Theory & Psychology

Response

Theory & Psychology
2019, Vol. 29(1) 144–151

DOI: 10.1177/0959354318824414
journals.sagepub.com/home/tap
SAGE

Measurability, systematic error, and the replication crisis: A reply to Michell (2019) and Krantz and Wallsten (2019)

Günter Trendler
Worms, Germany

Abstract

Although Krantz and Wallsten (2019) claim that interval and ratio scales abound in psychology, they miss the opportunity to deliver specific evidence for their existence. Michell (2019), on the other hand, misconstrues my objection against the practical usefulness of conjoint measurement (Trendler, 2019). Furthermore, he underestimates the critical role humans play as measurement instruments—that is, as detectors of magnitudes of psychological attributes as derived quantities—and he also misunderstands the meaning of the Millean Quantity Objection. Finally, in answer to Krantz and Wallsten, I specify my position with regard to the connection between scientific stagnation, measurability, and reproducibility.

Keywords

conjoint measurement, derived measurement, psychometrics, Rasch model, representational theory

If confronted with a quantity objection the opponent is in quite a comfortable position. He or she has to point out just one single case where measurement has been successfully established. In essence, what must be demonstrated, for at least one psychological attribute A (e.g., ability), is that the ratio between two magnitudes of quantity of A_1/A_2 is constant. That is, for example, in the case of the Rasch hypothesis $\theta = A/D$, it must be shown that for two persons A_1 and A_2, and for different items D_1, D_2, D_3, … , the ratio $A_1/A_2 = \theta_{11}/\theta_{21} = \theta_{12}/\theta_{22} = \theta_{13}/\theta_{23} = \ldots = const.$ This is what measurement is all about. What should be added as a supplementary requirement to the invariance criterion is that a reported finding must be replicated by at least one independent researcher or research

Corresponding author:
Günter Trendler, Robert-Schuman-Str. 15, 67549 Worms, Germany.
Email: guenter@trendler.org

group (Trendler, 2013; see also Nozick, 2001). Therefore, giving only one example of a firmly established and generally accepted metric measurement scale—and not just present, in the manner of Krantz and Wallsten (2019), a list of publications, potentially containing the evidence—would not only clarify the matter substantially, but it would also set a standard for the attainability of measurement in psychology.[1] It is not up to the critics to search for proof of the existence of measurement in psychology.

Furthermore, it should be noted that the view expressed by Krantz and Wallsten (2019) that "[i]nterval or ratio scales *abound*" (p. 130) is not generally shared among psychologists; in contrast to physics, where the existence of ratio measurement is not contested. For example, in his assessment of the three volumes of *Foundations of Measurement*, Schönemann (1994) does not recognize the abundance described. On the contrary, what he detects is a "virtually perfect absence of empirical support" (p. 150) for axiomatic measurement theories and, in particular, with regard to conjoint measurement he notes that "[w]hatever utility such measurement may have, it is a far cry from 'FM [fundamental measurement] in the same sense that it is possible in physics'" (p. 154). In his comprehensive study of measurement in psychology, Michell (1999) also does not note the alleged plentiful availability of metric scales. This does of course not mean that Krantz and Wallsten's view is incorrect; it may only be a communication problem or some kind of bias on the part of the majority to acknowledge that "scales of the highest repute: interval and ratio scales" (Luce & Tukey, 1964, p. 4) are already available in psychology.

With regard to Michell's (2019) criticism of my argumentation (Trendler, 2019), I would first like to point out that the purpose of investigating the abstract-mathematical and the practical-concrete role of the standard sequence procedure in the representational measurement theory is to illuminate its relation to the classical or traditional concept of measurement. The result of the comparison is that, in essence, the concept of measurement is the same in both theories. This is the light in which my treatment of the method of solving inequalities should be viewed; i.e., in abstract form the standard sequence procedure underlies this "measurement procedure" as well. Neither is it my thesis that, in general, constructing standard sequences is the only practical method to discover ratios between magnitudes of quantity nor, in particular, that the method of solving inequalities necessarily presupposes the construction of standard sequences. Nonetheless, *every measurable attribute can be imagined as a standard sequence*.

So, what is my thesis? My argument is that, *since magnitudes of derived quantities cannot be determined without the help of quantitative indicators, derived measurement is preferable to conjoint measurement, because it is simpler in practical application*. I have also pointed out that the reason why the problem with psychological attributes as derived attributes—i.e., that they are not fundamentally measurable—is not immediately recognized in psychology, is because it is more or less tacitly assumed that humans have the capabilities of measurement instruments. This is, I believe, the main cause of the illusion that all quantities are fundamental quantities; a view which is endemic to the representational measurement theory.

