Revisiting old ideas for scale construction

William Revelle and Kayla M. Garner

Department of Psychology
Northwestern University
Evanston, Illinois USA

European Conference on Personality, Berlin, Germany

July, 2024

Slides available at https://personality-project.org/sapa

Abstract

In the mid 1950s, scale construction for academic personality research turned from predicting behaviors to measuring constructs. Emphasizing construct validity has led the field from being practically useful to being theoretically "pure". This was a mistake. We will address the advantages of using broad scales with low to moderate internal consistency (fishing nets) as contrasted to narrow, highly internally consistent (spears) scales. We will suggest that one catches more fish with nets than with spears. One builds up nets by using lower level items and nuances rather than more internally consistent high level factor score estimates. Predicting multivariate data requires multivariate models rather than factorially pure measures.

Examples in predicting gender, health, and exercise will be taken from open source material in the `psychTools` package with analysis using the `psych` package.

NORTHWESTERN UNIVERSITY

Outline

Two broad approaches in personality assessment and theory

1. For the past 70 years, personality assessment has been split between those who emphasize theoretical constructs thought to explain behavior and those who "merely" want to predict it. See (Möttus et al., 2020; Revelle, 2024c; Yarkoni and Westfall, 2017).
   - The psychological construct approach is most associated with the formative work of Jane Loevinger (1957) and Lee Cronbach and Paul Meehl (Cronbach and Meehl, 1955).
   - The straight predictive approach is best represented today by Robert Hogan and the success of the Hogan Personality Inventory.
2. Bob Hogan (2024) emphasizes that the successful tests for predicting real world outcomes such as educational attainment, occupational status or income are based upon empirical scoring of items that work. Perhaps the two most well known examples of this technique include the Strong Vocational Interest (Strong Jr., 1927) and the MMPI (Hathaway and McKinley, 1943). (Hogan, 2024).

Construct versus predictive validity

1. Constructs as explanations are much more fun to talk about (e.g., Extraversion and performance under stress, Revelle et al., 1976, 1980, 1987; Revelle and Anderson, 1992) and lead to an emphasis upon sharply defined, unidimensional tests. These tests are then correlated with other tests to better define the nomological network of a domain.
   - Psychometric techniques such as factor analysis and assessment of reliability are seen as important skills to master.
   - Psychometric tools including R packages (e.g, psych, Revelle, 2024a) can be developed to help do the analysis and tutorials on how to construct scales and assess reliability are well cited
2. Tests as predictive devices requires writing (choosing) good items, forming predictive scales, and then cross validating them. This is not as much fun, but results in higher predictive validity (but less parsimony) than construct pure measures (Revelle et al., 2021; Stewart et al., 2022).

These ideas are not new, merely forgotten

1. Gulliksen (1950) suggested validity varies independently of internal consistency.
2. Humphreys (1994) examined the phenotypic trait of intelligence and equated it to the total breadth of the cognitive repertoire. He argued that short and homogenous measures can not measure the breadth of intelligence.
3. Nandakumar (1991) citing many papers by Humphreys, forcefully argues that, from the validity viewpoint, tests should be deliberately constructed to include numerous minor factors.
4. Tellegen and Waller (2008) explained how to measure personality traits, one should not focus on maximizing internal consistency but rather focus on breadth.
5. (Condon et al., 2020; Stewart et al., 2022; Möttus et al., 2020) show how nuances (items) predict better than higher order scales.
6. Yarkoni and Westfall (2017) discuss the power of machine learning for prediction.

Recent work

1. We have previously encouraged researchers to resist the seductive beauty of latent variables which we equate to believing in the tooth fairy (Revelle, 2024c).
2. We have also shown how internal consistency trades off with predictive validity such that less internally consistent tests may actually be more valid (Eagly and Revelle, 2022; Revelle and Garner, 2024; Garner, 2024).
3. We have also shown the power of choosing items using classical scale construction techniques (now called "machine learning using k-fold cross validation") to predict real world criteria (Elleman et al., 2020; Revelle et al., 2021).
4. The central theme of these papers is that predictive validity should be a major goal of personality researchers and that we should rediscover some of the techniques that have long been known, but unfortunately, long forgotten.

But what are these "forgotten" techniques?

