frontiers in
PSYCHOLOGY

OPINION ARTICLE
published: 06 May 2013
doi: 10.3389/fpsyg.2013.00246

Additive conjoint measurement and the resistance toward falsifiability in psychology

Moritz Heene *

Department of Psychology, Learning Sciences Research Methodology, Ludwig Maximilian University of Munich, Munich, Germany
**Correspondence: heene@psy.lmu.de*

Edited by:
Andrew S. Kyngdon, NSW Office of the Board of Studies, Australia

Reviewed by:
Joshua A. McGrane, The University of Western Australia, Australia

The history of the past four decades of the theory and application of additive conjoint measurement (ACM) is characterized by vivid developments of its theoretical foundation (cf. Luce and Tukey, 1964; Krantz et al., 1971, 2006; Narens, 1974), industrious developments of statistical and computational implementations (cf. Karabatsos and Ullrich, 2002; Karabatsos and Sheu, 2004; Karabatsos, 2005; Myung et al., 2005) and heated debates about its applicability and significance in psychology (cf. Michell, 1997, 2009; Borsboom and Mellenbergh, 2004; Barrett, 2008; Borsboom and Scholten, 2008; Kyngdon, 2008a; Trendler, 2009). What started as a promising foundation to solve the everlasting debate about the quantitative nature of psychological attributes (Ferguson et al., 1939) ended in perseverative debates with very little transfer to mainstream psychological science still being dominated by structural equation modeling (SEM) and item response theory (IRT). After reading the aforementioned articles, and comparing their implications with the day-to-day business of mainstream psychological science, even an unbiased reader would certainly agree with Cliff (1992) that ACM was a "…revolution that never happened" (p. 186).

It is not the aim of this article, to discredit the efforts of mathematical psychology and proponents of ACM in particular. I just want to address the naïve but relevant question why ACM as a stringent way to formalize and to test the requirements of quantitative measurement in psychology has not been embraced by mainstream psychology as a means to an end to test what they always claim: that most of the attributes (e.g., intelligence and personality factors) are quantitative.

An attribute possessing a quantitative structure is required to satisfy the three conditions of ordinality (transitivity, antisymmetry, and strong connexity) *and* the six conditions of additivity (associativity, commutativity, monotonicity, solvability, positivity, and the Archimedean condition; cf. Michell, 1990, p. 52f.). Most of these conditions are *testable* hypotheses but I have never seen any empirical test in psychological articles before data were analyzed with SEM or IRT models, which already *assume* the quantitative structure of the attributes under consideration as argued below. Somewhere during my psychology studies at the university I learned that psychology is an empirical science and that there is therefore no room for claims that should just be believed. However, given the assumed but almost never tested quantitative nature of most of the psychological attributes as reflected in factor analysis, SEM and IRT models, I must have missed or misunderstood something.

RESISTANCE TOWARD INCONVENIENT TRUTH

The question arises why debates about testing the assumption of quantitative measurement more rigorously emerge from time to time without any broader impact on psychological measurement with a few exceptions (Luce, 2000; Kyngdon, 2011). Any attempt to answer this question will, of course, be incomplete, so that I will suggest a factor that might be of special importance: psychologist's avoidance toward falsifiability and hence, toward inconvenient truth.

A number of authors state (cf. Borsboom and Mellenbergh, 2004; Borsboom and Scholten, 2008; Fisher, 2011) that the axiomatic structure of ACM is too restrictive with respect to the regularities in the order relations of the items, the examinees, and an ordinal index of the probability of a correct response. ACM relates to situations in which one attribute (P; e.g., the probability of getting an item correct) is related *additively* to two others (A the ability and B the item difficulty) such that $P = f(A + B)$ (where f is any positive monotonic function). In fact, the requirements of ACM are rarely fulfilled in applied psychological data (Cliff, 1992; Michell, 2009) because the data must satisfy the highly restrictive conditions of double cancelation, solvability, and the Archimedian axiom (cf. Michell, 1990). Satisfaction of these requirements implies that A and B are additive and are therefore quantitative (cf. Krantz et al., 1971).

