The seductive beauty of latent variable models: or why I don't believe in the Easter Bunny

William Revelle *

Abstract

Seduced by their mathematical beauty, psychologists have been using latent variable models for more than a century. Whether discussing a general factor of cognitive ability, personality, or psychopathology there has been an unfortunate tendency to reify hierarchical structures without examining the utility of alternative models. To some of us, the use of latent variables was an unfortunate mistake. By emphasizing internal consistency rather than validity, parsimony of fit rather than function, the use of latent variables has led psychological measurement and theory down a beautifully seductive garden path rather than focusing on the real problem of actually being useful. I will address some of these alternatives and suggest that it is time to think more critically of the use of latent variable models in our theorizing and applications.

Keywords: Latent Variables, Reliability, Validity, Massively Missing Completely at Random (MMCAR), Scale construction, Factor analysis, Item analysis; Open Source

To receive an award for a lifetime contribution to the study of individual differences is a great honor and an opportunity to review the history and prognosticate on the future of our field. To do so, I am not going to talk about my work so much as challenge a basic assumption that we as a field have been making for the past 80 years, and that is the belief in the power of construct validity and of latent variables. To challenge latent variable models at an ISSID meeting or in its journal is a daunting (foolish?) task and seems to fly in the the face of the amazing contributions of the three prior winners of this award. For all three of them, Hans Eysenck, Arthur Jensen, and Ian Deary were leaders in promoting the power of latent variable models and the theoretical richness that involved.

Hans Eysenck, as a student of Cyril Burt, searched for the latent variables of personality. One of his earliest studies was of the factor structure of behavioral measures among hospitalized soldiers (Eysenck, 1944), subsequent publications continued in this tradition as he married the power of factor analytic techniques to the study of structure and dynamics of personality (Eysenck and Himmelweit, 1947; Eysenck, 1952, 1967; Eysenck and Eysenck, 1985). Besides founding the International Society for the Study of Individual Differences he also founded its flagship journal, *Personality and Individual Differences*. Indeed it was reading his popular publications emphasizing factor analysis and other quantitative techniques (Eysenck, 1953, 1964, 1965) that led me to study personality as a way to combine my interests in mathematics and psychology.

The second winner of this award was Arthur Jensen whose emphasis was upon the 'g' factor of cognitive ability as a higher level latent variable that could organize and explain the structure of cognitive ability (Jensen, 1998). Jensen emphasized the g factor of cognitive ability in terms of the effect of early childhood interventions (Jensen, 1969). From a psychometric point of view, his discussion of what makes a good g remains an essential example of a higher order factor structure (Jensen and Weng, 1994).

Ian Deary (2001) remains a leader in intelligence research, with his collaborators on the MidLothian study of cognition over the life span (Deary, 2009; Johnson et al., 2010). He is both a critic and a supporter of factorial models of cognition. He brought back (Bartholomew et al., 2009) the concept of sampling theory (Thomson, 1935) as a plausible alternative to the hierarchical factor structure so beloved by Spearman.

1. Latent Variables

All three of these researchers worked in the grand tradition of psychometrics and made use of factor analytic techniques. These techniques go back to 1904 with the amazing insights of Charles Spearman. In his two influential papers written while a graduate student of Wundt in Leipzig, Spearman translated the correlation coefficient from the insights of Galton (1888) and the mathematics of Bravais (1844) and subsequently Pearson (1896) to be understandable to psychologists (Spearman, 1904b). In a second article in the same journal, he further developed the basic concepts of reliability, and laid the foundations for factor analysis (Cudeck and MacCallum, 2007; Spearman, 1904a).

*Based upon the Distinguished Contribution Lecture to the International Society for the Study of Individual Differences, July 2023.
Correspondence concerning this article should be addressed to William Revelle, Department of Psychology, Northwestern University, Evanston IL, 60208. Telephone number: 847-491-7700. Email: revelle@northwestern.edu. I would like to thank David Condon, David Funder, Kayla Garner, Lew Goldberg, Robert Hogan, René Mõttus, and Daniel Ozer for their comments and suggestions. This is the author's version as submitted to PAID.

Spearman emphasized the distinction between observed (manifest) and true (latent) correlations and showed how "correcting" for the attenuation due to unreliability (Spearman, 1904a) converted observed correlations ($r_{p'q'}$) into estimates of the "true" correlation (r_{pq}) between various measures of cognitive ability.

$$r_{pq} = \frac{r_{p'q'}}{\sqrt{r_{p'_1 p'_2} r_{q'_1 q'_2}}} \quad (1)$$

This insight of correcting for attenuation and searching for a common factor was used by Webb (1915) in his amazing analysis of ability and character (finding factors of a 45 x 45 correlation matrix by hand was a monumental effort.)

Although not referring to it, Spearman's use of manifest and latent correlations is reminiscent of Plato's Allegory of the Cave (Plato, n.d.). Manifest variables are equivalent to shadows cast on the wall of the cave by people moving in front of a fire. This metaphor is useful when we consider the effect attributed to Flynn (1984, 1987) by Herrnstein and Murray (2010) of manifest intelligence scores increasing by .3 sd per decade which could be seen as analogous to a change in shadow length as people move closer to the fire. That is, manifest variables can change over time with no real change in latent scores.

Spearman's main use of latent variables was to show that the correlations between a number of cognitive abilities showed a remarkable consistency which suggested a latent common factor. This was the introduction of factor analysis as well as test theory. The basic idea was that each observed score reflects a common factor and a specific factor as well as some error. In modern notation this is

$$\mathbf{X} = \lambda_i{}'\theta_i + \S_i + \epsilon \quad (2)$$

where $\mathbf{X}$ is an observed score, λ_i is the correlation of the general factor with a specific item, θ_i is the latent value of an item, $\S_i$ is the item specific factor, and ϵ is a random disturbance. Subsequent work by Thurstone (1934, 1935) introduced matrix algebra to Spearman's tables, and generalized the single factor to multiple factors. Further extensions of Thurstone led to general factors (g), group factors (G), specific factors (S) and random error

$$\mathbf{X} = \lambda'_g g + \lambda'_G G + \lambda'_S S + \epsilon. \quad (3)$$

Because if tests are measured on just one occasion, the specific factors and error are confounded and as the number of group factors increases the relative importance of the general factor will increase. Thus evaluation of the saturation of the general factor was used as a measure of the test's adequacy and estimates were known as measures of internal consistency. With the assumption of just one general factor and no group factors, tests could be evaluated by the amount of general factor saturation as a percentage of total variance

$$\rho_{xx} = \frac{\mathbf{1'}\lambda_i\mathbf{1}}{\mathbf{1'C1}}. \quad (4)$$

where $\mathbf{C}$ is the covariance of the items and $\mathbf{1}$ is a vector of ones. With the further assumption that all λ_i are equal (so called τ equivalence) this estimate is known as λ_3 (Guttman, 1945) or α (Cronbach, 1951). When calculations were done with desk calculators, and finding correlations was tedious and finding factors was even more tedious the charm of these estimates was they could be found from the variance of the total test ($\sigma^2_X = \mathbf{1' \, C \, 1}$) and the variances of the k items ($\Sigma_1^k(\sigma^2_i)$) and did not require finding k * (k-1)/2 covariances. For with k items, and the assumption that λ_i are identical for all items, equation 4 becomes

$$\lambda_3 = \alpha = \frac{k}{k-1}\frac{\sigma^2_X - \Sigma\sigma^2_i}{\sigma^2_X} = \frac{k\bar{c}_i}{1 + (k-1)\bar{c}_i}. \quad (5)$$

If the interitem covariances are found then $\lambda_3 = \alpha$ are functions of the the average interitem covariance ($\bar{c}_i$) and the number of items (k).

Why are these various equations relevant? Equation 2 suggests that items are made up of a latent true score and error and because errors are thought to be uncorrelated, aggregating items increases the internal consistency of the test (equation 5).

With the assumption that items were very noisy Equation 2 led to the tendency to emphasize aggregating items and using a test's internal consistency as an index of factorial validity. Items were thought to be composed on one true factor and error. This belief was supported by the relatively low correlations of items with each other, suggesting that the common variance was low and the error was large. But this ignored the surprisingly high test-retest correlations of items even over several weeks. For instance, when examining the 9 items of the Impulsivity subscale from the EPI (Eysenck and Eysenck, 1964) in the `epiR` data set in the *psychTools* package for R the inter-item correlation is just .11 but the average test-retest correlation over several weeks is .52. (These items are dichotomous. If we find the average tetrachoric values they are .19 inter-item and .74 for test retest.) This pattern of higher test-retest interitem correlations is is also true even for a presumably better set of items (the items measuring Neuroticism) with average inter-item correlations of .15, but test-retest correlations also averaging .52 (.27 and .74 for tetrachorics). Similar findings have been reported for 100 items of the HEXACO with item test-retest correlations over 13 days having a mean value of .65 (Henry et al., 2022). In an unusual design Condon (2022) reports that the stability of items over 15 minutes with 143 intervening items between .6 and .7 for most items. All of these findings suggest that the unique variance of an item is much more stable than previously thought and that aggregating them leads to more than just a pure factor measure for it also includes some of the unique but stable item variance.

1.1. *Common factor analysis*

At the data level, the basic equation for the factor model is that

$$\mathbf{X} = \lambda_i\theta_i + \epsilon \quad (6)$$

where $\mathbf{X}$ is an observed score, λ_i is the correlation of the general factor with a specific item, and θ_i is the latent value of an item, and ϵ is a random disturbance. which can be generalized to general factors (g) , group factors (G), specific factors (S) and random error.

Equation 6 may also be expressed in terms of the factors of a covariance matrix:

$$\mathbf{C} \approx \lambda \lambda' + \Theta^2. \tag{7}$$

Generalizing Equation 6 to the include general, group and specific variance, the observed score on a test item may be modeled in terms of the sum of the products of factor scores (**g, f, s, e**) and loadings (**c, A, D**) on these factors:

$$\mathbf{x} = \mathbf{cg} + \mathbf{Af} + \mathbf{Ds} + \mathbf{e} \tag{8}$$

Ignoring the contribution of specific variance (**Ds**) the reliable variance of the test is that which is not error, the reliability of a test with standardized items should be

$$\omega_t = \frac{\mathbf{1'cc'1} + \mathbf{1'\ A\ A'\ 1}}{V_x} = 1 - \frac{\Sigma(1 - h_i^2)}{V_x} = 1 - \frac{\Sigma u_i^2}{V_x} \tag{9}$$

where h_i^2 is the item communality and u_i^2 is the item uniqueness. The percentage of the total variance that is due to the general factor (ω_g, McDonald, 1999) is

$$\begin{aligned}\omega_g &= \frac{\mathbf{1'cc'1}}{V_X} \\ &= \frac{\mathbf{1'cc'1}}{\mathbf{1'cc'1} + \mathbf{1'AA'1} + \mathbf{1'DD'1} + \mathbf{1'ee'1}} \\ &= 1 - \frac{(\Sigma c_i)^2}{V_x},\end{aligned} \tag{10}$$

where the total test variance (V_x) is the sum of the elements of all the item variances and covariances and $(\Sigma c_i)^2$ is the squared sum of the loadings on the general factor.

Writing such a set of equations reinforces the unfortunate separation between psychometrics and psychology. For, as a leading psychometrician suggests

> Historically, psychological issues have been the driving force behind the development of psychometric methods, beginning most convincingly with the work of Spearman on intelligence, factor analysis, and test-score reliability, and continued by Thurstone, Cronbach, Guilford, and many others. As psychometrics developed into a more mature area, psychometricians began looking for new topics, and these were found in statistics and computer science perhaps more than in psychology. This not only weakened the connection between psychological impetus and psychometric method but also created a psychometrics that was mathematically more demanding for psychologists. The result of this loosened tie in combination with more demands caused many new psychometric tools to go unnoticed in psychology. (Sijtsma, 2009b, p 172).

To which I will add that psychometrics drifted away from the primary mission of helping psychologists develop useful measures and instead became seduced by the beauty of latent variables.

1.2. Scepticism about factors

Although a major contributor to studies of the factorial structure of ability and temperament (Guilford, 1954, 1956), late in his carer J. P. Guilford (1975) suggested that factor analytic results should not be taken without caution.