1. Empirical scale construction (choosing items that work) was the hallmark of the successful scales of the 1930s-1950s.
   - Occupational scales on the Strong Vocational Interest (Strong Jr., 1927) were formed of items endorsed by people in specific occupations that were not as strongly endorsed by people in general.
   - The clinical scales of the Minnesota Multiphasic Personality Inventory Behavioral and emotional items were formed from those items that discriminated a crtierion group from "normal" controls. (But see Helmes and Reddon, 1993, for a thoughtful critique, which criticize the MMPI for lack of theory as well as lack of cross validation of the original scale construction.)
2. Similar empirical scale construction (with an overlay of socioanalytic theory) (Hogan, 1982, 2024) has driven the construction of the HPI (Hogan and Hogan, 1995).
3. Hase and Goldberg (1967); Goldberg (1972) compared the utility of empirical and factor based techniques and found systematic advantages and disadvantages to both.

Validity and reliability: a short digression

1. Although we know from Spearman that we can correct for reliability to find the "True" relationship between two variables, this does not help us in the real world.
2. Reliability is incorrectly associated with internal consistency which leads to such derivations as coefficients KR20 (Kuder and Richardson, 1937), λ_3 (Guttman, 1945) or α (Cronbach, 1951).
3. Expressed terms of inter-item correlations, α is just $\frac{k\bar{r}}{1+(k-1)\bar{r}}$ and increases with test length (k) and the average interitem correlation ($\bar{r}$).
4. However, *validity* of a k item test (r_{y_k}) or the correlation with an external criterion, Y, also increases with test length, and the average item validity ($\bar{r_y}$) but decreases as the inter-item correlation increases $r_{y_k} = \frac{k\bar{r}_y}{\sigma_x} = \frac{k\bar{r}_y}{\sqrt{k+k*(k-1)\bar{r}}}$.

Reliability and Validity

1. Lets unpack these two equations.
   Internal consistency varies by number of items and average correlation (redundancy).

$$\lambda_3 = \alpha = \frac{k\bar{r}}{1 + (k-1)\bar{r}} \tag{1}$$

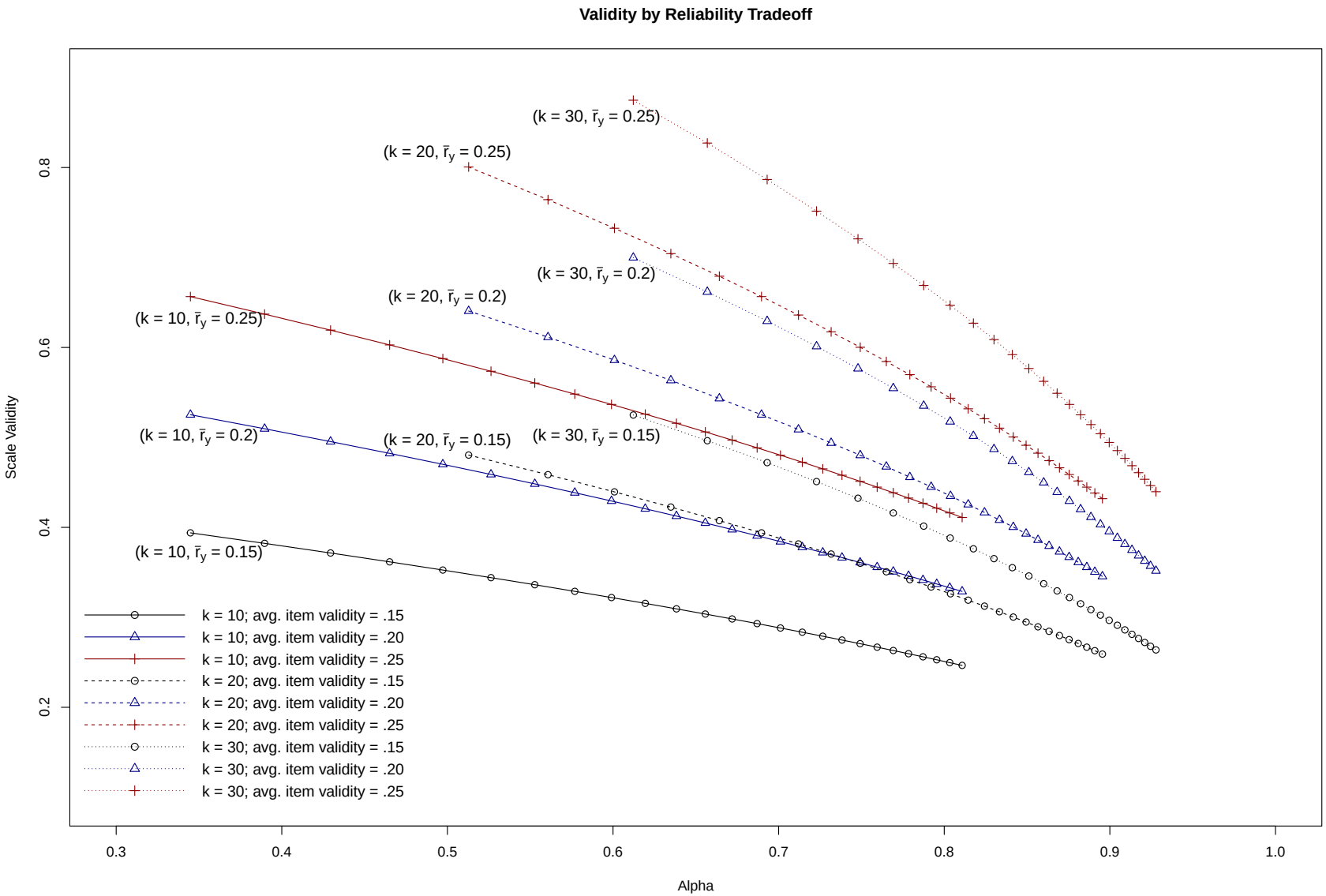
2. But validity varies by number of items, average within test correlation and average item validity