In response to Michell's (2019) objections some specifications are therefore necessary: first, the question is not whether the human body can serve as measurement instrument (e.g., the heart rate for time measurement) or whether the human participant can differentiate between magnitudes of physical stimuli (e.g., light intensity, sound intensity, or

length), but whether humans can unequivocally identify magnitudes of psychological attributes *qua* derived quantities.[2] Second, in order to challenge my view, it is sufficient to indicate one single psychological attribute that is fundamentally measurable, at least on a nominal scale. As pointed out (Trendler, 2009, 2013), if it comes to quantities, nominal measurement is far from a trivial matter. That is, the following question must be answered by specifying a concrete measurement procedure: how can we determine, for instance, if two persons A_1 and A_2 possess the same amount of ability (i.e., $a_1 = a_2$) or how can we find out if the same person A_1 has the same amount of ability at different times (i.e., $a_1 = b_1 = c_1 = \ldots$), so that we can confidently conclude that the same point on the quantitative dimension has been identified (for how the task of identifying "fixed points" is accomplished in physics, see Chang, 2004)?

What also seems to escape Michell's (2019) attention is that, when the quantitative hypothesis is tested, the issue investigated is not only of whether the relevant psychological attributes are quantitative, but what inevitably enters as an auxiliary hypothesis is the question of whether humans have the capabilities of measuring devices, no matter if the test participant or the researcher is aware of this or not. In what sense are humans conceived as measuring instruments? They are considered as such not under any circumstance, but only when it is assumed that the observed behavior conveys directly or indirectly quantitative information about the relevant psychological attributes. This is in general the case when the quantitative hypothesis is tested by asking test participants questions about the position of magnitudes on a quantitative dimension (e.g., Michell, 1990, 1994).[3] More precisely, what is tacitly assumed is that, first, humans have "internally" the capability to determine magnitudes of psychological attributes, compare them for more or less, or determine ratios between them and, second, that they are able to communicate, partly or completely, the result of the "internal" measurement operations "outwardly" to the experimenter. Accordingly, Sixtl (1982) notes, methods for data collection can be differentiated into direct and indirect methods. In the case of *direct methods of data collection*, test participants are required to provide metric information about psychological factors directly (e.g., estimations of ratios between levels of psychological attributes). If *indirect methods* are used, then test participants are merely required to deliver nominal (e.g., yes/no answers) or ordinal data (e.g., judgments about more or less). In this case it is assumed that metric information is provided implicitly.

It is important to understand that if the verification of the quantitative hypothesis fails, it does not necessarily follow that the investigated factors are non-quantitative; it is also conceivable that humans do not have the capabilities of measurement instruments or that as such they are impaired in their function. In short, *the validity of inferences about the theoretical meaning of negative empirical results depends on the issue of the undisturbedness of humans as measuring instruments* (for details on the theory of measuring devices see Janich, 1985). Therefore, in the face of negative empirical evidence, if one does not want to abandon the hypothesis that the investigated psychological attribute is quantitative (e.g., in the cases described in Michell, 1990, Chapters 5–7), one will have to make sure before repeating an experiment that the test participants are valid and undisturbed devices for measurement. In the case of artificial, man-made instruments it is clear how this can be done. But how are we to proceed with human beings? We cannot simply call the craftsman or the mechanic to check and, if necessary, fix them. The only

alternative consists in the assumption that humans are by nature perfect, i.e., undamageable measuring devices. In my view this hypothesis is problematic because in the real world where disturbances abound there are no such things as perfect instruments; i.e., they can always break down, in which case they must be repaired or replaced.

These, then, are in essence the reasons why I think that the hypothesis that humans have the capabilities of measuring devices is unrealistic; though it is logically coherent and though, when considered superficially, it has the appearance of a testable empirical hypothesis. Note that this is a variant of what I have called the Millean Quantity Objection (Trendler, 2009). Michell (2019) questions the power of the objection by stating that "mental phenomena are captured via experimental apparatus (viz. psychological test items), not with the precision physics displays, but with a useful degree of verisimilitude" (p. 141). This misrepresents the meaning of the objection: *The question is not if psychologists need experimental apparatus to capture mental phenomena, but if humans themselves, as test participants, can satisfy the role of experimental or measuring machines, as prescribed by measurement theory*.

