomenclature, we define:

$$\beta_n = log(ability[person\ n])$$
$$\delta_i = log(difficulty[item\ i])$$
$$p_{ni} = probability\ of\ success\ of\ person\ n\ on\ item\ i$$

So,

$$\beta_n - \delta_i \equiv log\frac{p_{ni}}{(1-p_{ni})}$$

The difference between the ability of person *n* and the difficulty of item *i* is defined to be the log of the probability of success over the probability of failure, also called the "log odds" of success. Note that this is a *definition*—we are defining the difference between a person ability and an item difficulty. It amounts to saying: the difference between a person and an item is defined to be zero when the probability of success is 0.50.

With a little algebra, the formula can be rearranged to calculate the probability that person *n* will succeed on item *i*, which is the canonical form of the Rasch model for dichotomous data:

$$p_{ni} = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}$$

It says the probability of success of a person on an item is a function of the difference between that person's ability and the item's difficulty. To analyze a test data set, we adjust the person betas and item deltas until the probabilities they imply for each cell in the matrix most closely fit the observed data, a process called "maximum likelihood estimation". Beyond such details, Rasch's model is just a probabilistic way to implement Newton's force equation.

There are significant benefits to this way of thinking about measurement:

- Minimal statistical assumptions.
- Works for small (e.g. classroom-sized) datasets.
- Missing data is not a problem.
- Students are comparable across test forms, even if one form is much harder.
- Items are comparable across student groups, even if one group is much more advanced.
- Flags items that do not belong on the test, for whatever reason.
- Flags students who were not properly measured by the test, for whatever reason.
- Clarifies whether items embody a coherent construct.
- Fit to the model implies that student measures are not influenced by the performance of other students or by which items were selected for the test.
- And that means *the test is fair*.

The main caveats are:

- Unidimensionality. The items should all get at the same underlying construct embodying a well-specified content domain. The Rash model has been generalized to include multidimensional models, but they still must use items that reflect a clearly defined content domain or combination of content domains.
- Local item independence. Each item should be statistically independent of the others.

In summation, by applying a simple physical analogy to human behavior in a controlled situation (a test), it is possible, in principle, to make precise scientific measurements of a person's interior state and mental traits, and many other things besides. These measurements have the potential to be every bit as valid as, and, indeed, are mathematically indistinguishable from, the physical measurement of force and mass. As a tool of measurement, the model is applicable to any dataset, in any field, where some kind of "force" can be imagined to

encounter some kind of "resistance" to produce an outcome within a well-specified domain. Indeed, inasmuch as reproducible measurement is a prerequisite of science, it is unclear how any scientific field can progress without some comparable palette of measurement techniques. It is this intuition that has kept the Rasch psychometrics community going for the past 50 years and which shines through the lines of the four articles in this special issue.

In their paper, *Developing a Survey to Explore of Sense of Belongingness related to Language Learning Using Rasch Measurement Theory*, Knisely & Wind show how measurement theory can apply to something as subtle as one's sense of belonging in a community. In particular, they explore the question of whether "sense of belonging" is a coherent construct that can be measured. The Rasch model is often used in this way—not just to confirm and implement a well-known construct (e.g. language ability) but to identify and refine new, poorly understood constructs. In this case, a procedure to measure "belongingness" makes it possible to explore a variety of interesting questions, for instance, whether a strong feeling of belonging facilitates academic achievement among college students.

Wang & Engelhard, in *Using a Multifocal Lens Model and Rasch Measurement Theory to Evaluate Rating Quality in Writing Assessments*, develop an analogy between rater-mediated assessments and optics—the idea that raters are like "lenses" that both reveal a pattern of student achievement and distort it in systematic ways. Building on work on "lens models" by Brunswick in the 1950s, they show that the "many-facet Rasch model" can be used in lieu of traditional regression analysis to model rater effects. Facets models build on the idea that an observed phenomenon can be, and almost certainly is, the effect of *more than two causes*, exploiting the property that Rasch models, due to their simple additive nature, can be used to separate and disentangle each layer of causality. Going back to physics, the observed acceleration of a spacecraft can be modeled as the result of: a) thrust from rocket engines (facet 1), b) inertial mass of the rocket (facet 2) and c) gravitational force of the planet (facet 3). Similarly, performance on a writing assessment can be modeled as the result of: a) writing proficiency of a student (facet 1), b) difficulty of the writing domain (facet 2) and c) severity of the rater (facet 3). By separating out these forces, Wang & Engelhard are able to compare the performance of expert and operational raters to predict the fairness of the test when it goes operational. They show how using the many-facet Rasch model to implement Brunswick's lens concept linearizes estimates of writing proficiency into interval measures, supports analysis of rating scale data (regression requires interval data) and enables an apples-to-apples comparison between two classes of rater.