> In spite of all the negative appearances that factor analysis may give to the critical investigator, I am prepared to reiterate that the method can be a powerful tool to aid in deriving useful psychological constructs. But it cannot do so without theoretical psychological thinking to go with it. There has been entirely too much blind faith, on the part of many who factor analyze, in what factor analysis can do. I sometimes think that its chief value is to enable us to turn data around so we can look at them, from which new insights may arise. But more than that, it can be used to test those insights in a kind of hypothetico-deductive manner. Admittedly, this may not be in a way some investigators would demand. Fortunately, other ways of testing the validity of factorial constructs are available by more ordinary experimental methods. (Guilford, 1975, p 802)

As much as we would want our theories to represent factorially defined constructs and to claim a correspondence between factors and psychological systems (Royce, 1983), it is important to remember that factors are convenient fictions that are merely one way to organize the structure of covariance matrices (Revelle, 1983; Revelle and Ellman, 2016).

> The trend of this discussion suggests a hiatus between the orientations of psychologists who factor analyze. The focus seems to be either in the direction of data or of psychological Constructs, for the empirical versus the theoretical analyst. The empiricist is likely to take the data structure to be the psychological structure. The theorist looks to the data to suggest the psychological structure, recognizing that the two may lack complete isomorphism. The theorist also requires replications with invariance of psychological factors, under somewhat varied conditions, with variations in samples of tests as well as in tested populations. He may also be concerned about relations among factors and possibly about superstructures. "Push-button" factor analysis has not yet achieved a fool-proof program for grinding out invariant, generalized constructs under varied conditions. (Guilford, 1975, p 803)

Indeed, to some, to believe in latent variables is to believe in the Easter Bunny (R. Hogan, personal communication).

2. Construct validity

In partial response to the plethora of scales developed to predict various criteria using e.g., the MMPI (Hathaway and McKinley, 1943) or the Strong Vocational Interest Test (Strong Jr., 1927) and to try to marry psychological theory with scale construction, the 1950's saw three monumental efforts considering the measurement of psychological constructs. Of these, perhaps the best known is that of Lee Cronbach and Paul Meehl (Cronbach and Meehl, 1955) who tried to define a new type of validity: construct validity. This was in striking contrast at the time when validity was typically taken to be how well the test predicted some criterion.

Constructs, as embedded in nomological networks, were seen as theoretical concepts and could only be evaluated in terms of the pattern of correlations. Criterion-oriented validation procedures, on the other hand, harkened back to the operational definitions of behaviorism. Concurrent validity is the correlation with a current criterion. Predictive validity is the correlation with a future criterion. Content validity was established by showing that the test items were a sample of a universe in which the investigator is interested. Construct validation was seen as a never ending process:

> A construct is defined implicitly by a network of associations or propositions in which it occurs. Constructs employed at different stages of research vary in definiteness. ... Many types of evidence are relevant to construct validity, including content validity, interitem correlations, intertest correlations, test-"criterion" correlations, studies of stability over time, and stability under experimental intervention. High correlations and high stability may constitute either favorable or unfavorable evidence for the proposed interpretation, depending on the theory surrounding the construct. (Cronbach and Meehl, 1955, p 200).

An even stronger argument against predictive validity and in favor of constructs was Jane Loevinger (1957) who suggested that to study prediction was not science :

Favorably quoting the economist and statistician Jacob Marshak in his discussion of decision making, Loevinger said: (p 641)

> "A theory provides us with solutions which are potentially useful for a large class of decisions. [...] Hence, the more we know about its properties the better. If we merely want to know how long it takes to boil an egg, the best is to boil one or two without going into the chemistry of protein molecules. The need for chemistry is due to our want to do other and new things " (Marschak, 1954, p 214). She went on to say "The argument against classical criterion-oriented psychometrics is thus two-fold: it contributes no more to the science of psychology than rules for boiling an egg contribute to the science of chemistry. And the number of genuine egg-boiling decisions which clinicians and psychotechnologists face is small compared with the number of situations where a deeper knowledge of psychological theory would be helpful.

To which I will suggest that boiling an egg is sometimes more practically important than spending years studying chemistry.

2.1. The Multi-Trait-Multi-Method Matrix

The third paper in this series emphasizing constructs was by Donald Campbell and Donald Fiske (Campbell and Fiske, 1959) who elaborated on the nomological network and introduced the concept of the Multi-Trait-Multi-Method Matrix (MTMM). They emphasized that it is the pattern of correlations with measures of the same construct measured in the same way (reliability) as well as different ways (convergent validity) as contrasted to measures of different constructs (divergent validity). They were specifically not interested in testing the utility of their measures so much as the convergence of multiple measures of the same construct as indications of validity.

An early example of a MTMM correlation matrix was the set of correlations between self ratings, self report test scores, and and peer ratings on 5 dimensions taken from the (Guilford, 1940) inventory of factors reported by Carroll (1952). As would be hoped, higher convergence was found for traits across methods than for different traits within method. A similar approach to assess the validity of scales was proposed by McCrae et al. (2011) who reported the long term stability of NEO facets, as well as the agreement of self rated facet scores with peer and spouse ratings on those same facets. Although they do not report the discriminative validity presumably they thought of these correlations as the diagonal values of a MTMM and thus as convergent mono-trait-hetero-method validities.

A more recent example of a Multi-Trait-Multi-Method Matrix considers the results of a validation study of traits measured by self report as well as by peer ratings (Zola et al., 2021). From an online sample using Massively Missing Completely at Random sampling of items (roughly 100-200 items per subject from a pool of almost 700 items) data were collected from 158,631 anonymous volunteer participants on items from the SAPA Personality Inventory (spi-135) (Condon, 2018). Correlations were found using the Noah's Ark procedure (pairwise complete). In addition, all participants were asked if they would nominate peers to supply ratings on their personality. Peer ratings were thus collected on 1,554 individual participants who rated 921 of the original participants on a short form of 30 items measuring 8 constructs. Table 1 shows the correlations between five trait measures (α reliabilities on the diagonal). The upper left quadrant of the table shows the correlations of the self report scales, the lower right quadrant the peer ratings. Except for the diagonal elements, these are all multi-trait-mono-method correlations. The lower left quadrant shows the raw correlations of the multi-trait-hetero-method correlations. The values above the diagonal reflect correlations corrected for attenuation. The two minor diagonals reflect the mono-trait-hetero-method validities.

Table 1: Self report and peer report from the SAPA-project. Correlations reported by Zola et al. (2021). Reliabilities on the main diagonal. Raw correlations below the diagonal. Correlations corrected for reliability above the diagonal. Upper left quadrant reflects SAPA Personality Inventory scores (Condon, 2018) for 158,631 participants, mean n/item = 18,180. Other quadrants reflect 908 peer rated participants. Values > .4 are highlighted. Data from the `zola` dataset in the *psychTools* package.

	Self Report					Peer Ratings				
Variable	Agrbl	Cnscn	Nrtcs	Extrv	Opnnn	Agrbl	Cnscn	Stblt	Extrv	IntlO
Agreeableness	**0.87**	0.32	-0.14	0.28	0.09	**0.75**	0.21	0.18	0.34	0.22
Conscientiousness	0.28	**0.87**	-0.20	0.13	0.06	0.16	**0.78**	0.22	**0.42**	0.13
Neuroticism	-0.12	-0.18	**0.90**	-0.28	-0.10	-0.01	-0.16	**-0.78**	-0.40	-0.25
Extraversion	0.25	0.12	-0.25	**0.90**	0.14	0.01	-0.01	0.07	**0.71**	0.14
Opennness	0.08	0.05	-0.09	0.13	**0.86**	-0.14	-0.06	0.10	0.17	**0.49**
Agreeableness	**0.47**	0.10	-0.01	0.00	-0.09	**0.45**	0.36	**0.47**	0.15	**0.44**
Conscientiousness	0.15	**0.55**	-0.12	-0.01	-0.04	0.18	**0.58**	**0.42**	**0.41**	**0.47**
Stability	0.13	0.16	**-0.58**	0.05	0.07	0.25	0.25	**0.60**	0.38	**0.52**
Extraversion	0.23	0.28	-0.27	**0.49**	0.11	0.07	0.23	0.22	**0.52**	0.32
IntellectOpenness	0.14	0.08	-0.15	0.09	0.30	0.19	0.24	0.27	0.15	**0.44**

2.2. *Test Theory*

With the emphasis upon constructs, much of the work in test theory became how to design tests to maximize internal consistency measures of reliability. In contrast to the earlier work by Gulliksen (1950) and Nunnally (1978) which emphasized validity much of the past 60 years has emphasized reliability and internal structure and equated validity with factorial validity. For a discussion of the move towards construct validity and away from simple prediction, see Slaney (2017).

Developments in test theory emphasized unidimensional constructs to be measured with "the New Psychometrics" of Item Response Theory (Embretson, 1996; Embretson and Hershberger, 1999; Reise, 1999) and considered validity in terms of Structural Equation Models (Bollen, 1989; Jöreskog, 1978; Wiley, 1973). IRT is based upon the concept of a latent variable causing the manifest responses to items, SEM is regression with latent variables (observed variables corrected for measurement error). These new approaches have enshrined latent variables without considering the consequences.

Although originally requiring knowing how to code and having familiarity with matrix algebra IRT and SEM procedures have become easier to use without necessarily understanding when and why to use or not use various methods. "One side of the problem is that psychologists have a tendency to endow obsolete techniques with obscure interpretations. The other side is that psychometricians insufficiently communicate their advances to psychologists, and when they do they meet with limited success" (Borsboom, 2006, p 428). The critiques are written in matrix notation in journals such as *Multivariate Behavioral Research* and *Psychometrika* and seem to most nonexperts as debating the number of angels who can dance on the head of a pin.

Our users are taught to push buttons on menu driven programs and to report the statistics that are seen as necessary. They are not taught to think about what these various measures mean in their endless search for construct validity. For "construct validity functions as a black hole from which nothing can escape: Once a question gets labeled as a problem of construct validity, its difficulty is considered superhuman and its solution beyond a mortal's ken." (Borsboom, 2006, p 431)

3. Prediction versus theory

Although classic texts on measurement (e.g., Gulliksen, 1950; Nunnally, 1978) devote entire chapters to issues of validity, more recently there has been less emphasis upon the practical problem of prediction and more on the beauty of equations specifying latent variables. As Hogan (2009) put it "Mainstream psychometrics concerns measuring entities (i.e., determining "true scores"). But applied assessment has a job to do, and that is to predict outcomes."

Although criticizing construct validity Borsboom and Mellenbergh (2004) add an even stronger criticism of criterion validity:

> " the idea of construct validity was introduced to get rid of the atheoretical, empiricist idea of criterion validity, which is a respectable undertaking because criterion validity was truly one of the most serious mistakes ever made in the theory of psychological measurement. The idea that validity consists in the correlation between a test and a criterion has obstructed a great deal of understanding and continues to do so. " p 1065.

They go on to say

> "Therefore, not just criterion validity but any correlational conception of validity is hopeless. The double-headed arrows of correlation should be replaced by the single-headed arrows of causation, and these arrows must run from the attribute to the measurements".

> "Validity is a property of tests: A valid test can convey the effect of variation in the attribute one

> intends to measure. This means that the relation between test scores and attributes is not correlational but causal." p 1067

3.1. *In defense of predictive validity*

In striking contrast to these critiques of predictive validity is the success of several groups of researchers concerned with vocational interests (Dawis, 1992; Donnay, 1997; Holland, 1959; Strong Jr., 1927), psychopathology (Hathaway and McKinley, 1943), or the analysis of "folk concepts" of social interaction (Gough, 1965). Strong Jr. (1927) championed the predictive power of scales formed from items that distinguished members of a particular occupation from "People In General". This completely empirical procedure was adapted by the developers of the MMPI (Hathaway and McKinley, 1943) and the CPI (Gough, 1965). Harrison Gough was interested in predicting such varying criteria of socialization ranging from those seen as "best citizens" to incarcerated felons (Gough, 1965). Whether using the California Psychological Inventory (Gough, 1957) or an Adjective Check List (Gough, 1960) the goal was not a clean factor structure so much as scales that worked.