I therefore agree with the argument that it is more than questionable why such rigorous measurement structures could be found in psychological data. As illustrated elsewhere (cf. Schönemann, 1994; Heene, 2011) psychology seemed to be overwhelmed by the successful application of mathematics in classical physics and invented "…models with close reference to those of classical physics, which were *then* applied to psychological observations" (Heene, 2011, p. 53; italics in the original). This approach ignores that the development of mathematical models has been closely interwoven with the empirical observation of invariant phenomena in physics implying that the mathematical models have often been derived *from* those phenomena (see also Sherry, 2011).

On the other hand, the tools of mainstream psychology such as SEM and IRT make exactly these strong assumptions about the quantitative structure of psychological attributes. But avoiding any tests of quantitative measurement but applying methods making the assumption of quantity appears to be nothing more than a self-delusion that one bears something valuable instead of being in fact empty-handed. This all too strong tendency to avoid falsification is probably deeply rooted in the scientifically unhealthy political/economical aspiration of psychology (Vautier et al., 2012) which keeps the machine for paper-producing and grant-funding well-oiled but also leading to a severe publication bias. Consider Levine et al. (2009) who showed that effect size and sample size are negatively correlated in 80% of meta-analyses. Consider Fanelli (2010, p. 4) who found that "…the odds of reporting a positive result were around five times higher for papers published in Psychology and Psychiatry and Economics and Business than in Space Science" (see also Fanelli, 2009, 2012; Bones, 2012). Despite these numbers, the possibly best evidence of my claims comes from a logical argument: has anyone ever seen articles using SEM, IRT, or Rasch models in which the author admitted the *falsification* of his/her hypotheses? On the contrary, it appears that stringent model tests are mostly carefully avoided in favor of insensitive "goodness-of-fit indices" (cf. Karabatsos, 2001; Heene et al., 2011).

Given that the empirical foundation for ACM might seldom be given it is then reasonable to apply more flexible measurement models such as the Rasch model (Rasch, 1981) which some authors regard as a *probabilistic* formulation of ACM (Perline et al., 1979) and also leading to interval-level measurement. Kyngdon (2008b), however, argues that there is no basis for this claim by showing that parameters of IRT and Rasch models are only invariant against positive monotone transformations. Thus, if both the Rasch model and the more general three-parameter logistic model fit a data set, only the *order* upon the person ability estimates produced by these models remains invariant. Hence, as only order is preserved under positive monotone transformation (Narens, 1981), the fit of an IRT or a Rasch model, respectively, may in fact not be indicative of quantity, but of order.

Moreover, justification for using the Rasch model relates frequently to the argument that random error forms a fundamental that is, non-ignorable feature of every psychological response process and must therefore be included in any model formulation (cf. Borsboom and Scholten, 2008; Fisher, 2011). Since the Rasch model as a probabilistic model accounts for random error it seems to be the panacea of the measurement problems in psychology. However, the magic of obtaining an interval-scale for items and examinees comes with a price because the Rasch model's status as a quantitative theory is derived exclusively through the error term as Michell (2008) pointed out. With the Rasch model, if the error was eliminated, the slope of the item response curves would become infinite, resulting in step-functions of the Guttman model and the "measurements" of the Rasch model reduce only to mere order. But eliminating error must by definition lead to better measurement, not the impossibility of measurement. Nevertheless, Sijtsma (2012) has recently argued that this reasoning is incorrect:

> The Guttman model divides the latent variable scale into disjoint and exhaustive intervals in which differences $\Theta - \delta_j$ do not affect response probabilities. The Rasch model assumes these differences to have a monotone relationship to response probabilities. From the viewpoint of IRT, the Guttman model ignores the information contained in the intervals, thus paying the price of a lower measurement level. (p. 14)

I do not see why this line of argumentation refutes Michell's (2008) "Rasch paradox". Sijtsma's reasoning *presupposes* that the latent trait is continuous. Furthermore, we can only ignore information "…contained in the intervals" when there already is interval-level information, but this is not at all self-evident but simply an assumption of IRT.

