House of Cards?

A critique of the methodology of the "Big Five"

John Raven[1]

Version date: 21 April 2024

Abstract

This paper discusses problems with the "Big Five". These occur at every level from the wording of the items through the deployment of factor analysis, Likert scales, and coefficient Alpha to the naming of the resultant scales. In many cases the problems stem from failure to follow the guidance of the APA Task Force on Statistical Inference at the most basic level before applying more "sophisticated", but poorly understood, (mostly computerised), statistical procedures.

*****

So far as I can see, the whole, multi-million $, Big Five "personality" structure suffers from severe structural deficits at every level from the ground up. (Pierce, 2019, has made a somewhat similar claim).

Wording of the questions themselves (i.e. the foundations of the enterprise). And some implications.

In their most pervasive format, most of the questions in the questionnaires on which this structure is built take the form of:

"How **strongly do you agree or disagree** with the following question 'I am overweight"?

At root, such a question makes little sense. What does it mean to say that "I **strongly** agree with this statement? What additional information does it convey over simply agreeing?

I can say that "I am very overweight", "I am somewhat overweight", "My weight is OK", "I am underweight", "I am very underweight", or "I don't know whether I am overweight or not – it depends on whose standards"[i].

So, if I am forced, in order to move on, to answer the question as it stands, I have to misconstrue the question and answer a question that I had not actually been asked. (See Humpty Dumpty and Alice).

Having done this once it induces a set to, without much thought, just select some answer to the remaining confusing questions based on some hasty unformulated image of what the question behind the question might have been.

This may be OK for simple items like the above, but many of the questions deal with more complex issues.

Few researchers enquire into what people understand by the questions or what they mean by their answers.

The result is the collection of data which tells us little about what the respondent actually wanted to say.

In effect, the information provided by people's answers to such questions is (rightly) deemed to have little meaning in its own right: Only the scale scores are deemed to have meaning.

[1] With thanks to Steve Hughes.

It is difficult to believe that the results of any further analyses based on such garbage data can be much other than garbage. (Although I have to admit that I have sometimes been surprised by the apparent meaningfulness of what does emerge.)

But this is not merely a surface problem.

What are people to do when they want to say something like "I don't know" or "I do not understand the question", but have no option but choose one of the options they are offered?

Maybe they develop a habit of just selecting the central alternative.

If they do that across all items, it generates item distributions having a similar pattern.

And thus high correlations between those items (see Wilkinson *et al*'s [1999] extract from the unpublished final report of the APA Task Force on Statistical Inference for a detailed discussion).

This makes the interpretation of any use made of those correlations (eg in the course of factor analyses and multiple-regression analyses) problematic.

Failure to examine the basic correlation matrices arranged by factor.

Long before the APA Task Force (and represented by one sentence in Wilkinson's paper) urged us to do so, we had made a habit of printing out the basic correlation matrices with the items ordered as per the output from a factor analysis (see eg Raven et al, 1971). This usually revealed that factor analysis had done surprising things. This deterred us from embarking on discussions, and generating "scales", based solely on factor loadings.

Failure to understand factor analysis itself.

(i) Roots in analysis of forces in physics.

The programs used today have their roots in the programs used by physicists to resolve networks of forces (vectors) acting on an object into a smaller number which could, in turn, be used to identify the direction in which the object would move and/or to identify groups of forces having something in common with those in a cluster but differentiating them from those in other clusters. (In fact, we used these programs running on the National Physical Laboratory's original Atlas computer for our early factor analyses.)

As every schoolboy knows, or used to know, calculating the magnitude and direction of the vector resulting from the operation of a network of forces could be tackled by setting up two arbitrary orthogonal axes, dropping perpendiculars onto them and adding them up. One could then rotate the axes to align one of them with direction in which the object would move (or anywhere else). One could also create oblique axes and align them through clusters of forces to see what happened. So one could, in effect, choose any solution one wanted. As Pierce has shown (actually this is not the reference I really want), this is still the case. The results are thus primarily determined by the preferences of the investigator. Eysenck had great fun with these programs in reducing the vast amounts of data collected by psychologists in the military in WW2 to something one could get hold of. Hence the unscrambling of Introversion from Neuroticism.