Perhaps more well known to readers of this journal or members of ISSID is the success of the Hogan Personality Inventory (Hogan and Hogan, 1995). These tests are validated by their success in predicting real world outcomes.

4. Aggregation should be purposeful

We have known since Spearman that test reliability goes up with test length (Figure 1 panel A), as does validity (Figure 1 panel B). This leads us to form progressively longer scales in a hope that irrelevant variance will diminish as a source of test variance.

The classic example of the effects of aggregation is seen with the most used statistic in psychology "coefficient α" (Cronbach, 1951) (Equation 5).This measure is also known as KR-20 (Kuder and Richardson, 1937) or λ_3 (Guttman, 1945). Part of the appeal of α/λ_3 is that it can be found from the item variances and total test variance and is available in commercial software (Sijtsma, 2009a). Although this was convenient in the period of the desk calculator, this is no longer important and so-called model based estimates can be found from the covariances (Equations 9, 10). For fixed average correlation, both α/λ_3 increase with the number of items.

Aggregation can also increase validity by combining k items with average validity $\bar{r}_y$

$$r_{y_k} = \frac{k\bar{r}_y}{\sigma_x} = \frac{k\bar{r}_y}{\sqrt{k + k * (k-1)\bar{r}}}. \tag{11}$$

But there is an interesting contrast between Equations 5 and 11: "What one selects when optimizing predictive utility are items that are mutually uncorrelated but highly correlated with the criterion. This is not what one expects or desires in measurement. Note that this does not preclude that tests constructed in this manner may be highly useful for prediction. It does imply that optimizing measurement properties and optimizing predictive properties are not convergent lines of test construction." (Borsboom and Mellenbergh, 2004, p 1067). That is, there is a tradeoff between internal consistency and validity. This tradeoff may be seen when comparing (Figure 1 panel A) with (Figure 1 panel B). For while both internal consistency and validity increase with the number of items. The highest validity is found for those items that lead to the lowest internal consistency.

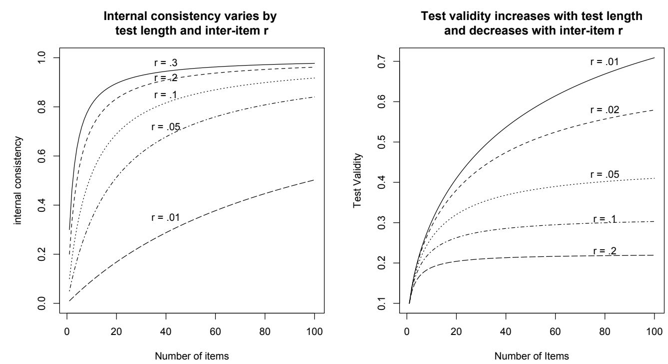

Figure 1: α and validity as a function of the number of items and the average correlation showing the tradeoff between internal consistency and predictive validity;

The power of aggregation is that composite scales can include important variance and reduce the contribution of extraneous error. However, aggregation to maximize internal consistency (Equation 5) will tend to minimize variance that is not random and not common with other items. My colleagues and I refer to such aggregation as spear-fishing – developing sharp, pointed instruments with high internal consistency (Garner, in press; Revelle and Garner, 2023). The alternative approach is to use a net – diffuse scales that include multiple items with criterion validity, even if not highly associated with each other. As we suggest, you can catch more fish with a net than a spear.

Consider the correlations of 10 items from Athenstaedt (2003) that are discussed by Eagly and Revelle (2022) (Figure 2). These items are included in the `Athenstaedt` data set in the *psychTools* package (Revelle, 2023b) for the R statistical system (R Core Team, 2023). The analyses and graphics were done using the *psych* package (Revelle, 2023a) in R. Using the inter-ocular trauma test for the number of factors, these 10 items clearly represent 2 independent factors. Although the sets of items are basically orthogonal, they all correlate with gender. We can find composite scales of these items by combining the first 2, 3, 4 or 5 from each factor (F2 ..., F5, M2 ... M5) or composite scales of 1, 2, 3, 4, 5 from each set (MF2, MF4, MF6, MF8, MF10). (Table 2). Just M or just F scales are very internally consistent (ω_h = .72 ... .85) and reasonably valid (r_{gender} = .52 ... .58). But the composite (MF) scales are much less internally consistent (ω_h = .11 ... .23, α = .11 ... .77) and more valid (r_{gender} = .67 ... .75).

It is interesting to compare the two indicators of internal consistency. The conventional measure for the 10 item MF scales, α, is by conventional criteria (Nunnally, 1978) "acceptable" with values of .77. That is to say, we would expect such a 10 item scale to correlate .77 with a parallel measure. But from the point of view of whether these scales measure one thing, they clearly do not. The ω_h values of .15 suggests that just 15% of the variance is due to one latent factor.

That is, from a traditional measurement point of view, the MF scales are clearly inadequate for they do not represent one construct. Just 11 to 15 % of their variance is common to the scale. But their predictive validity is far superior to that of the "better" scales that are purer measures of a single construct. As Eagly and Revelle (2022) said " the patterning of psychological gender/sex differences can be difficult to discern in narrowly defined attributes but emerges more strongly in general trends. It follows that neither similarity nor difference prevails but instead a more complex intertwining of these two types of findings". This tradeoff between validity and internal consistency is seen in Figure 3 which plots the validity correlations against the ω_h measures of general factor saturation.

We have previously reported similar findings (Eagly and Revelle, 2022) using a data set from Gruber et al. (2020) which also show the power of aggregation and the benefit of aggregating independent dimensions. Whether considering scales of personality, cognitive or behavioral activity, combining uncorrelated measures with high internal consistency produced scales that were much more valid but were clearly not measures of a single latent factor.

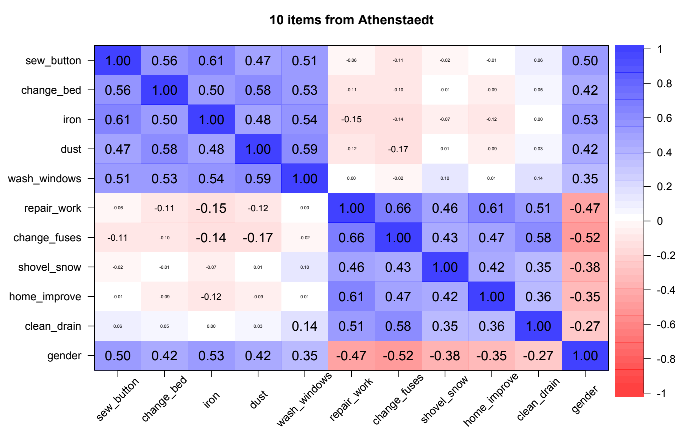

Figure 2: 10 items from Athenstaedt (2003) show a clear two factor structure representing 5 items reflecting feminine activities and five representing masculine activities. Although the first and second set of five items are clearly independent, both sets correlated with gender.

5. Structure of ability and temperament

5.1. Ability

One of Spearman's great contributions was the recognition of the positive manifold of cognitive ability. That is, that measures of cognitive ability are all positively correlated and could

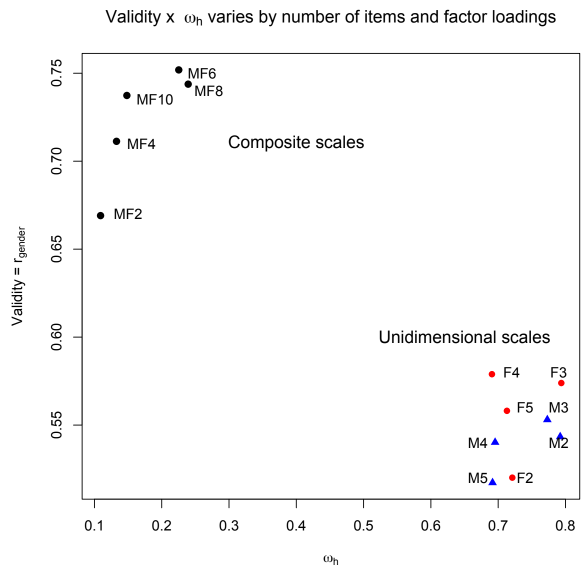

Figure 3: Showing the tradeoff between prediction and internal consistency as indexed by ω_h. The values are taken from Table 2 and are the correlations of 8 unidimensional scales and 5 multidiminsional scales with gender as a function of the general factor saturation ω_h of each scale. The composite scales, although not reflecting a single latent variable, are clearly more valid but less internally consistent than are the unidimensional scales.

be identified by having positive loadings on a general factor (Borg, 2018). This observation should not, however, be taken to imply that there is a general causal factor of ability, for factors are merely one way of representing correlational structure. There are interesting alternative explanations for the positive manifold other than Spearman's g. For as Thomson (1916) pointed out with his independent "bonds" model, rather than one overarching g, tests can correlate because they represent a number of overlapping features. This important idea has been discussed by Bartholomew et al. (2009) and can be simulated by the `sim.bonds` function in *psych*. The Thomson bonds model has also been applied to discussions of the factor structure of temperament items (McCrae, 2014).

Yet another way to achieve a positive manifold has been proposed by Kovacs and Conway (2016, 2019) as multiple processes that grow together. A different development perspective of the meaning of the positive manifold is the observation that that scores on various cognitive measures change at different rates over time (Flynn, 1987). This set of findings calls into question the simple g as primary cause model. The discussion in the last part of that article should be required reading to all who study ability.

Although any positive manifold can be factored to produce lower level (group) and a higher level (g) factor, this says nothing about causality. Higher order factors no more imply causality than the positive manifold of size variables implies a common factor of "bigness" (Figure 4 panel B). As an example of a higher level factor structure in cognitive ability consider the 16 items from the "ICAR sample items" found in the *psychTools*

Table 2: Correlations of item composites corrected for item overlap. α reliabilities on the diagonal. The F and M scales show high correlations within and low between the two sets of scales. e.g., the five F scale correlates .06 with the five item M scale. The data are from Athenstaedt (2003) and are available in the `athenstaedt` dataset in the *psychTools* package. The bottom two lines report the correlations with gender, and the ω_h measure of general factor saturation. See Figure 3 to see the validity and internal consistency trade off.

Variable	F2	F3	F4	F5	M2	M3	M4	M5	MF2	MF4	MF6	MF8	MF10	gendr
F2	*0.72*													
F3	0.75	*0.79*												
F4	0.77	0.80	*0.82*											
F5	0.77	0.81	0.84	*0.85*										
M2	0.12	0.15	0.16	0.14	*0.79*									
M3	0.09	0.12	0.13	0.10	0.75	*0.76*								
M4	0.09	0.12	0.13	0.10	0.77	0.78	*0.81*							
M5	0.06	0.09	0.10	0.06	0.79	0.80	0.81	*0.82*						
MF2	0.36	0.46	0.48	0.48	0.38	0.41	0.45	0.46	*0.11*					
MF4	0.48	0.55	0.58	0.57	0.52	0.51	0.53	0.53	0.46	*0.59*				
MF6	0.52	0.56	0.58	0.58	0.55	0.54	0.56	0.56	0.56	0.66	*0.69*			
MF8	0.54	0.58	0.60	0.59	0.58	0.57	0.58	0.57	0.61	0.71	0.73	*0.75*		
MF10	0.54	0.59	0.61	0.60	0.59	0.57	0.58	0.57	0.63	0.73	0.75	0.77	*0.77*	
gender	0.52	0.57	0.58	0.56	0.54	0.55	0.54	0.52	0.67	0.71	0.75	0.74	0.74	1.00
ω_h	0.72	0.79	0.69	0.71	0.79	0.77	0.7	0.69	0.11	0.13	0.23	0.24	0.15	

package. These items are part of a larger project (the ICAR project) to develop open source ability items. Originally developed by Condon and Revelle (2014) and then working with colleagues in the UK and Germany, the ICAR project now has 17 item types and a database of several thousand items (Dworak et al., 2021; Revelle et al., 2020). These items show the traditional hierarchical structure of ability items (Figure 4 panel A).