On a final note, some clarifying words about the connection between scientific stagnation, measurability, and reproducibility may be permitted. My claim is not that in the history of psychology no real discoveries have ever been made. What I have in mind, when describing contemporary experimental psychology as a stagnant science, is what has been called the "neo-Galtonian research paradigm" (Lamiell, 2003, p. 185). Lamiell explicates that "what is actually analyzed through the statistical techniques proper to neo-Galtonian inquiry (i.e., the data analysis procedures issuing in the putatively explanatory models) is variation around [an] overall mean" (p. 185).[4] Since the advent of the so-called reproducibility debate, it should be clear to everyone that the number of real (i.e., replicable) effects claimed to have been discovered may be strongly inflated.

It is noteworthy that in the meantime the often scientifically questionable quality of "psychological knowledge" is also acknowledged outside the ivory tower of academia as a problem to be dealt with. As was already noted by Ziskin (1970): "psychiatric and psychological evidence … frequently does not meet reasonable criteria of admissibility and should not be admitted in a court of law" (as cited in Faust, 2012a, p. xiii). In particular, as a consequence of the introduction of the *Daubert* standard—which specifies guidelines for admitting scientific expert testimony—"there has been a dramatic increase in litigation concerning whether expert testimony in many different scientific disciplines should be admitted into evidence in courts of law. Psychological expert testimony is frequently the subject of such litigation, in both civil and criminal cases" (Petrosinelli, 2012, p. 36). The reason for this is the finding that, "[m]ental health professionals may claim that their field is a science, with all the weight and prestige connoted by that assertion. In many cases, however, the imputed knowledge of the discipline is based on foundations that are either nonscientific or represent weak or problematic science" (Faust, 2012b, p. 42).

In modern test theory, the problem of the lack of reproducibility and its connection to the question of measurability has been known for a long time. In particular, two scholars, Gerhard Fischer (1968, 1974) and Friedrich Sixtl (1980, 1981, 1982, 1985, 1993, 1998), have addressed the problem. Sixtl (1985), for instance, points out that the arithmetic mean—n.b., under the premise that the relevant psychological attributes are

measurable (e.g., that numbers of items solved N is directly proportional to ability A)—"can indicate the real central value of a parameter" (p. 338) only if the influence of systematic disturbances is negligible. But since "every person represents a unique individual" (p. 338), it can be ruled out that systematic disturbances are in general under control. In consequence, whenever systematic disturbances are active, the mean does not represent the "true value" of a random distribution anymore, but it is "not further interpretable" (p. 322).[5]

Furthermore, the means obtained by repeating the same experiment with different samples will unpredictably fluctuate depending on the unique composition of each sample. As Sixtl (1985) notes, depending on the distribution of the organism variable O in a sample, one can "produce almost any mean" (p. 321), so that with different samples even antithetical hypotheses may be found to be empirically "true." The reason for this is that instead of depending on a specific value of O, the observed variations in reaction "depend on the distribution of the organism variable; they are therefore artifacts of the respective population or sample of individuals. This explains the lack in replicability of empirical findings in the behavioral sciences" (Sixtl, 1981, p. 63).[6] Accordingly, Sixtl calls the commonly shared view that the mean is "a reliable measure of a stable characteristic" (Speelman & McGann, 2013, heading 5), "the fundamental error of contemporary psychology" (Sixtl, 1998, p. 525) or the "myth of the mean" (Sixtl, 1993, p. 399). As argued, *a solution to the problem of measurement intrinsically implies a solution to the problem of systematic error* (Trendler, 2009).

These are, in short, the reasons why measurement matters. Unfortunately, the problem of measurability is not perceived as the primary cause of the failure to replicate, but what has been identified instead as the main issue is an inappropriate and dysfunctional use of established methods of statistical analysis (Asendorpf et al., 2013; Francis, 2012). Therefore, what will be found if the neo-Galtonian path is pursued—even if updated and refurbished (Asendorpf et al., 2013; Borsboom & Cramer, 2013; Epskamp, Rhemtulla, & Borsboom, 2017; Resnick, 2018; Zwaan, Etz, Lucas, & Donnellan, 2018)—is that the signals formerly believed to have been discovered, will eventually vanish in the noise.[7] But until then, many articles will be published, much taxpayer's money will be spent, and great academic careers will be made and yet, justifiably so, the public's perception of psychology (Ferguson, 2015) will not improve.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. As a model of how the evidence may be presented in a simple and readily understandable manner, I would like to suggest, for example, Pouillet's (1856, pp. 629–632) description of the experimental confirmation of Ohm's law.