(ii) Successive corrections to attempts to re-create whole correlation matrix from smaller number of variables.

I suspect that few people know that, by multiplying the loadings of all items on all factors across the full table of loadings one gets back exactly to the squares of the correlations.

So what one has in an unrotated (Principal Components) solution is an attempt to re-create the whole correlation (co-variance) matrix on the assumption that there is one underlying variable (cause) behind it *and then progressively* ***correcting*** *the estimates so obtained* via the addition of further factors. The subsequent factors are thus not "independent" factors but each dependent on what has been extracted before.

This misunderstanding is reproduced in step-wise multiple regression, which the Task Force claimed had, "like a Siren, lured thousands to their doom".

Given the implications for the careers of the hundreds of thousands of researchers who work/ed in the field, it is therefore perhaps not surprising that Wilkinson made scant reference to these problems … and perhaps not surprising that the report of the Task Force itself has never been published.

(Wilkinson did make a vague attempt to alert people to them by urging researchers to "first look at your data at the most basic level to see whether the application of more "sophisticated" statistical programs is justified").

Failure to understand the measurement model behind "Likert" scaling and the tendency to fish out, and re-brand, any collection of items emerging from factor analysis as a "scale"[ii].

There is no a-priori reason to believe, for example, that someone who repeats 10 times that he or she is over-weight is more pre-occupied with weight than someone who does not. It only shows that he/she is consistent in his/her statements. To make a claim that he/she is "more" preoccupied with this issue than are others, one would have to show that this consideration had overcome a variety of other considerations which might have affected his/her response to a series of items. This means that one would have to show that the items in the proposed scale are tapping into a *variety* of other considerations. In practice, it would be unusual for such a specific factor to emerge from the collection of items that have been fed into a factor analysis of "personality" items. What is the probability that this collection of items satisfies the requirement that responding positively to all the items really means that respondent has an enhanced general disposition to behave in ways indexed by whatever trait that collection of items is said to measure? Can the collection of items really be considered a trait?

Coefficient Alpha.

Despite the general air of scepticism generated by the above discussion, it might have been expected that I would have been reassured by the widespread citings of Coefficient Alpha as "proof" of the internal consistency, or meaningfulness, of the "variables". But this was not the case. Unfortunately, I could not remember why I find this statistic unconvincing. Eventually, I remembered that calculating Alpha for a scale having the properties required by Item Response Theory would produce results as meaningless as those produced by applying factor analysis to such a scale (Raven & Fugard, 2008/2020). However, before I had done that, I did a Google search which came up with McNeish [2017]). This reminded me of the basic problems involved in calculating correlations between items like those cited above (Stevens, 1958). These include the fact that, to form a legitimate basis for calculating correlations, the rating scales for the items must have the properties of interval scales – that is, the difference between eg. "Strongly Agree" and "Agree" must be similar to that between "Agree" and "Neither Agree nor Disagree" (etc.). Furthermore, the distributions should be Gaussian (which, as we have seen, they are not). McNeish then went on to list various critiques of the Alpha statistic … but then, unfortunately (from my point of view), went on to come to a more positive conclusion. I myself continue to doubt the value of this statistic as

confirmation of the meaningfulness of these scales, particularly without examination of the statistics behind the calculation.

The cladding: Over-inclusive and misleading naming of factors and 'Likert scales'.

My favourite example of misleading naming of a constellation of items is "Narcissism". At root, the term means "Excessive love of oneself". But, by a process of progressive extension (see links below), it has, in the Big Five, come to refer to an extraordinary constellation of items having to do with, including, other things, the malicious destruction of people who are vital to the survival of the organisation.

The case provides us with a wonderful example of concept creep[iii].

But the problem is actually pervasive. Small collections of items are given names which suggest that they have wider (theoretical) significance than could possibly be the case. This enhances the impression that the author has made a significant contribution to the advancement of science and promotes wider use of the concept in discussions in the journals.