This hierarchical structure is remarkably similar to that of 19 measures of physical size taken from the United States Airforce which also show a higher level factor structure (Figure 4 panel B). This factor, best summarized as physical size can not be said to be a cause of arm length or chest diameter. For size is a formative sum of the component measurements.

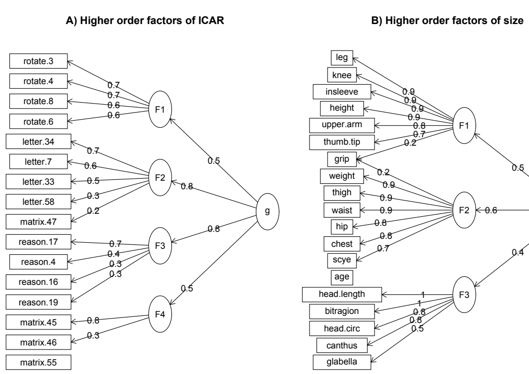

Figure 4: Hierarchical analysis of 16 ability from the ICAR (panel A) and 19 size measures from the United States Airforce (panel B). Data sets in the *psychTools* package are `ability` and `USAF` respectively. Measures of internal consistency: $\omega_h = .66, .53, \alpha = .83, .90, \omega_t = .86, .95$ for ability and size respectively.

5.2. Temperament

Although in the late 1960s, some Americans thought personality did not exist, this was not true in Europe where researchers continued to discuss the genetic and physiological basis of personality (Eysenck, 1967; Revelle, 1989)[1]. Finally, recognizing that perhaps personality traits did indeed show consistency across situations and over time, debates between alternative structural models focused on three (Eysenck, 1990; Peabody, 1967), five (Digman, 1990; Goldberg, 1990; Costa and McCrae, 1992), seven (Comrey, 2008), and even sixteen (Cattell and Stice, 1957) basic dimensions. After a consensus upon a five factor model became somewhat accepted, the debate continued as to whether one general factor (Musek, 2007; Revelle and Wilt, 2013), or two higher order (Digman, 1997) better captures the personality space. The debate continues to this day with some suggesting that the consensual Big Few structure is a useful organizing framework (Bainbridge et al., 2022) while others discuss how this structure is not replicable across cultures, or even within the natural language (Condon, 2023; Cutler and Condon, 2023).

Analogous to the questions of structure in personality is the debate about the structure of psychopathology. Influential work suggesting common factors to personality disorders was based on converting "comorbidities" of diagnostic categories into tetrachoric correlations and then factoring the resultant matrices (Krueger and Markon, 2006a,b; Markon et al., 2005). These findings led to the "HiTOP" model (Forbes et al., 2021) as an attempt of organizing all of psychopathology into a single hierarchical model. However, this organization is not without its critics who suggest the analogy of the 'p' factor of psychopathology with the 'g' factor of ability is incorrect and not helpful (Watts et al., 2023).

Furthermore, that measures of personality and psychopathology can be described as formative rather than reflective indicators (Jonas and Markon, 2016) has major implications to their use. For if they are formative, our latent variables are just descriptive summaries of the items rather than causal (Bollen, 2002; Howell et al., 2007).

[1] For a history of the "dark ages of personality," see Revelle et al. (2011).

6. Prediction

But how much did these debate about personality structure help our understanding of the causes and consequences of personality? Science is about prediction and understanding. The use of latent variables which are factorially pure supposedly helps us understand our variables and further our theories. But how well do these latent variables actually help us predict real criteria? The distinction between prediction and understanding is not new, for it has been raised before (e.g., Mõttus et al., 2020; Yarkoni and Westfall, 2017), but it is worth reminding those of us who were seduced by latent variable that there are important alternatives to theory driven approaches.

Prediction of real world phenomena is hard and effect sizes tend to be small (but important). In their extensive review of the power of personality to predict meaningful criteria (life span, occupational attainment, and divorce) Roberts et al. (2007) showed robust, but small effects. They point out, however, these effects are equivalent in magnitude to the effects of Social Economic Status or cognitive ability. Although it is not clear what specific trait theories predict that prudent and conscientious people tend to live longer and have more stable marital relationships these results are important. They are, however, more descriptive than theory driven findings. They do show that there is something about the aggregation of items assessing prudent behavior that enhances prediction.

Unfortunately, in reviewing the power of personality to predict real outcomes, Roberts and his colleagues ignored an important part of personality: interests. People spend most of their lives working. Knowing what influences their choice of occupation is not just the Big Few or even the Facets or Nuances of traditional personality instruments (Anni et al., 2023). Impressive as the analysis of 263 occupation in terms of personality profiles (Anni et al., 2023) is, they continue in the unfortunate tradition in academic personality research to ignore interests, perhaps because they are seen as too practical and useful.

Seemingly less known to most academic personality researchers is a substantial literature in counseling as well as industrial-organizational psychology that discusses the power of interests to predict job choice (Armstrong et al., 2004; Donnay and Borgen, 1996; Su et al., 2019). Much of this work is "dustbowl empiricism" inspired by Strong Jr. (1927) who spent a lifetime developing scales that predicted satisfaction with jobs. A fairly common organization of the Strong scales (Donnay et al., 2005) is the Realistic, Investigative, Artistic, Social, Enterprising and Conventional (RIASEC) model of Holland (1996) which suggested the six personality "types" flourish in appropriate environments. The six types are said to be able to be summarized in a circumplex with the axes of ideas versus data and people versus things. An alternative representation of the axes is that of Hogan (1982) who posited sociability and prudence as the primary axes. Su et al. (2019) points out that "Interests have also been shown to have incremental validity over cognitive ability and personality traits in predicting job performance" (p 1) and then went beyond the traditional six clusters of the RIASEC to introduce an eight dimensional model (SETPOINT) based upon factor analysis of interest items. Their work is an example of the seductive beauty of latent variables for they go beyond simple empirically derived scales in their attempt at finding a clean CFA structure.

In a practical sense, the question about the utility of theory versus prediction has been answered by the success of companies that develop instruments to predict employee success by using proprietary instruments. Rather than adopt factorially pure instruments with high construct validity, these companies emphasize scales that discriminate successful from unsuccessful workers. Criteria of interest include absenteeism, theft, malacious behaviors and general dishonesty or lack of integrity (Hogan et al., 1996; Hogan and Sherman, 2020). Predictive validity is shown for truck drivers, service dispatchers, or machine operators. The success of this approach may be seen by the number of companies that use these proprietary instruments. Their instruments are broadly theory relevant, e.g., socioanalytic theory suggests that we should study the interpersonal challenges of getting along, getting ahead and finding meaning in life (Gottlieb et al., 2021; Hogan, 1982; Hogan and Blickle, 2018) and emphasize predictive rather than factorial validity. Combining multiple dimensions is better than any single dimension. Thus Hogan et al. (1994) in their review of personality and leadership effectiveness cite literature that surgency, emotional stability, and conscientiouness predict better leadership performance.

The debate about scale construction procedures between those favoring latent variable models, those favoring theory driven models, and those using criterion oriented scales was addressed by Hase and Goldberg (1967) who reached the conclusion that all of these procedures were about equally effective when predicting a variety of criteria. In a monumental followup which also addressed basic scale construction principles, Goldberg (1972) came to somewhat different conclusions, showing how factorially based scales worked better on easy to predict criteria, but that criterion oriented techniques were better with harder to predict criteria. Hase and Goldberg (1967); Goldberg (1972) examined 468 unique items taken from the CPI to predict 13 different criteria for a total sample of just 152 subjects. Being firm believers in the need to cross validate their results, the derivation and cross validation samples had just 76 participants. Using much larger samples, my colleagues and I have found that empirical item level and lower level factor scales dominate high level factor based prediction (Revelle et al., 2021). Here I elaborate on those findings.

6.1. *Examples of prediction at the scale level*

At a more micro level, I have already used the example of predicting gender from various stereotypical gender items (Table 2, Figure 3) to show that increasing internal consistency does not necessarily lead to increases in validity. In fact, there is a well known (but forgotten) tradeoff between the two. I now consider a more complicated example which uses dimensions that are commonly seen in personality research and examine predicting a set of 8 criteria using three levels of analysis (Figure 5).

For reproducibility of my results, I use data from the `spi` dataset in the *psychTools* package and include the relevant R

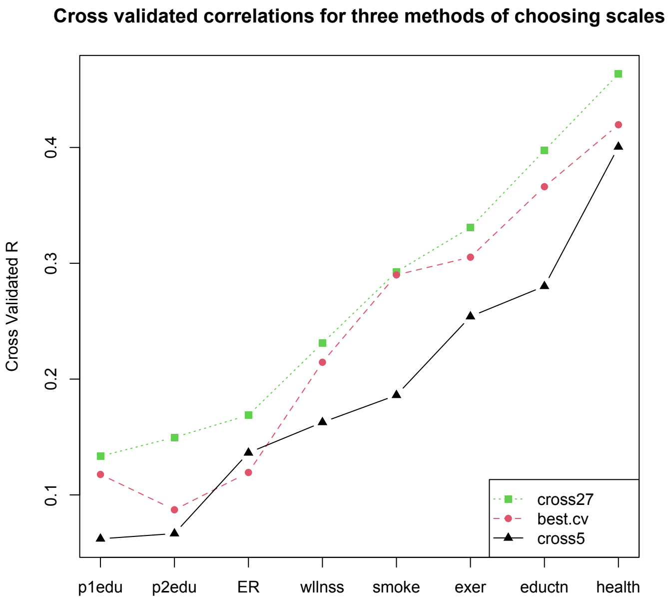

Figure 5: Predicting 8 criteria from the `spi` data set. The values shown are the cross validated Multiple Correlations from five higher order factors, 27 lower level factors, and the `bestScales` solutions. N derivation =2,000; N cross validation=2,000.

code in the appendix. The `spi` dataset was collected as part of the SAPA project discussed earlier and includes 135 items from Condon (2018). These 135 were carefully curated from a larger set of 696 items which in turn were taken from the more than 2,000 items in the International Personality Item Pool (Goldberg et al., 2006). Of these 135 items, 70 may be formed into 5 higher level composites representing the Big Few, while all 135 items can be scored for 27 different lower level item composites. Conventional estimates of internal consistency ($\omega_h, \alpha, \omega_t$) as well as various measures of unidimensional structure (Revelle and Condon, 2023) are shown in Table 3. As expected (Widaman and Revelle, 2023a,b) scale scores found by unit weighting of the keyed items match factor score estimates with all correlations > .97.

Because of the well known need to cross validate any empirical finding (Cureton, 1950), all analyses were done on a randomly chosen 50% of the data and then the resulting β weights were applied to the other 50% of the sample. With the sample sizes I am using, (derivation N = 2,000, cross validation N = 2,000) the amount of shrinkage in the cross validation samples was minimal (compare the multiple R values for the derivation and cross validation samples in Table 5).

For each of these eight criteria, Figure 5 shows the cross validated multiple correlations for scales representing the Big Few, the "little 27", as well as scales formed from finding the best cross validated items using the `bestScales` function. Although all the β values for the 5 and 27 predictors on the 8 criteria are shown in Table 5, for conciseness, I just discuss self ratings of wellness and reported exercise. The three largest β weights suggest that Exercise is done more by people who are high on conscientious, emotional stability and more extraverted. These same three factor based scales predict self ratings of health, but with a bigger effect for emotional stability and an overall larger R. When examining these relationships in more detail, by looking at the lower level factor/scales, we see that Exercise is associated with not being easy going, but being sociable and a seeking stimulation. Health is also associated with not being easy going, but is particularly associated with well being, low anxiety, self control and sensation seeking.

6.2. *Prediction at the item level*

In addition to using higher level and lower level factors/scales, it is also possible to use the items themselves. A graphical demonstration of how subsets of items from each of these higher level or lower level factors relate to the criteria is shown as a pair of "Manhattan" plots (Figure 6). These two plots show the zero order correlations for each item in each scale with the criteria. Thus, although Neuroticism correlates -.27 with health, we can see that this is due to about seven of the 14 items in the scale and the high correlation of well being with health reflects the high correlations of all of the items in that short scale.

