Manuscript ID: behavsci-243792
Type of manuscript: Perspective
Title: **The EFPA Test-Review Model: When good intentions meet a methodological thought disorder**
Received: 2 November 2017

Reviewer #1 report

1. To make the paper have a high impact (and I think it could), I would suggest coming at "re-doing" sections 10 and 11 of the EFPA Test Review Model from the standpoint of a practitioner. In the US, this means they have had <= 1 course on measurement, <= 1 course on statistics, and maybe 1-2 courses on psychological assessment. For example, I come to p. 17 and read "When working with orders or types, existing or new methods directly suited to demonstrating repeatability of order or type are used." How do I examine this? I don't think the limited measures and arbitrary thresholds in section 10 of the EFPA TRM are necessarily the way to go, but they are easy for practitioners to get an idea of acceptable vs. unacceptable. How would I know if the "repeatability of order or type" from a scale is sufficient? Or even how would "repeatability of order or type" even be indexed?

[Author Response]: I'm not quite sure how to respond here. The argument seems to be twofold:

❶ Practitioners are currently trained in methods and procedures which, as Michell puts things, are a pathology of science, but these methods "are easy for practitioners to get an idea of acceptable vs. unacceptable."

But, the whole point of my article is to show that if one understands the constituent properties of measurement, then the psychometric methods and concepts taught within so many practitioner courses as 'required practices' is no longer advisable except for historical and statistical interest. The methods, the thresholds, the precision accorded to 'psychometrics-based' estimates is illusory. However, if one were to treat the estimates as 'rough'n'ready guides, that's fine – which is probably what many practitioners do in day-to-day practice. However, that is not how these methods are taught, referred to, or advertised among those who teach or promote psychometric methods.

If psychometrics courses began with: 'these methods are simply to enable you to get *an idea of* acceptable vs. unacceptable", then this would not allow the ITC and others to recommend their use as 'best practice' or lay down firm guidelines as to what's acceptable and what isn't. This matters greatly when test scores are used as point-estimates in legal and commercial settings, where the psychometric indices (including standard errors/confidence intervals) are presented as accurate estimates of magnitude of some attribute. The additive-unit metric assumption runs throughout these computations, without any realisation that there is no evidence supporting it, and with no idea expressed as to how such evidence might be forthcoming. So, the wise practitioner/assessment expert would employ methods more suited to the data – approximating by using quantitative methods and treating the results as convenient and pragmatic *(with no claims made about true scores etc.),* using actuarial or other methods which relate a test score to the probability of an outcome, using order-statistics, or as with types, using class-based analytical methods. None of which invoke a true-score or any of the other assumptions of conventional psychometrics.

❷ "For example, I come to p. 17 and read "When working with orders or types, existing or new methods directly suited to demonstrating repeatability of order or type are used." How do I examine this?". The argument here is that practitioners do not know how to assess repeatability of orders and types, because they only know how to assess reliability using the methods of psychometrics (mostly correlations and alpha coefficients). My article is not meant to be a didactic exposition of analytical and computational methods of which many practitioners are unaware. There is a substantive literature of statistical methods working with orders and classes, likewise actuarial methodologies. Furthermore, it is the principle that is important here; that embodied in the question required to be answered by

any practitioner, which I put into context for assessing the reliability of a 'type' designation such as might be required for the MBTI:
"*will an individual obtain the same type-assignment if retested today, tomorrow, next week, next month, next year, over whatever duration is considered relevant?"* p. 17. The same question applies to a test score: "*will an individual obtain the same test -score if retested today, tomorrow, next week, next month, next year, over whatever duration is considered relevant?"*
It is about assessing the discrepancy between two or more scores, ranks, or classes. No more, no less. Sometimes as with my organization's own values assessment, the test result is a sequence of types (value preferences with can consist of one, two, or three values). Here, special computational algorithms have to be created which embody the concept of similarity between 'sequences/sequence components) see:
http://www.cognadev.com/publications/Cognadev_Technical_Report_2_VO_Retest_20_Aug_14.pdf
However, I do also assess "strength and separability" using quantitative methods, but look how they are described and interpreted.

The point being that such an approach gets you the answer you require – without assuming anything that simply makes what you do look more 'precise' than it really is given the properties of the data at hand. Even a novel computational algorithm such as that created for these data is open for scrutiny and sense-checking as to whether or not it really does convey a useful pragmatic assessment of similarity.