It was mentioned earlier that the core Rasch model requires unidimensionality but that multidimensional extensions of the model have been developed. Perhaps the most prominent of these is the obscurely named Multidimensional Random Coefficient Multinomial Logit (MRCML) model (Adams, Wilson & Wang, 1997). Steffen Brandt's *Concurrent Unidimensional and Multidimensional Calibration Within the Item Response Theory* makes an original contribution to this body of work by addressing a simple, yet embarrassing, problem faced by testing practitioners everywhere.

Test designers are often asked to design tests that report one overall score for each student, plus a set of "diagnostic" subscale scores. The request *sounds* innocuous, but it presents a serious dilemma. If we design the test to be *unidimensional*, the overall score on the main dimension makes sense and fits the Rasch model, but the subdimension scores are meaningless since they are, in fact, just less precise copies of the main dimension, however imaginatively we may name them. On the other hand, if we design the test to be *multidimensional*, we get a richer view of student abilities and meaningful subdimension scores (plus better handling of "local item dependence", which means some items are not fully independent of each other in a statistical sense), but there is no clearly defined Rasch-compliant overall score on the main dimension, just a hobbled together weighted mean of subdimension statistics. Brandt figured out a way to get *both* types of scores—a single overall score and subdimension scores—out of one test using what he calls the generalized subdimension model (GSM). Besides the obvious usefulness of the model in dealing with a troublesome test design dilemma, there are other advantages: a) the resulting overall score is free of local item dependence artifacts; b) the estimation of the overall score is performed within the item response theory estimation framework and not externally to it, allowing

the calculation of reliable individual person scores and standard errors; c) the overall score has a clear intuitive meaning as a mean of the subdimension student abilities.

In *Assessing assessment literacy: An item response modeling approach for teacher educators*, Duckor, Draney & Wilson tackle the problem of measuring teacher "assessment literacy"—an aspect of teacher craft that sounds harmless but leads into a labrynth of state and federal policy goals and pedagogical and measurement theory. The authors hypothesize at least three teacher "learning progressions" that are part of assessment literacy: a) understanding learning targets, b) understanding assessment tools, and c) understanding data interpretation. Within each learning progression they seek to detect a continuum of tasks that distinguish performance levels of teacher understanding as part of a general Classroom Assessment Literacy (CAL) construct space. The CAL instrument was administered to a sample of teachers and the results analyzed for evidence of reliability and validity and to determine whether the conceptual structure of the learning progressions makes sense. The article is a detailed example of the work involved in conceptualizing, building, refining, and synthesizing Rasch-compliant constructs and, as such, pulls together themes raised in the preceding three articles.

By such studies, we hope to broaden the reach and practicality of Rasch's vision of scientific measurement. To breathe the air of this kind of research, attend the next IOMW conference in New York City in April 2018 (www.iomw.org).

References

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. Applied Psychological Measurement, 21, 1–23.

Belkind, O. (2012). Physical Systems: Conceptual Pathways between Flat Space-time and Matter. Springer (Chapter 5.3) .

Décret relatif aux poids et aux mesures du 18 germinal an 3 (7 avril 1795) [Decree of 18 Germinal, year III (April 7, 1795) regarding weights and measures]. Grandes lois de la République (in French). Digithèque de matériaux juridiques et politiques, Université de Perpignan. Retrieved November 3, 2011.

Mach, E. (1919). "Science of Mechanics" .

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.