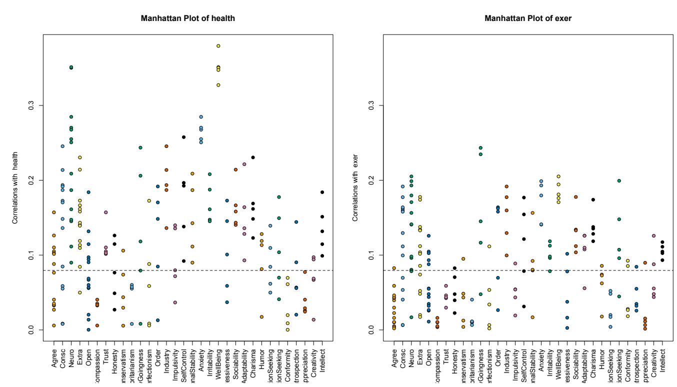

Figure 6: Manhattan plots organize individual item validities by 5 higher order Agree .. Open and 27 lower order factors. The data are the derivation sample from the spi. N=2,000. The dashed line represents the Bonferroni adjusted level of significance at the $p < .01$ level.

A more detailed pattern for exercise and health is found by looking at the items that are most descriptive. A simple "machine leaning" algorithm, implemented in the `bestScales` function identifies those items which are most related to a criterion in each of 10 "folds" of the data. K-fold cross validation splits the data into k folds, and treats N*(k-1)/k participants as the derivation sample and N/k as the cross validation sample. Pooled cross validation coefficients are then used to choose the "best" items. We have compared `bestScales` to more conventional techniques such as LASSO regression and finds that it performs about as well (Elleman et al., 2020). The advantage of `bestScales` is that it is completely transparent and produces a list of the best items for any criteria. Given that SAPA data normally has a high degree of missingness (by design) and that

Table 3: Various estimates of internal structure for 5 "Big Few" and 27 lower level scales from the `spi` dataset. For a list of the items and scoring keys for these scales, see the help page for the `spi` dataset in the *psychTools* package. Calculations done using the `reliability` function in the *psych* package. The first three columns are the traditional measures of internal consistency, the next three represent three measures of unidimensionality, the next two are results of split half analyses and represent the best and worst split half reliabilities. The final three columns report the mean and median inter-item correlations and the number of items per scale.

Variable	ω_h	α	ω_t	Uni	τ	ρ_p	max split	min split	$\bar{r}$	median r	N items
Agree	0.55	0.87	0.89	0.69	0.80	0.86	0.91	0.66	0.32	0.25	14
Consc	0.58	0.86	0.88	0.75	0.84	0.90	0.91	0.70	0.30	0.27	14
Neuro	0.61	0.90	0.92	0.84	0.90	0.94	0.94	0.75	0.40	0.36	14
Extra	0.66	0.89	0.91	0.82	0.89	0.92	0.94	0.77	0.38	0.34	14
Open	0.47	0.84	0.86	0.68	0.77	0.88	0.89	0.62	0.27	0.22	14
Compassion	0.80	0.88	0.89	0.99	0.99	1.00	0.87	0.82	0.59	0.58	5
Trust	0.80	0.87	0.89	0.99	0.99	1.00	0.87	0.81	0.58	0.58	5
Honesty	0.71	0.81	0.84	0.96	0.97	0.99	0.83	0.70	0.46	0.46	5
Conservatism	0.56	0.78	0.85	0.82	0.90	0.91	0.84	0.61	0.41	0.35	5
Authoritarianism	0.63	0.81	0.86	0.89	0.93	0.95	0.85	0.63	0.46	0.46	5
EasyGoingness	0.45	0.68	0.76	0.90	0.92	0.98	0.73	0.58	0.29	0.29	5
Perfectionism	0.34	0.70	0.74	0.82	0.83	0.99	0.72	0.53	0.31	0.33	5
Order	0.62	0.81	0.85	0.92	0.94	0.99	0.83	0.66	0.46	0.42	5
Industry	0.72	0.84	0.86	0.99	0.99	1.00	0.84	0.76	0.52	0.50	5
Impulsivity	0.72	0.87	0.90	0.98	0.98	1.00	0.87	0.80	0.58	0.58	5
SelfControl	0.49	0.76	0.83	0.90	0.94	0.96	0.80	0.60	0.39	0.36	5
EmotionalStability	0.65	0.85	0.89	0.98	0.98	1.00	0.84	0.76	0.52	0.50	5
Anxiety	0.83	0.90	0.91	0.99	0.99	1.00	0.89	0.83	0.64	0.62	5
Irritability	0.78	0.89	0.91	0.98	0.99	0.99	0.89	0.79	0.61	0.60	5
WellBeing	0.80	0.90	0.92	0.99	0.99	1.00	0.90	0.81	0.63	0.63	5
EmotionalExpressiveness	0.73	0.80	0.83	0.92	0.93	0.99	0.83	0.68	0.45	0.43	5
Sociability	0.66	0.85	0.89	0.97	0.98	0.99	0.85	0.75	0.53	0.50	5
Adaptability	0.62	0.80	0.84	0.92	0.93	0.99	0.82	0.68	0.44	0.42	5
Charisma	0.67	0.82	0.86	0.94	0.96	0.98	0.84	0.72	0.47	0.43	5
Humor	0.68	0.78	0.82	0.91	0.92	0.99	0.81	0.64	0.42	0.40	5
AttentionSeeking	0.80	0.88	0.90	0.92	0.93	0.99	0.89	0.77	0.58	0.67	5
SensationSeeking	0.77	0.86	0.89	0.97	0.98	0.99	0.87	0.77	0.55	0.54	5
Conformity	0.67	0.82	0.87	0.89	0.93	0.96	0.85	0.67	0.47	0.47	5
Introspection	0.56	0.78	0.84	0.92	0.93	0.99	0.81	0.68	0.41	0.41	5
ArtAppreciation	0.68	0.80	0.83	0.89	0.90	0.99	0.81	0.65	0.44	0.46	5
Creativity	0.70	0.85	0.86	0.97	0.97	1.00	0.85	0.77	0.52	0.53	5
Intellect	0.81	0.86	0.87	0.99	0.99	1.00	0.84	0.78	0.54	0.52	5

Table 4: Descriptive statistics for the eight criteria used in the examples from the `spi` dataset. The trimmed mean represents the mean with the top and bottom 10% removed. The Mad is the median absolute difference from the median. For a discussion of the estimates of skewness and kurtosis see the help pages for `describe` in the *psych* package.

Variable	vars	n	mean	sd	median	trmmd	mad	min	max	range	skew	krtss	se
health	1	3536	3.51	0.98	4	3.54	1.48	1	5	4	-0.25	-0.42	0.02
p1edu	2	3051	4.72	2.39	5	4.77	4.45	1	8	7	-0.11	-1.33	0.04
p2edu	3	2896	4.33	2.32	5	4.28	4.45	1	8	7	0.09	-1.33	0.04
education	4	3330	4.10	2.21	3	4.00	1.48	1	8	7	0.41	-1.04	0.04
wellness	5	3311	1.54	0.50	2	1.55	0.00	1	2	1	-0.17	-1.97	0.01
exer	6	3310	3.57	1.60	4	3.60	1.48	1	6	5	-0.35	-1.06	0.03
smoke	7	3348	2.19	2.04	1	1.70	0.00	1	9	8	1.83	2.19	0.04
ER	8	3347	1.16	0.48	1	1.03	0.00	1	4	3	3.42	12.74	0.01

Table 5: Standardized β weights for 5 and 27 predictors of 8 criteria. Also shown are the multiple R values for the derivation sample (N=2,000) and cross valdiation sample (N=2,000). Although values $r > .075$ have Bonferroni adjusted probabilities of $< .01$, I highlight those $\beta > .1$. Calculations done with the `lmCor` and `crossValidation` functions in the *psych* package.

Variable	p1edu	p2edu	ER	wllns	smoke	exer	edctn	helth
(Intercept)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Agree	0.02	0.01	-0.03	0.03	**-0.10**	-0.03	**0.11**	0.02
Consc	-0.02	-0.04	0.01	**0.11**	-0.06	**0.15**	0.04	**0.16**
Neuro	-0.04	-0.03	**0.12**	0.02	0.06	**-0.15**	**-0.14**	**-0.27**
Extra	0.05	0.07	0.04	**0.11**	0.07	**0.11**	-0.09	**0.14**
Open	0.09	0.10	-0.03	0.00	0.08	0.05	**0.13**	0.04
R - derivation	**0.13**	**0.14**	**0.12**	**0.17**	**0.17**	**0.28**	**0.24**	**0.41**
R - cross valid	0.06	0.07	**0.14**	**0.16**	**0.19**	**0.25**	**0.28**	**0.40**
(Intercept)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Compassion	0.04	-0.02	0.05	0.05	0.00	-0.02	0.03	-0.03
Trust	0.03	0.07	-0.06	-0.02	-0.09	0.01	0.04	0.03
Honesty	**-0.10**	-0.06	0.01	-0.04	0.02	0.00	0.10	-0.03
Conservatism	0.02	-0.01	0.04	0.04	0.00	0.02	-0.03	-0.01
Authoritarianism	-0.02	-0.04	-0.01	0.05	**-0.16**	-0.08	-0.09	-0.01
EasyGoingness	-0.08	-0.05	-0.04	-0.07	0.05	**-0.17**	-0.05	**-0.10**
Perfectionism	0.03	0.05	0.01	0.02	-0.02	0.01	-0.03	0.02
Order	-0.06	-0.05	-0.04	-0.02	0.00	0.09	0.04	0.05
Industry	-0.06	-0.05	0.00	0.01	**0.11**	-0.02	0.03	-0.01
Impulsivity	-0.05	-0.04	-0.01	0.00	0.04	-0.03	-0.02	-0.05
SelfControl	0.05	0.07	0.03	0.01	**-0.18**	0.08	-0.10	**0.14**
EmotionalStability	-0.07	-0.04	-0.04	0.00	0.06	-0.06	0.05	-0.08
Anxiety	-0.01	0.04	0.06	0.01	0.03	-0.08	**-0.11**	**-0.12**
Irritability	-0.08	**-0.11**	-0.01	0.03	0.01	-0.04	0.00	-0.05
WellBeing	0.10	0.05	-0.02	0.05	-0.05	0.09	0.04	**0.29**
EmotionalExpressiveness	-0.02	-0.03	-0.02	0.05	0.07	-0.07	**0.13**	-0.03
Sociability	0.07	0.06	-0.03	0.00	-0.03	**0.10**	**-0.14**	0.05
Adaptability	-0.03	-0.05	**-0.12**	-0.06	-0.04	-0.02	0.09	0.00
Charisma	-0.07	-0.07	0.05	0.07	**0.15**	0.04	-0.03	-0.04
Humor	0.01	0.05	0.04	0.07	-0.06	0.05	**-0.14**	0.04
AttentionSeeking	0.02	0.09	-0.01	-0.04	-0.01	-0.07	0.10	0.03
SensationSeeking	-0.04	-0.02	**0.14**	0.04	0.01	**0.10**	**-0.18**	**0.11**
Conformity	-0.04	-0.04	0.04	0.06	0.04	0.01	0.07	0.02
Introspection	-0.01	0.05	-0.06	-0.02	0.03	0.03	0.09	0.08
ArtAppreciation	0.06	0.02	-0.04	0.05	0.02	-0.01	0.04	-0.05
Creativity	0.04	0.00	0.09	-0.01	0.02	-0.01	0.00	-0.06
Intellect	0.06	0.06	-0.08	0.04	-0.02	0.00	**0.11**	0.01
R - derivation	**0.23**	**0.25**	**0.24**	**0.24**	**0.32**	**0.37**	**0.41**	**0.49**
R - cross valid	**0.13**	**0.15**	**0.17**	**0.23**	**0.29**	**0.33**	**0.40**	**0.46**

it works on both raw data as well as covariance matrices, we have found `bestScales` to be particularly useful.