I have not included any references to such work in my article because this is just one solution to a particular kind of awkward data. For the current Hogan HDS manual, I used a normalised distance euclidean measure which indexed score discrepancy for retest analyses, and non-metric MDS to compare the configural similarity between applicant and volunteer test-takers.

This is what I refer to as 'case-building' when speaking about validation, part of which includes reliability-repeatability assessment.

Yes, it requires a different kind of measurement/assessment/evaluation course taught to both students and practitioners, which is all about how to construct an evidence-base for an assessment, and includes the context-specific approaches to answering your question: "How would I know if the "repeatability of order or type" from a scale is sufficient?"

But I feel I can't put all of this into my current article, whose only purpose is to set out the background, the reasoning, and recommendations for change. It's long enough as is!

2. p. 17-18 Validation. The only thing discussed here is, for lack of better terms, concurrent/predictive validity. This is important, but what about content or construct types of validity evidence? How does this look coming from a perspective that the what we are measuring is not quantitive?
As an aside, the sentence "No more or less than providing empirical evidence in support of any claim that is made regarding the results reported by an assessment" (p. 17) should be engraved on the door of every test publisher. The latest editions of Wechsler Intelligence Scales are great examples of how this is not being done and no one is holding them accountable.

[Author Response]: I have added in new text on pages 18-19:

For many reading the above, the terms: construct, content, face, predictive, concurrent, and ecological validity are missing from the validation process. At best, the validation seems to be all about concurrent/predictive validity. This is because these terms are now entirely obsolete and irrelevant, along with 'nomological nets'. Borsboom, Mellenbergh, & van Heerden [64] provided the definitive arguments and clarifications for what constitutes validity and validation:

> "Validity is not complex, faceted, or dependent on nomological networks and social consequences of testing. It is a very basic concept and was correctly formulated, for instance, by Kelley (1927, p. 14) when he stated that a test is valid if it measures what it purports to measure …
>
> A test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure." (p. 1061)

And, with regard to the distinction between *validity* and *validation*:

> "This is clear because validity is a property, whereas validation is an activity. In particular, validation is the kind of activity researchers undertake to find out whether a test has the property of validity. Validity is a concept like truth: It represents an ideal or desirable situation. Validation is more like theory testing: the muddling around in the data to find out which way to go. Validity is about ontology; validation is about epistemology. The two should not be confused. "

Finally, construct validity as a concept was shown to be deeply flawed by Maraun [3] in 1998, and was finally rendered obsolete by the powerful arguments in two chapters authored by Borsboom and colleagues and Michell [21, 22] in a 2009 book on test validity. All of which has been completely ignored by psychometricians and those teaching practitioners.

Now I realise the reviewer is probably wanting to know how all these terms could be rejected so robustly by the various authors, and how those teaching psychometrics to others could fail to recognize the implications for how they approached validity; especially given the recent slew of additional work published on this matter:

- Vautier, S., Veldhuis, M., Lacot, E., & Matton, N. (2012). The ambiguous utility of psychometrics for the interpretative foundation of socially relevant avatars. *Theory and Psychology,* 22, 6, 810-822.
- Slaney, K.L., & Racine, T.P. (2013). Editorial: Constructing an understanding of constructs. *New Ideas in Psychology*, 31, 1, 4-12;
- Slaney, K.L., & Garcia, D.A. (2015). Constructing psychological objects: The rhetoric of constructs. *Journal of Theoretical and Philosophical Psychology*, 35, 4, 244-259.;
- Boag, S. (2015). Personality assessment, 'construct validity', and the significance of theory. *Personality and Individual Differences* (http://www.sciencedirect.com/science/article/pii/S0191886914007673), 84, ,36-44;
- Slaney, K. (2017). Validating Psychological Constructs: Historical, Philosophical, and Practical Dimensions. London, UK: Palgrave Macmillan).

Trying to describe why and how construct validity and types of validity have been criticised and eventually set aside would take another article! The best I can do is show the reader where to go to read the most powerful justifications as to why these 'types of validity' are no longer relevant as validity-type 'standards' to which tests must adhere or show. For example, whether content validity is relevant at all depends upon the specific context of an assessment, and whether or not the test-taker

is required to adjudge the content as relevant to something or otherwise for fear of compromising the assessment. i.e. the test-taker's view of the content in the context of the assessment is thought to affect the veracity of their responses.