This uncomfortable situation that psychometric models cannot work without "error," has lead in my opinion, to great statistical hand wringing and argumentative acrobatics to avoid falsification of the quantitaty assumption. This line of argumentation is often linked to the demonstration of correspondences between psychology and physics.

For instance, Fisher (2011) claims that the probabilistic nature of the Rasch model reflects the physical phenomenon of stochastic resonance (SR) within a biological system. Simply put, SR states that an output signal-to-noise ratio of a nonlinear threshold system is improved by moderate values of input noise intensity (cf. McNamara and Wiesenfeld, 1989). The weak and normally undetectable signal becomes then detectable due to resonance between the signal and the added stochastic noise because the added noise will occasionally lead to an exceeding of a threshold value of the periodic force (see Gammaitoni et al., 1998, for illustrative examples). A plethora of physical, biological and neurophysiological systems, as well as some phenomena from linguistics and visual perception can be described by SR which has been indirectly shown by applying both the signal and the noise externally to receptors and neurons or by data simulations (cf. Simonotto et al., 1997; Gammaitoni et al., 1998; Moskowitz and Dickinson, 2002).

Although it is intriguing to regard SR as a valid justification for probabilistic item response models in order to capture randomness, such an extrapolation is far-fetched because it is not at all self-evident why and how such micro-level phenomena can be extrapolated to the macro-level of item responses. Moreover, because present results on SR in biological systems bear on indirect evidence, the general applicability of SR to such systems is far from being clear as noted by McDonnell and Abbott (2009):

> Adding noise to external stimuli cannot prove that neurons or brain function depend on consistently available internal sources of randomness, i.e., on endogenous neural noise. The challenge is to devise an experiment that can remove naturally occurring healthy variability and demonstrate that function is impaired solely due to that removal. (p. 6)

It appears that borrowing examples from the natural sciences and relating them to the (error) structure of probabilistic item response models might be a persuading analogy but is not a convincing justification for the probabilistic nature of item response models. Explicit cognitive theories of the test item response process are needed, but psychometrics is profoundly lacking in such theories (Kyngdon, 2011). Furthermore, no experimental evidence currently exists which shows why and how such system-inherent error might occur in the item response process.

Finally, I just wonder why psychometricians have yet ignored the success ACM has within theories of utility and decision making in psychology ("prospect theory"; Kahneman and Tversky, 1979) in which ACM served as a formal proof. While it is true that human choice behavior did not strictly follow the requirements of ACM and research has discovered paradoxes of human choice behavior (Birnbaum, 2008), it is also clear that these observations have led to falsifications of old theories of choice behavior and the development of new ones that account for persistent violations of coalescing and first order stochastic dominance (e.g., Birnbaum, 2008; Luce et al., 2008). Frankly speaking, I have very rarely seen such an attitude within mainstream psychometrics be it IRT/Rasch or SEM where items are omitted from tests, powerless but flattering item-fit statistics are commonly used (Karabatsos, 2001), and correlated error terms are specified (Cole et al., 2007) to get a reasonable model-fit and to construct support for one's own the theory despite doubtful consequences (cf. Bones, 2012; Ferguson and Heene, 2012).

CONCLUSION

Altogether, it is possible that human cognitive abilities and personality traits simply are not quantitative. ACM might be in fact too severe for practical testing purposes. However, psychometricians continue to argue that cognitive abilities are quantitative and measurable "latent traits" (Markus and Borsboom, 2012). If this argument is correct, then once item response error is controlled, test score response data should be consistent with the cancellation axioms of ACM. Thus, more direct experimentation is needed instead of more sophisticated IRT models.