2. The usefulness of the investigations into the sensory perception of physical stimuli is beyond dispute. Only because of such investigations we know that "our sensations of every kind depend upon so many variable conditions, that for all scientific purposes we prefer to form our estimate of the state of bodies from their observed action on some apparatus whose conditions are more simple and less variable than those of our own senses" (Maxwell, 1871, p. 3). And it is because of the uncontrollable interaction between individual and contextual variability that no progress towards the measurability of sensations will result from "tests of the conjoint commutativity axiom for additive conjoint measurement" (Luce & Steingrimsson, 2011, p. 379) or similar approaches; as, I believe, is already sufficiently demonstrated by the fact that since Fechner, psychologists have not gotten any closer to achieving the objective.
3. The most common cases are instances of *measurement by fiat.* This concept applies whenever methods of data analysis are used, which require that the data satisfy metric scale requirements (e.g., calculation of means), but without actually having or providing evidence that they really do.
4. The "statistical techniques" are usually summarized under the heading "methods of multivariate analysis." They are based on the calculation of basic statistics (i.e., mean, variance, correlation) and they comprise popular procedures like the *t*-test, analysis of variance, regression analysis, factor analysis, cluster analysis, or path analysis (Rencher & Christensen, 2012). The most recent technique—the latest thing, so to speak—which has to be added to this list, is network analysis (e.g., Borsboom & Cramer, 2013; Epskamp et al., 2017).
5. It should be noted that this is not a problem of an "optimal" sample size. When it comes to systematic disturbances it doesn't matter if a sample is "small" or "large." No matter how Big the Data, what counts is not the quantity, but the quality of the data. Therefore, paradoxically, increasing the sample size may actually decrease the power of an experiment.
6. In a similar vein, Krantz and Wallsten (2019) note: "A failed replication study may differ from the original one by sampling from populations that differ on variables such as age, sex, experience, or culture that only later are seen to be relevant" (p. 133). Systematic disturbances are also the reason why meta-analysis is useless.
7. The first analysis of "pre-registered" studies already seems to indicate that this is what really may be happening; namely, what it shows is a sharp rise in null findings (Warren, 2018).

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., … Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119. doi: 10.1002/per.1919

Borsboom, D., & Cramer, A. O. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*, 91–121. doi: 10.1146/annurev-clinpsy-050212–185608

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford, UK: Oxford University Press. doi: 10.1093/0195171276.001.0001

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika*, *82*, 904–927. doi: 10.1007/s11336–017–9557-x

Faust, D. (Ed.). (2012a). Preface. In *Coping with psychiatric and psychological testimony. Based on the original work by Jay Ziskin* (6th ed., pp. xiii-xviii). New York, NY: Oxford University Press. doi: 10.1093/med:psych/9780195174113.001.0001

Faust D. (2012b). Criteria for appraising scientific status I: *Daubert* factors. In D. Faust (Ed.), *Coping with psychiatric and psychological testimony. Based on the original work by Jay*

Ziskin. (6th ed., pp. 42-87). New York, NY: Oxford University Press. doi: 10.1093/med:psych/9780195174113.001.0001

Ferguson, C. J. (2015). "Everybody knows psychology is not a real science": Public perceptions of psychology and how we can improve our relationship with policymakers, the scientific community, and the general public. *American Psychologist*, *70*, 527–542. doi: 10.1037/a0039405

Fischer, G. H. (1968). Neue Entwicklungen in der psychologischen Testtheorie [New developments in psychological test theory]. In G. H. Fischer (Ed.), *Psychologische Testtheorie* [Psychological test theory] (pp. 15–158). Bern, Switzerland: Huber.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* [Introduction to the theory of psychological tests]. Bern, Switzerland: Huber.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*, 975–991. doi: 10.3758/s13423–012–0322-y

Janich, P. (1985). *Protophysics of time*. Dordrecht, the Netherlands: Reidel.

Krantz, D. H., & Wallsten, T. S. (2019). Comment on Trendler's (2019) "Conjoint measurement undone". *Theory & Psychology*, *29*, 129–137. doi: 10.1177/0959354318815767

Lamiell, J. T. (2003). *Beyond individual and group differences*. Thousand Oaks, CA: Sage.

Luce, R. D., & Steingrimsson, R. (2011). Theory and tests of the conjoint commutativity axiom for additive conjoint measurement. *Journal of Mathematical Psychology*, *55*, 379–385. doi: 10.1016/j.jmp.2011.05.004

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1–27. doi: 10.1016/0022–2496(64)90015-X

Maxwell, J. C. (1871). *Theory of heat*. London, UK: Longmans, Green, and Co.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Erlbaum.

Michell, J. (1994). Measuring dimensions of belief by unidimensional unfolding. *Journal of Mathematical Psychology*, *38*, 244–273. doi: 10.1006/jmps.1994.1016

Michell, J. (1999). *Measurement in psychology*. Cambridge, UK: Cambridge University Press.