I hope I've done enough with the new paragraph to meet the reviewer's observation/request … but as before, it may be that my response above remains unsatisfactory.

3. Does it make sense that the "Next Generation of Assessments" should even fall under being reviewed by EFPA TRM? The characteristics described on pp. 11-13 are so different from what is done in psychology (at least in clinical/educational assessment--perhaps I-O too?) that I would think this would call for an entirely different approach to ``test'' evaluation, esp. given the proprietary nature of them. So, while I agree that the EFPA TRM cannot be used for such assessments, I don't think it was ever (or will ever?) designed to evaluate them.

[Author Response]: Exactly my point – not only are the guidelines flawed from a measurement/psychometrics perspective; they are now increasingly inappropriate for a new generation of assessments, some of which no longer contain any items or even an assessment completed by an individual. But the underlying requirements remain: an evidence-base constructed for reliability and validation of an assessment. By setting aside the psychometrics test theory altogether, I have argued it is possible to create a broad framework which applies equally to both current and Next-Generation assessments. The onus is now on evidence-base/case-building for the use of an assessment rather than ticking boxes on forms by using specific psychometric indices/procedures. Because we are not making measurement in the manner of a natural scientist, relying upon base and derived unit quantities, with technical definitions of constructs guiding their use/measurement we have to accept things are more complex and finessed. By reducing the overarching questions to their essence, one for reliability and one for validation, and by clearly separating out investigation of these within a scientific vs a pragmatic perspective, I have tried to set out why contextually-relevant *case-building* is now the optimal pragmatic validation process.

4. The statements about legal challenges may be a bit overstated. I scanned through the latest version of Ziskin's text (6th edition, edited by Faust), and I couldn't find anything in there about the measurement debate (i.e., are the constructs even quantitative?) and I didn't see Michell listed as an "authority." This isn't to say that it won't be in the next edition or that the issues discussed in this paper are not informative for a legal context, just that I don't see the are-the-constructs-even-quantitative issue being part of many legal challenges in the immediate (Perhaps I am wrong, here; if so, the author should give some example legal cases where this has been an issue)

[Author Response]: The point here is that the Ziskin example showed what needed to be done in the context of clinical judgement being deployed as 'factual evidence' within criminal and parole courts/sessions. The acceptance by clinicians of the necessity to back up opinion with empirical facts was continually stalled by them, so Ziskin used the legal system as an indirect approach to 'forcing' clinicians to have to justify the evidential/factual status of their professional judgments, not to their peers and colleagues but to an impartial judge in a court of law. He used a substantive volume of empirical psychology experiment evidence to attack the expected/usual responses of clinicians under adversarial questioning of their judgements, presenting to the court the contrast between what a clinician concluded using 'expert' clinical judgement compared to the available empirical evidence concerning the capability of a clinician to form accurate clinical judgements. As the actuarial evidence concerning the relevant accuracies of clinical vs mechanical predictions within forensic and general psychology became more widely available, so did that enter the evidence-bases now used within the Ziskin approach. Within a few short years, the clinical judgement of risk of recidivism risk without overt reliance upon available empirical evidence of risk was rendered inadmissible as evidence within the forensic domain.

I used the Ziskin narrative as the exemplar of what might be required in areas of psychometric assessment, where psychometricians and those making claims about test scores are ignoring facts about measurement in order to pursue their claims of measurement. Within the internal professional 'conversations' and usages by other psychologists and researchers, the ignorance/avoidance of such facts has no real consequence other than to perhaps stifle scientific innovation and progress in an area of exploration. However, when test scores can have adverse real-world consequences for some, such as in employment settings, risk of adverse event prediction, or as I noted in the manuscript-literally a life or death outcome in US cases where a test score is used as a threshold for learning disability and subsequent offender punishment in murder cases, this is where the reliance on psychometric properties of scores may now be challenged. I have recently been an expert witness in a judicial review of learning disability in the New Zealand High Court (June, 2017) – where I argued (against the Crown and its experts) that no empirical evidence existed to support the use of an IQ score as a quantitative estimate of intelligence, and certainly not it's crude use as a point-estimate threshold indicator of learning disability *(even allowing for the use of confidence intervals, which are themselves predicated upon the attribute varying as a quantity);* presenting many of the articles quoted in my manuscript as background evidence substantiating my criticism. However, the review judgement is still pending so there is no case-judgement reference as yet to quote in my manuscript.