$$r_{y_k} = \frac{k\bar{r}_y}{\sigma_x} = \frac{k\bar{r}_y}{\sqrt{k + k * (k-1)\bar{r}}}. \tag{2}$$

Showing the reliability by validity tradeoff

1. Consider 9 scales formed from
2. 10, 20 or 30 items
3. Average validities of .15, .20, .25
4. Plot scale validity by scale α for $.3 < \alpha < .9$
5. Important to remember that $\alpha \neq$ reliability.

The trade off between test consistency and test validity



Several Examples

1. To show the power of items, one must give many items to many subjects. Thus:
2. We use data (952 items, N $\approx 255,000$)from the SAPA Project which uses a Massively Missing Completely at Random (MMCAR) approach.
3. Also use a subset of these data (the `spi`) included in the *psychTools* package.
4. We have previously reported similar results (Revelle et al., 2021) but now report some new analyses.
5. The basic theme of all of these results is that short, empirically chosen scales with low internal consistency (fishing nets) do a better job of prediction than do highly internally consistent scales (spears).

NORTHWESTERN UNIVERSITY

But what is SAPA?

1. SAPA (Synthetic Aperture Personality Assessment) presents random samples of 100-200 items sampled from 6600 items to volunteer participants interested in their personality (Condon, 2018; Revelle et al., 2010, 2017, 2021; Zola et al., 2017)
2. Name comes by analogy to Synthetic Aperture Radio Astronomy which combine many small radio telescopes to form image from a much larger telescope.
3. Although David Condon has released multiple tranches of 100-200,000 cases from the $> 2 * 10^6$ we have collected, here we just examine an earlier release of 255,000 case on 952 variables.
4. This set includes 19 demographic variables. 696 items from the International Personality Item Pool Goldberg (1999) and 60 items from the International Cognitive Ability Resource Condon and Revelle (2014); Condon et al. (2014).