It is still unclear and an unsolved problem what SEM and IRT models, notably the Rasch model, add to the clarification of the quantity problem in psychology. It is furthermore unclear what insights into *empirical phenomena* it provides as even attempts to explain the error structure seem to be premature. It is mostly forgotten that Rasch himself did not derive his model from empirical observations but "…within [Rasch's] own mathematical playground—with no relation to any actual item analysis problem!" (Rasch, 1979). It is not necessarily wrong to develop mathematical models *independently* from empirical observations. But, it is also not at all self-evident that empirical insights will result from such models, be it an IRT, SEM, or ACM. However, by avoiding tests of the assumption of a quantitative structure of psychological attributes, psychologists have yet failed to make progress on the basis of the fundamental scientific principle of falsification and in regard to their most fundamental assumptions of quantitative psychological attributes.

REFERENCES

Barrett, P. (2008). The consequence of sustaining a pathology: scientific stagnation—a commentary on the target article "Is psychometrics a pathological science?" by Joel Michell. *Measurement* 6, 78–123.

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychol. Rev.* 115, 463–501.

Bones, A. K. (2012). We knew the future all along scientific hypothesizing is much more accurate than other forms of precognition—A satire in one part. *Perspect. Psychol. Sci.* 7, 307–309.

Borsboom, D., and Mellenbergh, G. J. (2004). Why psychometrics is not pathological. *Theory Psychol.* 14, 105–120.

Borsboom, D., and Scholten, A. Z. (2008). The Rasch model and conjoint measurement theory from the perspective of psychometrics. *Theory Psychol.* 18, 111–117.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychol. Sci.* 3, 186–190.

Cole, D. A., Ciesla, J. A., and Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychol. Methods* 12, 381.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE* 4:e5738. doi: 10.1371/journal.pone.0005738

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE* 5:e10068. doi: 10.1371/journal.pone.0010068

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics* 90, 891–904.

Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., et al. (1939). Quantitative estimates of sensory events: final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Adv. Sci.* 1, 331–349.

Ferguson, C. J., and Heene, M. (2012). A vast graveyard of undead theories publication bias and psychological science's aversion to the null. *Perspect. Psychol. Sci.* 7, 555–561.

Fisher, W. P. Jr. (2011). Stochastic and historical resonances of the unit in physics and psychometrics. *Measurement* 9, 46–50.

Gammaitoni, L., Hänggi, P., Jung, P., and Marchesoni, F. (1998). Stochastic resonance. *Rev. Mod. Phys.* 70, 223.

Heene, M. (2011). An old problem with a new solution, raising classical questions: a commentary on Humphry. *Meas. Interdiscipl. Res. Perspect.* 9, 51–54.

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., and Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: a cautionary note on the usefulness of cutoff values of fit indices. *Psychol. Methods* 16, 319–336.

Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–291.

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *J. Appl. Meas.* 2, 389–423.

Karabatsos, G. (2005). The exchangeable multinomial model as an approach to testing deterministic axioms of choice and measurement. *J. Math. Psychol.* 49, 51–69.

Karabatsos, G., and Sheu, C.-F. (2004). Order-constrained Bayes inference for dichotomous models of unidimensional nonparametric IRT. *Appl. Psychol. Meas.* 28, 110–125.

Karabatsos, G., and Ullrich, J. R. (2002). Enumerating and testing conjoint measurement models. *Math. Soc. Sci. 43*, 485–504.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1971). *Foundations of Measurement, volume 1: Additive and Polynomial Representations.* New York, NY: Academic Press.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (2006). *Foundations of Measurement: Additive and Polynomial Representations.* Vol. 1. New York, NY: Dover Publications.

Kyngdon, A. (2008a). Conjoint measurement, error and the Rasch model. A reply to Michell, and Borsboom and Zand Scholten. *Theory Psychol.* 18, 125–131.

Kyngdon, A. (2008b). The Rasch model from the perspective of the representational theory of measurement. *Theory Psychol.* 18, 89–109.