This legal approach would not be in Ziskin's 3-volume series as Michell's work was published 15 or so years after the 5th edition of that series. Furthermore, this is not the focus of those three volumes, which were more concerned with the veracity of clinical and psychiatric judgements. I'm simply extending the principle into domains in which test scores based upon 'best practice psychometrics' might be presented as 'quantities', or the product of quantitative methodologies, in an employment or other court which requires expert testimony to be based upon fact rather than beliefs/opinion.

All I can say right now is that the judge found the adversarial questioning of the Crown experts based upon my affidavit of more than idle interest. Indeed under cross examination the opposing experts admitted they could no show empirical evidence to the court that IQ varied as an equal-interval attribute, but that they did not treat it like this as 'practitioners' anyway *(they had my affidavit weeks before the review, so they knew what was in it).* My point was that as a rough'n'ready indicator of cognitive functioning, an IQ score or even subtest scores were best interpreted as crude partial orders; good enough for most pragmatic/clinical work, but not sufficiently precise to use as point-estimates of capability.

Anyway, I hope I have provided a suitable response for the reviewer. I accept that the reviewer may still feel that "The statements about legal challenges may be a bit overstated", but this is all very new … in fact this article presents the first indications that the measurement claims of many psychologists may no longer be a matter for private discussion and debate among themselves, but might now be questioned by barristers (advised by measurement experts such as myself) in open court, where the usual hand-waving defences (as per Frank Schmidt's comment) will no longer suffice. However, the use of test scores as definitive indicators of attribute magnitude within the context of a legal challenge of potential adverse outcome is not common, but can happen as with challenges by disaffected employees required to undertake psychometric testing as part of having to re-apply for fewer jobs in a downsizing context. But that's' why there is no case-law as yet employing this line of attack as yet – I'm announcing its potential with this article.

Minor
1. p. 14 "the very issue that Trendler [43, 44, 19]...Trendler [19] has"
I think the 19 should be 20
Yes, sorry. I've corrected this.

2. Reference [20].. "Conjoint Measurement undone"

Is this accessible to the public somewhere? I tried to find it, but could not. What happens if it is not published?

It has passed 1st round reviewing with the journal Theory and Psychology – with a positive recommendation for publication subject to satisfactory responses to reviewer comments from Gunter. But, it is not openly available for download as yet, which is why all I could do was reference it as 'submitted'.

Anyway, thanks to the reviewer for taking time out to look at my article. It's not an easy one to consider. I've tried to respond as best I can – but accept that my responses may still be considered inadequate.

Reviewer #2 Report

The purpose of this study is to challenge the test standards and guidelines issued by the European Federation of Psychologists' Associations (EFPA) in 2013 and provide some new framework for researchers. In general, this paper is well written.

Some of the issues pointed out by the author have been discussed in the literature, such as the measurement scale and the latent variable models as validation tools. In this paper the author goes further and challenges the foundation of the whole psychometric filed, challenging the concepts and assumptions of reliability and validity as well as the methods and procedures widely adopted by researchers. The author has made many interesting and good points. Personally I also found I am troubled by the existent validation and reliability estimation methods in some of my psychometric work. I agree with the author that we should redefine reliability and validity and discuss the issues that we have been facing for years. We should also provide some solutions and call for new approaches to meet the needs of newly appeared or appearing assessment tools. I would like to discuss a few issues brought up by the author.

1. Can the author make it clear what is new about your "scientific framework?
Lines 557-570: The author proposed his "scientific framework", which basically includes phenomenon or indicators detection, causal explanatory theory construction, defining the construct under the study, and identifying the causes to explain the variations in the proposed measure of the construct. However, I don't find anything new here. This procedure is not much different from what researchers have been practicing. For example, researchers go through the same procedure to develop their theory models via EFA.