Based upon the zero order correlations, we see that Extraverts exercise more (r =.13) or that the linear regression of Extraversion + Conscientiousness combines the need for stimulation with the belief that exercise is healthy, (R = .22). Or we can use lower level constructs that suggest people with a high sense of well being, who are not easygoing and are high in industriousness exercise more (R=.33). Finally, we can find (and cross validate) the items that actually predict exercising (R=.33) (Table 6) or health (R=.43) (Table 7) . All of these are reasonable levels of understanding and prediction. It is important to point out the multiple regressions done with the little 27 were based upon 135 items (5 items per scale), the `bestsScales` results were based upon just the 20 items most related to each criteria.

7. Discussion and Conclusions

The tension between theory and prediction has been with us for many years. Empirically based scale construction using items to predict outcomes is not a new idea (e.g., Hathaway and McKinley, 1943; Stewart et al., 2022; Strong Jr., 1927, 1947) although it seems to have been forgotten by those who prefer constructs and latent variables. The elegance of the arguments for construct validity (Cronbach and Meehl, 1955; Loevinger, 1957) and the sheer pleasure of successfully doing a factor analysis or structural equation model has seduced us from the path towards predicting outcomes.

With the advent of very large data bases and recognizing the need for cross validation, the empirical approach has become popular in other fields. For knowing how to add (find sum scores) is, after all, the basic principle of polygenic risk scores used in Genome Wide Association Studies (GWAS) or in risk scores for medical outcomes. GWAS identifies the single SNPs correlated with outcomes as diverse as height or years of education which are then summed to produce a single score (the PRS). The effectiveness of PRS is evaluated by correlation with the criterion variable. While the effect of each SNP is trivial (but reliable given the sample sizes used), the combined scores have much larger effects. Thus Lee and his colleagues formed a PRS for years of education that could explain 11% of the variance (Lee et al., 2018) from the composite score of 1,271 unrelated SNPs. Not using GWAS, but just combining unrelated predictors is seen in the Environmental Risk Scores for psychosis (Vassos et al., 2020) or the Environment Wide Association Studies to quantify general health risks of environmental pollutants (Park et al., 2014). All of these studies are using SNPs as items in formative measures of risk. They do not posit a latent variable causing the SNPs.

Although most users of SEM think of the items as reflective indicators of latent variables, the alternative is to recognize that many of our latent variables are just formative sums of independent items. I am not denying the power of aggregation to form better measures, I am just suggesting that our measures need to be recognized for what they are: sums of independent items which do not necessarily, and frequently do not, have anything in common. That is, to think of a scale as more than a simple sum and to reify it as some latent variable is to mislead ourselves. With a finite number of items, factor score estimates are not latent variables, they are merely weighted sum scores. Focusing on measures of internal consistency at the cost of focusing on predictive validity is a mistake.

An alternative to the simple factor model of scale construction was proposed by McCrae (2014) in his distinction between scales as the intersection of items versus the union of items. Reconceptualizing our scales as formed from the union of multiple items that carry unique information makes problems in Differential Item Functioning and factorial invariance less challenging than thinking of homogeneous scales all meant to measure one latent construct. Consider the case of sex differences in depression. Items measuring depression (e.g., "In the past week I have felt downhearted or blue" or "In the past week I felt hopeless about the future" have roughly equal endorsement characteristics for males and females. But the item "In the past week I have cried easily or felt like crying" has a much higher threshold for men than for women (Schaeffer, 1988; Steinberg and Thissen, 2006) indicating a much higher level of depression for men who endorse the item. Similarly, lack of factorial invariance across cultures is not a reason to reject a scale, but is a reason to more carefully investigate the pattern of item differences across these cultures. Discussions of DIF in terms of relative versus absolute measurement help clarify the need to examine the meaning of items before leaping to conclusions about factor invariance at the scale level (Borsboom et al., 2002).

7.1. Conclusions

In the preceding pages I have taken the somewhat radical position that our emphasis upon latent variables and construct validity as an attempt to understand the structure of personality has been done at the cost of showing that personality is actually useful. Although it is much easier (and more enjoyable) to talk about theories of Extraversion and Neuroticism (Eysenck, 1967) or Impulsivity and Anxiety (Gray, 1981, 1987), to use these higher level dimensions in predicting real outcomes is difficult. For to predict specific outcomes it is better to resort to short, non-homogenous tests made up of the specific items that actually work. Such scales are formative measures that do not reflect some underlying latent cause, but are merely the observed sums of observed variables. We should stop believing in the Easter Bunny.

Table 6: 20 spi items that best predict exercise. The last two columns identify items that are markers (if they are) of the five higher order factors and then the 27 lower level factors. The item numbers correspond to those from Condon (2019). The item validities are the means of 10 folds. Estimates of internal consistency: $\omega_h = .62, \alpha = .88, \omega_t = .90, u = .69, r_{exercise} = .33$.

Variable	mean r	item	B5	L27
q_1024	-0.24	Hang around doing nothing.		EasyGoingness
q_1052	-0.23	Have a slow pace to my life.		EasyGoingness
q_811	-0.21	Feel a sense of worthlessness or hopelessness.	Neuro	WellBeing
q_1662	0.20	Seek adventure.		SensationSeeking
q_1505	-0.20	Panic easily.	Neuro	Anxiety
q_1371	0.19	Love life.		WellBeing
q_808	-0.19	Fear for the worst.	Neuro	Anxiety
q_1452	-0.19	Neglect my duties.	Consc	Industry
q_2765	0.18	Am happy with my life.		WellBeing
q_4249	-0.18	Would call myself a nervous person.	Neuro	Anxiety
q_312	-0.18	Avoid company.	Extra	Sociability
q_1444	-0.18	Need a push to get started.	Consc	Industry
q_56	0.18	Am able to control my cravings.		SelfControl
q_820	0.18	Feel comfortable with myself.		WellBeing
q_254	0.17	Am skilled in handling social situations.	Extra	Charisma
q_578	-0.17	Dislike myself.	Neuro	WellBeing
q_1254	-0.16	Leave a mess in my room.	Consc	Order
q_1483	-0.16	Often forget to put things back in their proper place.	Consc	Order
q_1979	0.16	Work hard.	Consc	Industry
q_1201	0.16	Keep things tidy.	Consc	Order

Table 7: 20 spi items that best predict health. The last two columns identify items that are markers (if they are) of the five higher order factors and then the 27 lower level factors. The item validities are the means of 10 folds. Estimates of internal consistency: $\omega_h = .64, \alpha = .90, \omega_t = .92, u = .37, r_{health} = .43$.

Variable	mean r	item	B5	L27
q_820	0.38	Feel comfortable with myself.		WellBeing
q_578	-0.35	Dislike myself.	Neuro	WellBeing
q_811	-0.35	Feel a sense of worthlessness or hopelessness.	Neuro	WellBeing
q_2765	0.35	Am happy with my life.		WellBeing
q_1371	0.33	Love life.		WellBeing
q_808	-0.28	Fear for the worst.	Neuro	Anxiety
q_1505	-0.27	Panic easily.	Neuro	Anxiety
q_4249	-0.27	Would call myself a nervous person.	Neuro	Anxiety
q_56	0.26	Am able to control my cravings.		SelfControl
q_4252	-0.26	Am a worrier.	Neuro	Anxiety
q_1989	-0.25	Worry about things.	Neuro	Anxiety
q_1452	-0.25	Neglect my duties.	Consc	Industry
q_1024	-0.24	Hang around doing nothing.		EasyGoingness
q_254	0.23	Am skilled in handling social situations.	Extra	Charisma
q_39	0.22	Adjust easily.		Adaptability
q_312	-0.21	Avoid company.	Extra	Sociability
q_1444	-0.21	Need a push to get started.	Consc	Industry
q_979	-0.21	Get overwhelmed by emotions.	Neuro	EmotionalStability
q_952	-0.21	Get angry easily.		Irritability
q_1052	-0.21	Have a slow pace to my life.		EasyGoingness

References

Anni, K., Vainik, U., Möttus, R., 2023. Personality profiles of 263 occupations. psyarxiv/ajvg2 doi:https://osf.io/preprints/psyarxiv/ajvg2.

Armstrong, P.I., Smith, T.J., Donnay, D.A., Rounds, J., 2004. The Strong ring: A basic interest model of occupational structure. Journal of Counseling Psychology 51, 299–313. doi:10.1037/0022-0167.51.3.299.

Athenstaedt, U., 2003. On the content and structure of the gender role self-concept: Including gender-stereotypical behaviors in addition to traits. Psychology of Women Quarterly 27, 309–318. doi:10.1111/1471-6402.00111.

Bainbridge, T.F., Ludeke, S.G., Smillie, L.D., 2022. Evaluating the big five as an organizing framework for commonly used psychological trait scales. Journal of Personality and Social Psychology 122, 749–777. doi:10.1037/pspp0000395.

Bartholomew, D., Deary, I., Lawn, M., 2009. A new lease of life for Thomson's bonds model of intelligence. Psychological review 116, 567–579. doi:10.1037/a0016262.

Bollen, K.A., 1989. Structural equations with latent variables. Wiley, New York.

Bollen, K.A., 2002. Latent variables in psychology and the social sciences. Annual Review of Psychology 53, 605–634. doi:10.1146/annurev.psych.53.100901.135239.

Borg, I., 2018. A note on the positive manifold hypothesis. Personality and Individual Differences 134, 13–15. doi:10.1016/j.paid.2018.05.041.

Borsboom, D., 2006. The attack of the psychometricians. Psychometrika 71, 425–440. doi:10.1007/s11336-006-1447-6.

Borsboom, D., Mellenbergh, G.J., 2004. Why psychometrics is not pathological. Theory & Psychology 14, 105–120. doi:10.1177/0959354304040200.

Borsboom, D., Mellenbergh, G.J., Heerden, J.V., 2002. Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. Applied Psychological Measurement 26, 433–450. doi:10.1177/014662102237798.

Bravais, A., 1844. Analyse mathématique sur les probabilités des erreurs de situation d'un point. Memoires Presentees al'Academie Royale des Sciences de L 'Institut de France .

Campbell, D.T., Fiske, D.W., 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin 56, 81–105. doi:10.1037/h0046016.

Carroll, J.B., 1952. Ratings on traits measured by a factored personality inventory. The Journal of Abnormal and Social Psychology 47, 626.

Cattell, R.B., Stice, G., 1957. Handbook for the Sixteen Personality Factor Questionnaire. Institute for Ability and Personality Testing, Champaign, Ill.

Comrey, A.L., 2008. The Comrey Personality Scales, in: Boyle, G.J., Matthews, G., Saklowfske, D.H. (Eds.), Sage handbook of personality theory and testing: Personality measurement and assessment. Sage, London. volume II, pp. 113–134.

Condon, D.M., 2018. The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model. PsyArXiv doi:10.31234/osf.io/sc4p9.

Condon, D.M., 2019. Database of Individual Differences Survey Tools. Harvard Dataverse doi:10.7910/DVN/T1NQ4V.

Condon, D.M., 2022. $RetestReliability = f(Stability, Memory, Personality) + \epsilon$. Presented at symposium in honor of Sarah Dubrow.

Condon, D.M., 2023. Big five replicability, in: osf.o/da59z (Ed.), ARP.

Condon, D.M., Revelle, W., 2014. The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. Intelligence 43, 52–64. doi:10.1016/j.intell.2014.01.004.

Costa, P.T., McCrae, R.R., 1992. Four ways five factors are basic. Personality and Individual Differences 13, 653–665. doi:10.1016/0191-8869(92)90236-I.

Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16, 297–334. doi:10.1007/BF02310555.

Cronbach, L.J., Meehl, P.E., 1955. Construct validity in psychological tests. Psychological Bulletin 52, 281–302. doi:10.1037/h0040957.

Cudeck, R., MacCallum, R.C., 2007. Factor analysis at 100: historical developments and future directions. Lawrence Erlbaum Associates, Mahwah, N.J.

Cureton, E.E., 1950. Validity, reliability, and baloney. Educational and Psychological Measurement 10, 94–96. doi:10.1177/001316445001000107.

Cutler, A., Condon, D.M., 2023. Deep lexical hypothesis: Identifying personality structure in natural language. Journal of Personality and Social Psychology 125, 173–197. doi:https://doi.org/10.1037/pspp0000443.

Dawis, R.V., 1992. The individual differences tradition in counseling psychology. Journal of counseling psychology 39, 7–19. doi:10.1037/0022-0167.39.1.7.