Kyngdon, A. (2011). Plausible measurement analogies to some psychometric models of test performance. *Br. J. Math. Stat. Psychol.* 64, 478–497.

Levine, T. R., Asada, K. J., and Carpenter, C. (2009). Sample sizes and effect sizes are negatively correlated in meta-analyses: evidence and implications of a publication bias against nonsignificant findings. *Commun. Monogr.* 76, 286–302.

Luce, R. D. (2000). *Utility of Gains and Losses: Measurement, Theoretical, and Experimental Approaches*. Mahwah, London. Lawrence Erlbaum.

Luce, R. D., Ng, C. T., Marley, A. A. J., and Aczél, J. (2008). Utility of gambling II: risk, paradoxes and data. *Econ. Theory* 36, 165–187.

Luce, R. D., and Tukey, J. W. (1964). Simultaneous conjoint measurement: a new type fundamental measurement. *J. Math. Psychol.* 1, 1–27.

Markus, K. A., and Borsboom, D. (2012). The cat came back: evaluating arguments against psychological measurement. *Theory Psychol.* 22, 452–466.

McDonnell, M. D., and Abbott, D. (2009). What is stochastic resonance? Definitions, misconceptions, debates, and its relevance to biology. *PLoS Comput. Biol.* 5:e1000348. doi: 10.1371/journal.pcbi.1000348

McNamara, B., and Wiesenfeld, K. (1989). Theory of stochastic resonance. *Phys. Rev. A* 39, 4854.

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

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *Br. J. Psychol.* 88, 355–383.

Michell, J. (2008). Conjoint measurement and the Rasch Paradox a response to kyngdon. *Theory Psychol.* 18, 119–124.

Michell, J. (2009). The psychometricians' fallacy: too clever by half? *Br. J. Math. Stat. Psychol.* 62, 41–55.

Moskowitz, M. T., and Dickinson, B. W. (2002). Stochastic resonance in speech recognition: differentiating between/b/and/v. in *IEEE Int. Symp. Circ. Syst.* 3, 855–858.

Myung, J. I., Karabatsos, G., and Iverson, G. J. (2005). A Bayesian approach to testing decision making axioms. *J. Math. Psychol.* 49, 205–225.

Narens, L. (1974). Minimal conditions for additive conjoint measurement and qualitative probability. *J. Math. Psychol.* 11, 404–430.

Narens, L. (1981). On the scales of measurement. *J. Math. Psychol.* 24, 249–275.

Perline, R., Wright, B. D., and Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Appl. Psychol. Meas.* 3, 237.

Rasch, G. (1979). *Letter to Ben Wright Regarding Development of the Rasch Model.* Available online at: http://www.rasch.org/memo1979.pdf

Rasch, G. (1981). *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago: University of Chicago Press.

Schönemann, P. H. (1994). "Measurement: the reasonable ineffectiveness of mathematics in the social sciences," in *Trends and Perspectives in Empirical Social Research*, eds I. Borg and P. Mohler (New York, NY: De Gruyter), 149–160.

Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Stud. Hist. Philos. Sci. A* 42, 509–524.

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory Psychol.* 22, 786–809.

Simonotto, E., Riani, M., Seife, C., Roberts, M., Twitty, J., and Moss, F. (1997). Visual perception of stochastic resonance. *Phys. Rev. Lett.* 78, 1186–1189.

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory Psychol.* 19, 579–599.

Vautier, S., Veldhuis, M., Lacot, É., and Matton, N. (2012). The ambiguous utility of psychometrics for the interpretative foundation of socially relevant avatars. *Theory Psychol.* 22, 810–822.

Received: 30 November 2012; accepted: 15 April 2013; published online: 06 May 2013.

Citation: Heene M (2013) Additive conjoint measurement and the resistance toward falsifiability in psychology. Front. Psychol. ***4****:246. doi: 10.3389/fpsyg.2013.00246*

This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.