[Author Response]: There is nothing new about the scientific framework I propose – it is just a restatement of the mindset and perspective required to investigate a concept as a scientist might, rather than someone more interested in pragmatic utility or as a psychometrician might. I am in essence contrasting how researchers go about developing theory models via EFA *(where at no time in that process do authors seek to determine and test what is causal for an attribute magnitude, how variation occurs i.e. what causes magnitude changes in an attribute)* and the kind of variation, as quantities, partial orders, synergistic etc.) vs how a scientist might go through the same process 'detecting' a phenomenon, but then seek to test theories of why it occurs and how it occurs. So, if we propose a concept such as 'grit' can be detected, then after the phenomenon detection process the scientific approach is to investigate a theory of what is thought to be causal for its observed variations, and whether those variations can be modeled as quantities or orders. None of this resembles how psychologists proceed with validation of 'detected' phenomena. Mostly, they end up correlating predefined equal-interval/quantitative scale scores with other scale scores or other behaviours, under the thematic process of creating a 'nomological net'. Then other psychologists will redefine the concept by adding in a few more items to some scales, call the concept the same name, and squabble over whose measure of 'grit' is the true measure, usually deploying different kinds of factor or latent variable models to justify their own views. This is what Michell refers to as the 'pathology of science'. It is as Tryon et al stated – the quotation I set out on page 5:

Tryon [13] in a response to Ferguson's [14] article on the negative public perceptions of psychology as a "real science", argued that for psychology to attain greater credibility:

> "We need more, not less, by way of modern causal mechanisms. Pennington (2014) asked and answered the question of what is required to provide a scientific explanation in the following way: "What does it mean to explain something? Basically, it means that we identify the cause of that thing in terms of relevant mechanisms" (p. 3, emphasis added). Psychologists claim to have mechanism information, but as Tryon (2014) and Tryon, Hoffman, and McKay (2016) explain, these claims are mainly false and illusory. For example, the very popular biopsychosocial model

> is just a list of relevant factors; it explains nothing more about psychology and behavior than a glass–metal–petroleum model would explain about how automobiles work. Listing variables, or ingredients, does not constitute explanation. We place variable names in boxes and draw arrows among the boxes, thereby imputing causality that is never explained. Drawing arrows does not constitute explanation. We say that squared correlations explain variance when they only account for variance. Accounting is not explaining. We use brain scans to identify brain lobes that are associated with psychological functions, but we cannot explain how those brain lobes do anything psychological any better than phrenologists could. Associations are not explanations. We identify mediators by correlational methods and discuss them as causal mechanisms. Correlation cannot establish causation. These "explanations" are "illusions of understanding". p. 505.

Let me ask you a question: who investigates the **causes** of magnitude variations within individuals varying in behaviours we call “conscientiousness”? Sure, many researchers provide correlations between scores and various other phenomena, including fMRI parameters etc., but who is truly investigating what is causal for those variations as a scientist might for something like temperature or electrical current? Psychologists have already predefined Conscientiousness as a quantity. But has anybody even tested the accuracy of this assertion?

As Michell (2008) put it:
“Survey the psychometric literature: It reveals a body of theories, methods, and applications premised upon the proposition that psychological attributes are quantitative but is devoid of serious attempts to consider relevant evidence for that premise. The theories proposed (such as the factor analytic theories of cognitive abilities and personality) are typically quantitative; mainstream psychometricians typically believe that they are able to measure abilities, personality traits, and social attitudes using psychological tests; and within applied psychometrics, tests are typically promoted using the rhetoric of measurement. Yet, there is little acknowledgment that this premise might not be true: No research programs that I know of are dedicated to testing it; no body of evidence is marshalled in its support (indeed, as far as I know, none exists); and no attempt has been made to devise methods for diagnosing the difference between quantitative and merely ordinal attributes. Psychometrics is premised upon psychological attributes being quantitative, but this premise is rarely treated as raising questions, usually only as answering them.” P. 8.
Michell, J. (2008). Is Psychometrics Pathological Science? *Measurement: Interdisciplinary Research & Perspective*, 6, 1, 7-24.

My feeling is that you perhaps disagree with much of the above, and my setting out what constitutes a scientific rather than pragmatic approach to seeking ‘validity’ of a proposed construct and its measurement. It may be you consider what currently takes place by so many researchers working with scales as scientific, and so cannot see any distinction. I beg to differ – as does Tryon et al, Maraun, and Michell at least. That’s fine – this is one of the points of my article – to set out a clear distinction that has legal implications when a score is introduced into a court of law as being ‘valid’ because it’s ‘construction’ meets certain professional guidelines and standards. I realize not everyone will see it as “clear”. But this is a ‘perspective’ article after all with a sting in its tail. That is, when asked to provide evidence that an attribute varies as a quantity in a court of law *(where the score has potential adverse real-world consequences for an individual)*, what will a psychologist or psychometrician provide in defence of that claim?