Michell, J. (2019). Conjoint measurement underdone: Comment on Günter Trendler (2019). *Theory & Psychology*, *29*, 138–143. doi: 10.1177/0959354318814962

Nozick, R. (2001). *Invariances: The structure of the objective world*. Cambridge, MA: Belknap Press of Harvard University Press.

Petrosinelli, J. G. (2012). *Daubert* and the psychological expert testimony: An attorney's perspective. In D. Faust (Ed.), *Coping with psychiatric and psychological testimony: Based on the original work by Jay Ziskin* (6th ed., pp. 28–38). New York, NY: Oxford University Press. doi: 10.1093/med:psych/9780195174113.001.0001

Pouillet, M. (1856). *Éléments de physique expérimentale et de météorologie* [Elements of experimental physics and meteorology] (7th ed., Vol. 1). Paris, France: Hachette.

Rencher, A. C., & Christensen, W. F. (2012). *Methods of multivariate analysis* (3rd ed.) Hoboken, NJ: Wiley. doi: 10.1002/9781118391686

Resnick, B. (2018, August 27). More social science studies just failed to replicate. Here's why this is good. What scientists learn from failed replications: How to do better science. *Vox*. Retrieved from: https://www.vox.com/science-and-health/2018/8/27/17761466/psychology-replication-crisis-nature-social-science

Schönemann, P. H. (1994). Measurement: The reasonable ineffectiveness of mathematics in the social sciences. In I. Borg & P. Mohler (Eds.), *Trends and perspectives in empirical social research* (pp. 149–160). Berlin, Germany: Walter de Gruyter.

Sixtl, F. (1980). Generalized laws of reaction, the average person, and interindividual variation. In E. D. Lantermann & H. Feger (Eds.), *Similarity and choice: Papers in honour of Clyde Coombs* (pp. 100–107). Bern, Switzerland: Huber.

Sixtl, F. (1981). Kritik des verhaltenswissenschaftlichen Experimentierens und Grundzüge einer wirksamen Forschungsstrategie [Critique of experimenting in the behavioral sciences and fundamentals of an effective research strategy]. In W. Janke (Ed.), *Beiträge zur Methodik in der differentiellen, diagnostischen und klinischen Psychologie* [Contributions to the methodology of differential, diagnostic and clinical psychology] (pp. 58–67). Königstein, Germany: Verlag Anton Hain.

Sixtl, F. (1982). *Messmethoden der Psychologie: Theoretische Grundlagen und Probleme* [Measurement methods in psychology: Theoretical foundations and problems] (2nd ed.). Weinheim, Germany: Beltz.

Sixtl, F. (1985). Notwendigkeit und Möglichkeit einer neuen Methodenlehre der Psychologie [The necessity and possibility of a new methodology in psychology]. *Zeitschrift für experimentelle und angewandte Psychologie*, *32*, 320–339.

Sixtl, F. (1993). *Der Mythos des Mittelwertes: Neue Methodenlehre der Statistik* [The myth of the mean: A new methodology for statistics]. München, Germany: Oldenburg.

Sixtl, F. (1998). Der Abschied von Homme Moyen alias Average Person [Taking leave of the homme moyen alias the average person]. In W. Hacker & M. Rinck (Eds.), *Bericht über den 41. Kongreß der Deutschen Gesellschaft für Psychologie in Dresden 1998* [Report on the 41st Congress of the German Psychological Association in Dresden, 1998] (pp. 519–526). Lengerich, Germany: Pabst Science.

Speelman, C. P., & McGann, M. (2013, July 23). How mean is the mean? *Frontiers in Psychology*, *4*, 451. doi: 10.3389/fpsyg.2013.00451

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology*, *19*, 579–599. doi: 10.1177/0959354309341926

Trendler, G. (2013). Measurement in psychology: A case of *ignoramus et ignorabimus*? A rejoinder. *Theory & Psychology*, *23*, 591–615. doi: 10.1177/0959354313490451

Trendler, G. (2019). Conjoint measurement undone. *Theory & Psychology*, *29*, 100–128. doi: 10.1177/0959354318788729

Warren, M. (2018, October 24). First analysis of "pre-registered" studies shows sharp rise in null findings. *Nature*. Retrieved from https://www.nature.com/articles/d41586-018-07118-1

Ziskin, J. (1970). *Coping with psychiatric and psychological testimony*. Beverly Hills, CA: Law and Psychology Press.

Zwaan, R., Etz, A., Lucas, R., & Donnellan, M. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, E120. doi: 10.1017/S0140525X17001972

Author biography

Günter Trendler has a degree in psychology from the University of Mannheim (Germany). Currently he is working as technical employee in the domain of plant design at a leading international industrial services provider.