As Pierce notes, this results in widespread variation in how professionals interpret the scores on the scales.

References

Humpty Dumpty and Alice https://www.thoughtco.com/humpty-dumpty-philosopher-of-language-2670315

Altgassen, E., Olaru, G., & Wilhelm, O. (2023) What if there were no personality factors? Comparing the predictability of behavioral act frequencies from a big-five and a maximal-dimensional item set. *European Journal of Personality*. 2023, Vol. 0(0) 1–15 https://www.researchgate.net/publication/370110725.

McNeish, D. (2017) Thanks Coefficient Alpha, We'll Take it From Here. Psychological Methods 23(3) DOI: 10.1037/met0000144 https://www.researchgate.net/publication/313852796_Thanks_Coefficient_Alpha_We%27ll_Take_it_From_Here

Pierce, M. (2019) A Critique of the Big Five. https://www.youtube.com/watch?v=ohi6cOZ3dmUhttps://www.youtube.com/watch?v=ohi6cOZ3dmU

Raven, J., Ritchie, J., & Baxter, D. (1971). Factor analysis and cluster analysis: Their value and stability in social survey research. *Economic and Social Review*, 2, 367-391. http://eyeonsociety.co.uk/resources/RRAB.pdf.PDF

Raven, J. (1984/1997). *Competence in Modern Society: Its Identification, Development and Release*. Unionville, New York: Royal Fireworks Press. www.rfwp.com (First published in 1984 in London, England, by H. K. Lewis.) Also available at: https://www.researchgate.net/publication/337925795_Competence_in_Modern_Society and http://eyeonsociety.co.uk/resources/Competence-in-Modern-Society-John-Raven.pdf

Raven, J. (2020). *Recent Research Supporting a Specific-motive Based Model of Competence.* (Extended version). https://www.researchgate.net/publication/344947791_Recent_Research_Supporting_a_Specific-_motive-based_Model_of_Competence_Extended_version

Raven, J., & Fugard, A. (2008/2020). What's wrong with factor-analyzing tests conforming to the requirements of Item Response Theory? http://eyeonsociety.co.uk/resources/fairttsts.pdf or https://www.researchgate.net/publication/344947966_What's_Wrong_with_Factor-Analysing_Tests_Conforming_to_the_Requirements_of_Item_Response_Theory

Stevens, S. S. (1946). On the theory of scales of measurement. Science,103, 677-80
Stevens, S. S. (1958). Measurement and man. Science, 127, 383-389
Wilkinson, L., & *Task Force on Statistical Inference*. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604. https://www.apa.org/science/leadership/bsa/statistical/tfsi-followup-report.pdf
Narcissism. https://en.wikipedia.org/wiki/Narcissism
https://en.wikipedia.org/wiki/Narcissistic_personality_disorder

[i] Here is an example from a current Questionnaire.
It comes from the Hogan Questionnaire which has three sets of 5 practice items including:

Question #2:
I regularly stay until the end when at parties

Strongly Disagree
Disagree
Agree
Strongly Agree

This question does not make sense to me.
I can say that I don't usually stay to the end at parties.
But saying that I strongly disagree with the statement does not logically say that - although through some kind of sloppy thinking ... not attending to the question and what one is saying (Humpty Dumpty) ... it can be inferred to say that.
Does disagreeing with it mean that I sometimes stay till the end?

[ii] Of course, there is no reason to believe that a collection of scores on such "variables" is the only, or even the most appropriate, way to capture individual differences (examples would include what Stevens (1958) might describe as "categorical" (viz 'descriptive') forms of "measurement".). Alternative, and perhaps more appropriate, ways of doing so are available as a result of the work of David McClelland. (see e.g. Raven 1984, 2020) and also Altgassen *et al* (2023).

[iii] Haslam, N. (2016). Concept creep: Psychology's expanding concepts of harm and pathology. *Psychological Inquiry*, 27, 1-17 file:///C:/Users/jrave/Downloads/pi16a.pdf