2. Can the author clarify the measurement scale issue?
The author challenges the measurement scales used currently in psychometrics. For example, in Lines 671-674 the author pointed out, "…*classical and modern test theory psychometrics was predicated on the assumption that all psychological attribute magnitudes vary as continuous or integer equal-interval entities;*" This is not true. For instance, researchers in psychometrics and statistics have developed discrete latent variables, such as latent class analysis and latent profile analysis, which assumes that the underlying variables are categorical.
Furthermore, in Lines 572-587 the description of his procedure for the scientific framework still falls in the conventional measurement scales though the procedure is modified.
[Author Response]: With regard to the latent class/profile methods, I stated:
classical and modern test theory psychometrics was predicated on the assumption that all psychological attribute magnitudes
I'm not referring to classes, but magnitudes. Latent classes and profile analysis are merely that, just ways of classifying individuals into types, no different from any inductive-classifier, whether Kohonen neural networks, exhaustive hunter-algorithm computational rule-based classifiers, and indeed any other machine-learning or other rule-based method for determining classes of objects or people. This is not test-theory' psychometrics or even 'measurement', merely classifier analysis. But since Maraun dealt with the nonsense of 'latent variables' years ago (Maraun, M.D. (2007). Myths and Confusions. http://www.sfu.ca/~maraun/myths-and-confusions.html; Maraun, M.D., & Halpin, P.F. (2008). Manifest and latent variables. *Measurement: Interdisciplinary Research & Perspective*, 6, 1, 113-117.) – I don't even consider latent class/profile analysis as particularly relevant techniques (compared to more direct rule-based methods). But that's just me, I do not accept any latent variable theory as being useful or even desirable. It's an unnecessary postulation that serves to confuse rather than assist the production of accurate explanatory psychological theory.

Again, I'm sure we are in complete disagreement on these issues. I'm taking a very harsh line but that's because with what has been published on these matters over the last 20 years or so, nothing is changing in psychology/psychometrics because no-one has actually said 'enough' so firmly. That is what I'm setting out to do in this article and that is why it will be disquieting to many. It will cause some to question their assumptions – but also take these arguments, facts about measurement, and logic to those who choose to remain aloof from such matters, because this time they may find themselves in an adversarial legal situation having to defend what I'm claiming is indefensible.

With regard to:
Furthermore, in Lines 572-587 the description of his procedure for the scientific framework still falls in the conventional measurement scales though the procedure is modified.
Point 1 and 2 in those lines are what can be proposed any investigator. But the scientific approach to exploring validity requires the kind of investigative activity as laid out in the sentence above these points:
it is a matter of empirical experimentation to demonstrate that what an investigator has proposed as causal for magnitudes of *grit* results in the numerical or otherwise defined magnitudes indicated by the assessment.

Let me repeat that quote from Michell (2008) again:
"Survey the psychometric literature: It reveals a body of theories, methods, and applications premised upon the proposition that psychological attributes are quantitative but is devoid of serious attempts to consider relevant evidence for that premise. The theories proposed (such as the factor analytic theories of cognitive abilities and personality) are typically quantitative; mainstream psychometricians typically believe that they are able to measure abilities, personality traits, and social attitudes using psychological tests; and within applied psychometrics, tests are typically promoted using the rhetoric of measurement. Yet, there is little acknowledgment that this premise might not be true: No research

programs that I know of are dedicated to testing it; no body of evidence is marshalled in its support (indeed, as far as I know, none exists); and no attempt has been made to devise methods for diagnosing the difference between quantitative and merely ordinal attributes. Psychometrics is premised upon psychological attributes being quantitative, but this premise is rarely treated as raising questions, usually only as answering them." P. 8.
Michell, J. (2008). Is Psychometrics Pathological Science? *Measurement: Interdisciplinary Research & Perspective*, 6, 1, 7-24.

This paper was always going to test the resolve of reviewers! The issue for me is whether what I'm saying is so wrong that it has to be corrected, or that as you put it: "I would like to discuss a few issues brought up by the author." My feeling is that some of what I've written has done exactly that I set out to do – engage the reader because some of what I'm saying does not accord with their own perspective. It's why I provided so many references – so interested readers can seek further clarification from the true experts in this area.