Deary, I.J., 2001. Intelligence: A very short introduction. OUP Oxford.

Deary, I.J., 2009. Introduction to the special issue on cognitive epidemiology. Intelligence 37, 517–519. doi:10.1016/j.intell.2009.05.001.

Digman, J.M., 1990. Personality structure: Emergence of the five-factor model. Annual Review of Psychology 41, 417–440. doi:10.1146/annurev.ps.41.020190.002221.

Digman, J.M., 1997. Higher-order factors of the big five. Journal of Personality and Social Psychology 73, 1246–1256. doi:10.1037/0022-3514.73.6.1246.

Donnay, D., Morris, M., Schaubhut, N., Thompson, R., 2005. Strong interest inventory manual (rev. ed.). palo alto.

Donnay, D.A., 1997. E.K. Strong's legacy and beyond: 70 years of the Strong interest inventory. Career Development Quarterly 46, 2 – 22. doi:10.1002/j.2161-0045.1997.tb00688.x.

Donnay, D.A., Borgen, F.H., 1996. Validity, structure, and content of the 1994 strong interest inventory. Journal of Counseling Psychology 43, 275–291. doi:0022-0167/96.

Dworak, E.M., Revelle, W., Doebler, P., Condon, D.M., 2021. Using the International Cognitive Ability Resource as an open source tool to explore individual differences in cognitive ability. Personality and Individual Differences 169. doi:10.1016/j.paid.2020.109906.

Eagly, A.H., Revelle, W., 2022. Understanding the Magnitude of Psychological Differences Between Women and Men Requires Seeing the Forest and the Trees. Perspectives on Psychological Science 17, 1339–1358. doi:10.1177/17456916211046006.

Elleman, L.G., McDougald, S., Revelle, W., Condon, D., 2020. That takes the BISCUIT: a comparative study of predictive accuracy and parsimony of four statistical learning techniques in personality data, with data missingness conditions. European Journal of Psychological Assessment 36, 948–958. doi:/10.1027/1015-5759/a000590.

Embretson, S.E., 1996. The new rules of measurement. Psychological Assessment 8, 341–349. doi:10.1037/1040-3590.8.4.341.

Embretson, S.E., Hershberger, S.L., 1999. The new rules of measurement : what every psychologist and educator should know. L. Erlbaum Associates, Mahwah, N.J.

Eysenck, H.J., 1944. Types of personality: a factorial study of seven hundred neurotics. The British Journal of Psychiatry 90, 851–861. doi:10.1192/bjp.90.381.851.

Eysenck, H.J., 1952. The scientific study of personality. Routledge & K. Paul, London,.

Eysenck, H.J., 1953. Uses and abuses of psychology. Penguin Books, London, Baltimore,.

Eysenck, H.J., 1964. Sense and nonsense in psychology. Penguin Books, Baltimore.

Eysenck, H.J., 1965. Fact and fiction in psychology. Penguin Books, Baltimore,.

Eysenck, H.J., 1967. The biological basis of personality. Thomas, Springfield.

Eysenck, H.J., 1990. Biological dimensions of personality, in: Pervin, L.A. (Ed.), Handbook of personality: Theory and research.. Guilford Press, New York, NY, pp. 244–276.

Eysenck, H.J., Eysenck, M.W., 1985. Personality and Individual Differences: A natural science approach. Plenum, New York.

Eysenck, H.J., Eysenck, S.B.G., 1964. Eysenck Personality Inventory. Educational and Industrial Testing Service, San Diego, California.

Eysenck, H.J., Himmelweit, H.T., 1947. Dimensions of personality; a record of research carried out in collaboration with H.T. Himmelweit [and others]. Routledge & Kegan Paul, London.

Flynn, J.R., 1984. The mean IQ of Americans: Massive gains 1932 to 1978. Psychological Bulletin 95, 29–51. doi:/10.1037/0033-2909.95.1.29.

Flynn, J.R., 1987. Massive IQ gains in 14 nations: What IQ tests really measure. Psychological Bulletin 101, 171–191. doi:10.1037/0033-2909.101.2.171.

Forbes, M.K., Sunderland, M., Rapee, R.M., Batterham, P.J., Calear, A.L., Carragher, N., Ruggero, C., Zimmerman, M., Baillie, A.J., Lynch, S.J., Mewton, L., Slade, T., Krueger, R.F., 2021. A detailed hierarchical model of psychopathology: From individual symptoms up to the general factor of psychopathology. Clinical Psychological Science 9, 139–168. doi:10.1177/2167702620954799.

Galton, F., 1888. Co-relations and their measurement. Proceedings of the Royal Society. London Series 45, 135–145.

Garner, K.M., in press. The forgotten trade-off between internal consistency and validity. (abstract). Multivariate Behavioral Research .

Goldberg, L.R., 1972. Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. Multivariate Behavioral Research Monographs. No 72-2 7.

Goldberg, L.R., 1990. An alternative "description of personality": The big-five factor structure. Journal of Personality and Social Psychology 59, 1216–1229. doi:10.1037/0022-3514.59.6.1216.

Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Cloninger, C.R., Gough, H.G., 2006. The international personality item pool and the future of public-domain personality measures. Journal of Research in Personality 40, 84–96. doi:10.1016/j.jrp.2005.08.007.

Gottlieb, T., Furnham, A., Klewe, J.B., 2021. Personality in the light of identity, reputation and role taking: a review of socioanalytic theory. Psychology 12, 2020–2041. doi:10.4236/psych.2021.1212123.

Gough, H.G., 1957. Manual for the California Psychological Inventory.

Gough, H.G., 1960. The adjective check list as a personality assessment research technique. Psychological Reports 6, 107–122. doi:10.2466/pr0.1960.6.1.107.

Gough, H.G., 1965. Conceptual analysis of psychological test scores and other diagnostic variables. Journal of Abnormal Psychology 70, 294–302. doi:10.1037/h0022397.

Gray, J.A., 1981. A critique of Eysenck's theory of personality, in: Eysenck, H.J. (Ed.), A Model for Personality. Springer, Berlin, pp. 246–277.

Gray, J.A., 1987. Perspectives on anxiety and impulsivity: A commentary. Journal of Research in Personality 21, 493–509. doi:10.1016/0092-6566(87)90036-5.

Gruber, F.M., Distlberger, E., Scherndl, T., Ortner, T.M., Pletzer, B., 2020. Psychometric properties of the multifaceted gender-related attributes survey (GERAS). European Journal of Psychological Assessment 36, 612–623. doi:10.1027/1015-5759/a000528.

Guilford, J.P., 1940. Inventory of factors STDCR. Sheridan Supply Co, Beverly Hills, Calif.

Guilford, J.P., 1954. Psychometric Methods. 2nd ed., McGraw-Hill, New York.

Guilford, J.P., 1956. The structure of intellect. Psychological Bulletin 53, 267 – 293. doi:10.1037/h0040755.

Guilford, J.P., 1975. Factors and factors of personality. Psychological Bulletin 82, 802–814. doi:10.1037/h0077101.

Gulliksen, H., 1950. Theory of mental tests. John Wiley & Sons, Inc.

Guttman, L., 1945. A basis for analyzing test-retest reliability. Psychometrika 10, 255–282. doi:10.1007/BF02288892.

Hase, H.D., Goldberg, L.R., 1967. Comparative validity of different strategies of constructing personality inventory scales. Psychological Bulletin 67, 231–248. doi:10.1037/h0024421.

Hathaway, S., McKinley, J., 1943. Manual for administering and scoring the MMPI.

Henry, S., Thielmann, I., Booth, T., Mõttus, R., 2022. Test-retest reliability of the hexaco-100—and the value of multiple measurements for assessing reliability. PLOS ONE 17, 1–14. doi:10.1371/journal.pone.0262465.

Herrnstein, R.J., Murray, C., 2010. The Bell Curve: Intelligence and class structure in American life. Simon and Schuster.

Hogan, R., 1982. A socioanalytic theory of personality, in: Nebraska Symposium on Motivation. University of Nebraska Press, pp. 55–89.

Hogan, R., 2009. John Holland. URL: https://www.hoganassessments.com/blog/john-holland/.

Hogan, R., Blickle, G., 2018. Socioanalytic theory: Basic concepts, supporting evidences, and practical implications, in: Shackelford, V.Z.H..T.K. (Ed.), The SAGE handbook of personality and individual differences: The science of personality and individual differences. Sage Reference. volume 1, pp. 110–129. doi:10.4135/9781526451163.n5.

Hogan, R., Curphy, G.J., Hogan, J., 1994. What we know about leadership: Effectiveness and personality. American Psychologist 49, 493 – 504. doi:10.1037/0003-066X.49.6.493.

Hogan, R., Hogan, J., 1995. The Hogan personality inventory manual (2nd. ed.). Hogan Assessment Systems, Tulsa, OK.

Hogan, R., Hogan, J., Roberts, B.W., 1996. Personality measurement and employment decisions: Questions and answers. American Psychologist 51, 469 – 477. doi:10.1037/0003-066X.51.5.469.

Hogan, R., Sherman, R.A., 2020. Personality theory and the nature of human nature. Personality and Individual Differences 152, 109561. doi:10.1016/j.paid.2019.109561.

Holland, J.L., 1959. A theory of vocational choice. Journal of Counseling Psychology 6, 35 – 45. doi:10.1037/h0040767.

Holland, J.L., 1996. Exploring careers with a typology: What we have learned and some new directions. American Psychologist 51, 397 – 406. doi:10.1037/0003-066X.51.4.397.

Howell, R.D., Breivik, E., Wilcox, J.B., 2007. Reconsidering formative measurement. Psychological methods 12, 205–218. doi:10.1037/1082-989X.12.2.205.

Jensen, A.R., 1969. How much can we boost iq and scholastic achievement. Harvard educational review 39, 1–123. doi:10.17763/haer.39.1.l3u15956627424k7.

Jensen, A.R., 1998. The g factor: The science of mental ability. Westport, CT: Prager .

Jensen, A.R., Weng, L.J., 1994. What is a good g? Intelligence 18, 231–258. doi:10.1016/0160-2896(94)90029-9.

Johnson, W., Brett, C.E., Deary, I.J., 2010. The pivotal role of education in the association between ability and social class attainment: A look across three generations. Intelligence 38, 55 – 65. doi:10.1016/j.intell.2009.11.008.

Jonas, K.G., Markon, K.E., 2016. A descriptivist approach to trait conceptualization and inference. Psychological Review. 123, 90–96. doi:10.1037/a0039542.

Jöreskog, K.G., 1978. Structural analysis of covariance and correlation matrices. Psychometrika 43, 443–477. doi:10.1007/BF02293808.

Kovacs, K., Conway, A.R., 2019. A unified cognitive/differential approach to human intelligence: Implications for iq testing. Journal of Applied Research in Memory and Cognition 8, 255 – 272. doi:10.1016/j.jarmac.2019.05.003.

Kovacs, K., Conway, A.R.A., 2016. Process overlap theory: A unified account of the general factor of intelligence. Psychological Inquiry 27, 151 – 177. doi:10.1080/1047840X.2016.1153946.

Krueger, R.F., Markon, K.E., 2006a. Reinterpreting comorbidity: A model-based approach to understanding and classifying psychopathology. Annual Review of Clinical Psychology 2, 111–133. doi:10.1146/annurev.clinpsy.2.022305.095213.

Krueger, R.F., Markon, K.E., 2006b. Understanding psychopathology: Melding behavior genetics, personality, and quantitative psychology to develop an empirically based model. Current Directions in Psychological Science 15, 113–117. doi:10.1111/j.0963-7214.2006.0041.

Kuder, G., Richardson, M., 1937. The theory of the estimation of test reliability. Psychometrika 2, 151–160. doi:10.1007/BF02288391.

Lee, J.J., Wedow, R., Okbay, A., et al., 2018. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nature Genetics 50, 1112–1121. doi:10.1038/s41588-018-0147-3.

Loevinger, J., 1957. Objective tests as instruments of psychological theory. Psychological Reports Monograph Supplement 9 3, 635–694. doi:10.2466/pr0.1957.3.3.635.