3. One minor suggestion: It may be better to rephrase the sentence, "To do requires a very technical/explicit definition of the construct…" (Lines 590-592)

[Author Response]: I've rephrased this as:
"To do so requires a technical/explicit definition of the construct; as within physics, a precise specification of what it is and how it should be measured, and what is causal or could conceivably cause such precise additive-unit variations in its magnitudes." – in blue in the manuscript text …

4. Can the author clarify what he suggested in "R2: Evaluate Reliability" (Lines723-753)? Would it be practical or sufficient if we only had one type of reliability estimation method, i.e., test-retest? If they didn't have repeated measures design, researchers would never be able to provide reliability evidence. In addition, if we don't use Pearson-based correlations, what specific methods or procedures would the author suggest? I think the suggestions from the author for reliability is fuzzy, not sufficient and the suggestion on eliminating all other types of reliability estimation methods is too extreme.

[Author Response]: Can't make it any clearer!! If they didn't have repeated measures design, researchers would never be able to provide reliability evidence. Exactly my point, they are not providing reliability evidence except in a clumsy obtuse fashion which is predicated upon specific versions of psychometric test theory being valid – which as I'm stating, is pathological science. Reliability is about estimating the discrepancy between observations over two occasions, in the same way as engineers, other scientists, and indeed the world over speak of assessing 'reliability'. A more formal explanation is provided in a Technical report I provided on Cognadev's Values Orientation assessment, where I went through the motions of providing Alpha, Theta, and Omega reliability estimates as 'commercial rubber-stamps' for those who want to see these kinds of indices (http://www.cognadev.com/wp-content/uploads/2017/02/Cognadev_Technical_Report_3_VO_Reliability_16_Sep_14.pdf). However, the first two pages set out what's so very wrong with these psychometrics conceptualisations, starting with Guttman's very first article on these matters. I also took apart the notion that ICCs and correlations were valid estimates of rater reliability (http://www.pbarrett.net/issid/issid2009.html), while introducing a class of distance-based and 'shaped-distance' coefficients. These latter have since been extended into waveform filter-function methods for determining discrepancy magnitudes between observations.

I'm caught here in considering whether this article should provide didactic material on methods – very much in line with the request and my response to Reviewer #1, point ❷. For me, it is the principle

which needs to be considered (retest or bust) – once over the intellectual hurdle of considering reliability estimation as an assessment over at least two occasions, then consideration of methods which can be used to assess discrepancy as 'repeatability' is straightforward *(for those who might teach 'methods' to others or claim to be 'quantitative' psychologists)*. I'm sure the reviewers are both seeing what this implies given I'm not invoking any test theory or indeed any statistical data models here, from which inferences might be drawn. This is very much a 'what you see in this sample of data is what you've got' approach.

I didn't include the technical report or conference references as these are not peer-reviewed publications, but I can add them in for exemplar/informational purposes if the reviewers think that these would be of assistance to the reader?

5. A general comment on quantitative latent variables:
Psychologists face a challenge—measuring something that cannot be measured directly, so that they have to construct latent variables and add meanings to these variables. I agree partly with the author on his view of latent variables. Sometimes, it is hard to quantify these variables and put them on equal interval or ratio scales, so we have to put them on ordinal or nominal scales. However, if researchers can collect enough evidence and make sense out of the quantitative latent variables, we should acknowledge this type of quantitative latent variables.

[Author Response]: Mike Maraun has previously dealt with all this (It is referenced in the manuscript (#24):

Maraun, M.D., & Gabriel, S.M. (2013). Illegitimate concept equating in the partial fusion of construct validation theory and latent variable modeling. *New Ideas in Psychology*, 31, 1, 32-42.

Abstract

There has come to exist a partial fusion of construct validation theory and latent variable modeling at the center of which is located a practice of equating concepts such as construct, factor, latent variable, concept, unobservable, unmeasurable, underlying, hypothetical variable, theoretical term, theoretical variable, intervening variable, cause, abstractive property, functional unity, and measured property. In the current paper we: a) provide a structural explanation of this concept equating; b) provide arguments to the effect that it is illegitimate; c) suggest that the singular reason for the presence of construct in the literature of the social and behavioral sciences is to mark an allowance taken by the social and behavioral scientist to obliterate the concept/referent distinction that is foundational of sound science.

And with reference to:

However, if researchers can collect enough evidence and make sense out of the quantitative latent variables, we should acknowledge this type of quantitative latent variables.