Markon, K.E., Krueger, R.F., Watson, D., 2005. Delineating the structure of normal and abnormal personality: An integrative hierarchical approach. Journal of Personality and Social Psychology 88, 139–157. doi:10.1037/0022-3514.88.1.139.

Marschak, J., 1954. Probability in the social sciences, in: Lazarfeld, P. (Ed.), Mathematical thinking in the social sciences. Free Press, pp. 166–215.

McCrae, R.R., 2014. A more nuanced view of reliability: Specificity in the trait hierarchy. Personality and Social Psychology Review 19, 97–112. doi:10.1177/1088868314541857.

McCrae, R.R., Kurtz, J.E., Yamagata, S., Terracciano, A., 2011. Internal consistency, retest reliability, and their implications for personality scale validity. Personality and Social Psychology Review 15, 28–50. doi:10.1177/1088868310366253.

McDonald, R.P., 1999. Test theory: A unified treatment. L. Erlbaum Associates, Mahwah, N.J.

Möttus, R., Wood, D., Condon, D.M., Back, M.D., Baumert, A., Costantini, G., Epskamp, S., Greiff, S., Johnson, W., Lukaszewski, A., Murray, A., Revelle, W., Wright, A.G., Yarkoni, T., Ziegler, M., Zimmermann, J., 2020. Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the big few traits. European Journal of Personality 34, 1175–1201. doi:10.1002/per.2311.

Musek, J., 2007. A general factor of personality: Evidence for the big one in the five-factor model. Journal of Research in Personality 41, 1213–1233. doi:10.1016/j.jrp.2007.02.003.

Nunnally, J.C., 1978. Psychometric theory. 2nd ed., McGraw-Hill, New York,.

Park, S.K., Tao, Y., Meeker, J.D., Harlow, S.D., Mukherjee, B., 2014. Environmental risk score as a new tool to examine multi-pollutants in epidemiologic research: an example from the nhanes study using serum lipid levels. PLoS One 9, e98632. doi:10.1371/journal.pone.0098632.

Peabody, D., 1967. Trait inferences: Evaluative and descriptive aspects. Journal of Personality and Social Psychology 7, 1 – 18. doi:10.1037/h0025230.

Pearson, K., 1896. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. Philisopical Transactions of the Royal Society of London. Series A 187, 254–318. doi:10.1098/rsta.1896.0007.

Plato, n.d. Plato The Republic : the complete and unabridged Benjamin Jowett translation (1892). 3rd ed., Oxford Univeristy Press, Oxford.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Reise, S., 1999. Personality measurement issues viewed through the eyes of IRT, in: Embretson, S.E., Hershberger, S.L. (Eds.), The new rules of measurement: What every psychologist and educator should know. Lawrence Erlbaum Associates, Mahwah, N.J., pp. 219–241.

Revelle, W., 1983. Factors are fictions, and other comments on individuality theory. Journal of Personality 51, 707–714. doi:10.1111/1467-6494.ep7380795.

Revelle, W., 1989. Personality theory is alive and well and living in europe. Contemporary Psychology: APA Review of Books 34, 235–236. doi:10.1037/027760.

Revelle, W., 2023a. psych:Procedures for Psychological, Psychometric, and Personality Research. 2.3.9 ed. Northwestern University, Evanston. https://CRAN.r-project.org/package=psych. R package version 2.3.9.

Revelle, W., 2023b. psychTools Tools to Accompany the Psych Package for Psychological Research. Northwestern University, Evanston. psychTools. R package version 2.3.9.

Revelle, W., Condon, D., 2023. Using unidim rather than omega in estimating undimensionality. submitted .

Revelle, W., Dworak, E.M., Condon, D.M., 2020. Cognitive ability in everyday life: the utility of open source measures. Current Directions in Psychological Science 29, 358–363. doi:10.1177/0963721420922178.

Revelle, W., Dworak, E.M., Condon, D.M., 2021. Exploring the persome: The power of the item in understanding personality structure. Personality and Individual Differences 169. doi:10.1016/j.paid.2020.109905.

Revelle, W., Ellman, L.G., 2016. Factors are still fictions [peer commentary on "towards more rigorous personality trait–outcome research," by R. Mõttus]. European Journal of Personality 30, 324–325.

Revelle, W., Garner, K.M., 2023. Measurement: Reliability, construct validation, and scale construction, in: Harry T. Reis, T.W., Judd, C.M. (Eds.), Handbook of Research Methods in Social and Personality Psychology (in press).

Revelle, W., Wilt, J., 2013. The general factor of personality: A general critique. Journal of Research in Personality 47, 493–504. doi:10.1016/j.jrp.2013.04.012.

Revelle, W., Wilt, J., Condon, D., 2011. Individual Differences and Differential Psychology: A brief history and prospect, in: Chamorro-Premuzic, T., Furnham, A., von Stumm, S. (Eds.), Handbook of Individual Differences. Wiley-Blackwell, Oxford. chapter 1, pp. 3–38.

Roberts, B.W., Kuncel, N.R., Shiner, R., Caspi, A., Goldberg, L.R., 2007. The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. Perspectives on Psychological Science 2, 313–345. doi:10.1111/j.1745-6916.2007.000.

Royce, J.R., 1983. Personality integration: A synthesis of the parts and wholes of individuality theory. Journal of Personality 51, 683–706. doi:10.1111/j.1467-6494.1983.tb00874.x.

Schaeffer, N.C., 1988. An application of item response theory to the measurement of depression. Sociological Methodology 18, 271–307. URL: http://www.jstor.org/stable/271051.

Sijtsma, K., 2009a. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika 74, 107–120. doi:10.1007/s11336-008-9101-0.

Sijtsma, K., 2009b. Reliability beyond theory and into practice. Psychometrika 74, 169–173. doi:10.1007/s11336-008-9103-y.

Slaney, K., 2017. Historical Precursors and Early Testing Theory. Palgrave Macmillan UK, London. doi:10.1057/978-1-137-38523-9_2.

Spearman, C., 1904a. "General Intelligence," objectively determined and measured. American Journal of Psychology 15, 201–292. doi:10.2307/1412107.

Spearman, C., 1904b. The proof and measurement of association between two things. The American Journal of Psychology 15, 72–101. doi:10.2307/1412159.

Steinberg, L., Thissen, D., 2006. Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. Psychological Methods 11, 402–415. doi:10.1037/1082-989X.11.4.402.

Stewart, R.D., Mõttus, R., Seeboth, A., Soto, C.J., Johnson, W., 2022. The finer details? the predictability of life outcomes from big five domains, facets, and nuances. Journal of Personality 90, 167–182. doi:10.1111/jopy.12660.

Strong Jr., E.K., 1927. Vocational interest test. Educational Record 8, 107–121.

Strong Jr., E.K., 1947. Vocational interests of men and women. Stanford University Press.

Su, R., Tay, L., Liao, H.Y., Zhang, Q., Rounds, J., 2019. Toward a dimensional model of vocational interests. Journal of Applied Psychology 104, 690. doi:10.1037/apl0000373.

Thomson, G.H., 1916. A hierarchy without a general factor. British Journal of Psychology 8, 271–281.

Thomson, G.H., 1935. The definition and measurement of "g" (general intelligence). Journal of Educational Psychology 26, 241 – 262. doi:10.1037/h0059873.

Thurstone, L.L., 1934. The vectors of mind. Psychological Review 41, 1. doi:10.1037/h0075959.

Thurstone, L.L., 1935. The vectors of mind: multiple-factor analysis for the isolation of primary traits. Univ. of Chicago Press, Chicago.

Vassos, E., Sham, P., Kempton, M., Trotta, A., Stilo, S.A., Gayer-Anderson, C., Di Forti, M., Lewis, C.M., Murray, R.M., Morgan, C., 2020. The Maudsley environmental risk score for psychosis. Psychological medicine 50, 2213–2220.

Watts, A.L., Greene, A.L., Bonfifay, W., Fried, E.I., 2023. A critical evaluation of the p-factor literature. PsyArXiv 7yrnp/. doi:10.31234/osf.io/7yrnp.

Webb, E., 1915. Character and intelligence: An attempt at an exact study of character. The British Journal of Psychology, Monograph Supplements I.

Widaman, K.F., Revelle, W., 2023a. Thinking thrice about sum scores, and then some more about measurement and analysis. Behavior Research Methods 55, 788–806. doi:10.3758/s13428-022-01849-w.

Widaman, K.F., Revelle, W., 2023b. Thinking about sum scores yet again, maybe the last time, we don't know, oh no . . .: A comment on McNeish (2023). Educational and Psychological Measurement 0, 00131644231205310. doi:10.1177/00131644231205310.

Wiley, D.E., 1973. The identification problem for structural equation models with unmeasured variables, in: Goldberger, A.S., Duncan, O.D. (Eds.), Structural equation models in the social sciences. Seminar Press, New York, pp. 69–83.

Yarkoni, T., Westfall, J., 2017. Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science 12, 1100–1122. doi:10.1177/1745691617693393.

Zola, A., Condon, D.M., Revelle, W., 2021. The Convergence of Self and Informant Reports in a Large Online Sample. Collabra: Psychology 7. doi:10.1525/collabra.25983.

8. Appendix: R code for analyses

R code

```
#first make the appropriate packages available
library(psych)
library(psychTools)

#select big 5 itemnames from the keys list
select <- selectFromKeys(spi.keys[1:5])
#factor the big 5 items
f5 <- fa(spi[select],5) #just factor the 70 Big Few iems
b5 <- scoreItems(spi.keys[1:5],spi) #find the raw meanscores
cor2(b5$scores,f5$scores)  #show they are basically the same
factor.congruence(f5,b5$item.cor)  #another way to show this


set.seed(47)  #to make reproducible results

sc <- scoreItems(spi.keys,spi)
spi.scales <- cbind(spi[,1:10],sc$scores)


n.obs <- NROW(spi)
ss <- sample(n.obs, n.obs/2,replace=FALSE)
derivation <- spi.scales[ss,]   #chose a random 50%
#linear regression
mod.5 <- lmCor(y=1:10,x = 11:15,data =spi.scales[ss,],
       plot=FALSE)
summary(mod.5)
#now do it for the little 27
mod.27 <-  lmCor(y=1:10,x = 16:42,data =spi.scales[ss,],
          plot=FALSE)
#cross validate
cv <- crossValidation(mod.5,data=spi.scales[-ss,])
cv.27 <- crossValidation(mod.27, data =spi.scales[-ss,])

bs <- bestScales(x = spi[ss, 11:145],
   criteria = spi[ss, 1:10], max.item = 60,
    n.item = NULL, wtd.n = 30, folds = 10,
    dictionary = spi.dictionary,cut=.1)

 bs.cv <- crossValidation(bs,data=spi[-ss,])
# bs.cv.w <-  crossValidation(bs,data=spi[-ss,],
#     options="optimal.weights")

  #combine all of these results into one data.frame
spi.reg <-  data.frame(deriv5=mod.5$R,cross5=cv$crossV[,1],
 deriv27=mod.27$R,cross27=cv.27$crossV[,1],
   best=bs$summary[,5],
 best.cv = bs.cv$crossV[,1])

 ord <- dfOrder(spi.reg,2)   #order it for a nice graphic

matPlot(ord[-c(8,9) ,c(4,6,2)],legend=1,col=c(3,2,1),
    lty=c(3,2,1),
    ylab="Cross Validated R",
     main="Cross validated correlations for
      three methods of choosing scales")

par(mfrow=c(1,2))   #two panel graph
manhattan(spi[ss,] ,cs(health,exer),spi.keys)

par(mfrow=c(1,1))  #reset to one panel


#g factors

om.ab  <- omega(ability,4)
om.af <- omega(USAF)
par(mfrow=c(1,2))   #two panel graph
 par("mar"= c( 0,0,0,0))  #set margins to be really wide
omega.diagram(om.ab,sl=FALSE,
   main="A) Higher order factors of ICAR",
    e.size=.05, rsize=.25)
 omega.diagram(om.af, sl=FALSE,
     main="B) Higher order factors of size",
     e.size =.05,rsize=.25)
  par(mfrow=c(1,1))  #back to one panels
```