[Author Response]: Why? What exactly constitutes empirical evidence for the proposition that an ad-hoc latent variable varies as a quantity? It is 'ad-hoc' because it is merely a derived mathematical function of data in a dataset. It's why we have so many 'latent variable' measures correlating moderately with one another but all claiming to 'measure' the same construct. The reality is we can make sense of simple sum scales or even ordered class scales just as well as we can a super-duper IRT/factored item scale. But the latter carries with it the notion of a measurement precision which is never tested, but merely asserted. That matters when reliance on that untested precision has an adverse impact on someone in the world outside of academic discourse.

Again, this issue has been discussed at length by Michell (2012):

Michell, J. (2012). "The constantly recurring argument": Inferring quantity from order. *Theory and Psychology*, 22, 3, 255-271.

"The inference from order to quantity is fundamental to psychometrics because the sorts of attributes that psychometricians aspire to measure are experienced directly only as ordered and, yet, it is concluded that such attributes are measurable on interval scales (i.e., that they are quantitative). This inference has been a feature of psychometrics since early last century, before which it permeated scientific thought and played a role in the development of psychophysics. Despite this, its cogency has been analysed only rarely. Elsewhere, I have argued that it is not deductively valid, a point that might be considered obvious except that attempts have been made to show otherwise. Its invalidity displayed, it is easily shown that it is not inductively reasonable either. However, it might still be urged that the inference from order to quantity is an inference to the best explanation: that is, that quantitative structure is reasonably abduced from order. I argue that the opposite is true: the most plausible hypothesis is that the sorts of attributes psychometricians aspire to measure are merely ordinal attributes with impure differences of degree, a feature logically incompatible with quantitative structure. If so, psychometrics is built upon a myth."

I included this as reference #12 in the manuscript.

6. A general comment on IRT model:

The author provided several examples to demonstrate that Rasch IRT model is a failure and suggested to abandon it. Personally I am not a fan of Rasch model either, but I do not want to exclude this tool. It is still useful in some situations. This is similar to how doctors diagnose a patient's problem. They will use different tools and collect a variety of evidence for their diagnosis.

When we conduct a scientific research project, we don't want to only show one piece of evidence and make a conclusion. The machinery numbers we obtain from Rasch model may just give us nuisance like what the author and Wood have demonstrated, but they may provide useful information. We should use them together with other evidence or information and make a judgement about whether the numbers make sense or not.

Ok, a perfectly reasonable comment. But note there is no claim being made here by the reviewer that Rasch scaling 'produces' quantitative measurement of a psychological attribute, but rather, such a scaling might have pragmatic value as a formalised/systematic method of cumulative scaling of item responses. In contrast, psychometricians who promote Rasch scaling do so as a way of showing how it produces a quantitative measurement 'latent variable' scale from item responses. As a useful approach to developing a magnitude scale which can provide 'good enough for practical use' estimates of ordered magnitudes, it would be silly to 'ban it'. But I was not arguing for that, but merely as I closed my discussion off on page 9:

Such IRT scaling models as the Rasch model may have pragmatic value, but they clearly hold no scientific value.

However, to soften this statement, I have rephrased it as:

Such IRT scaling models as the Rasch model may have pragmatic value, but it is unclear what scientific value they may possess, given they are blind to the semantic content of items, and as Michell [8] notes:

> "Item response modellers derive all quantitative information (as distinct from merely ordinal) from the distributional properties of the random 'error' component."

The more random error in a dataset, the more likely a model will fit that dataset.

(bottom of page 9).

I hope that is more in line with the reviewer's sentiments?

Finally:

I'm sorry to have written such a long response to both reviewers. I know it's a real pain for a reviewer having to plough through long responses like this. But the reviewers both made important and thought-provoking observations. I've had to respond accordingly, especially when trying to justify a lack of response in the manuscript to some of their observations.

I sometimes feel this needs a monograph or small book to make all of these points and address every reasonable concern/perspective expressed by the authors. But I just wanted to get across the almost 'chalk and cheese' position in which the EFPA and other professional guideline societies find themselves; address the key issues with referenced materials, and set out a framework and the potential for a legal challenge given the issues associated with the lack of 'evidence' for certain claims.

My current empirical work in this area is embodied within test manuals (Hogans and Cognadev), some presentations, and some personal computer programs I'm trying to make usable and useful for others (as with the Gower Index - http://www.pbarrett.net/Gower/Gower.html and http://www.pbarrett.net/KSD/KSD.html).

Anyway, I just hope I've done enough to meet with a possibly grudging "OK – I don't agree with all of what he's said in reply – but I can see what's he's driving at".