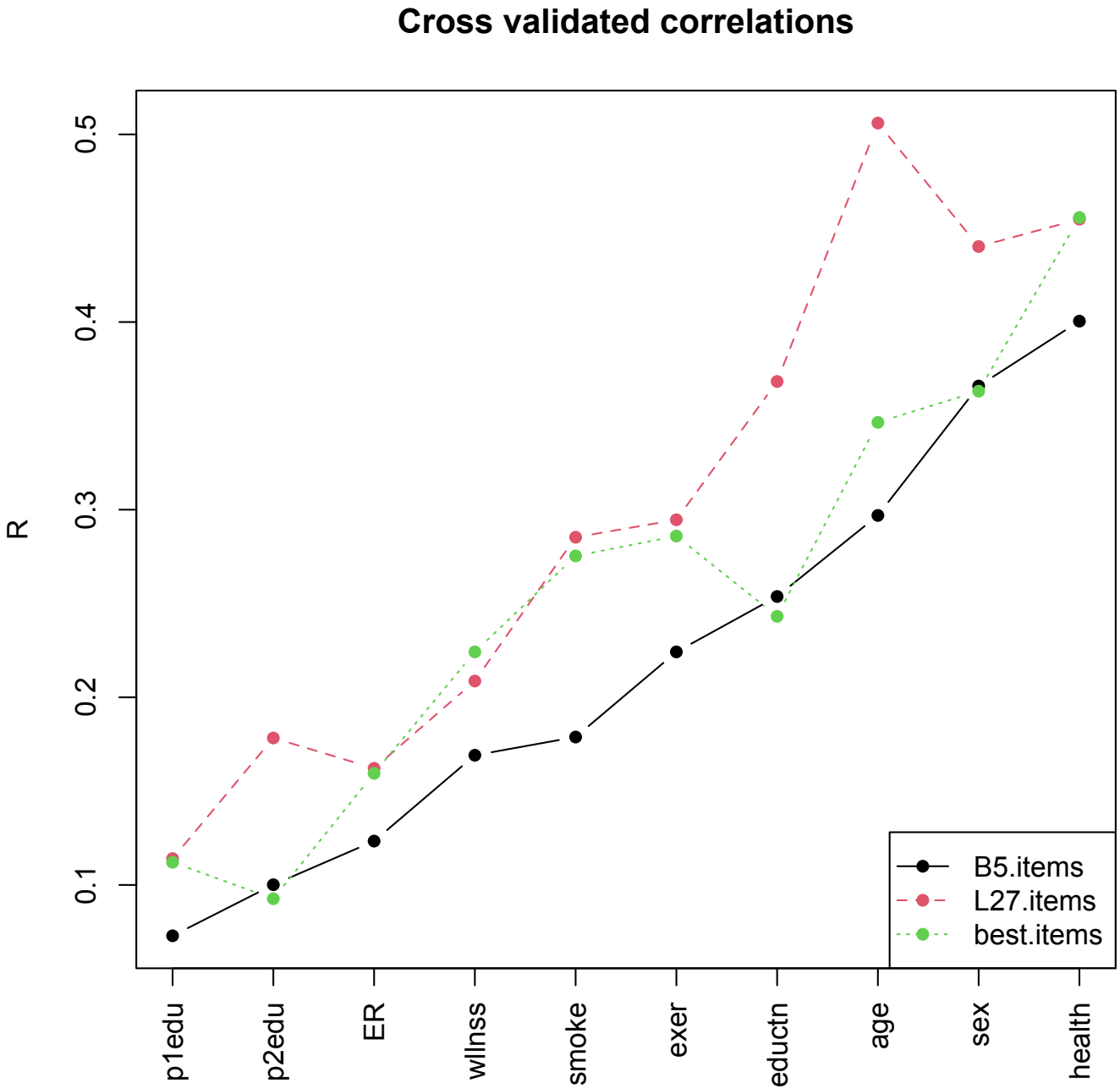
NORTHWESTERN UNIVERSITY

Design and Analysis

1. For ease of replicability and for demonstration purposes we use 135 items from the SAPA Personality Inventory (spi) Condon (2018).
2. We report 4000 subjects who took the spi as part of the SAPA project. These data are included as an example data set in the *psychTools* package (Revelle, 2024b) for R (R Core Team, 2024).
3. We randomly sampled 1/2 (2,000) subjects and then cross validated any analyses on the other 2,000 subjects.
4. the `spi` data set includes 10 criteria of interest.
5. We found 5 "Big Few" scales scores for 5 scales of 14 items each, 27 lower level/interstitial scales of 5 items each and then used a simple function `bestScales` to empirically find the best 10 items predicting each.
6. for the big Few and little 27, we used multiple correlation (using `lmCor`) to predict the criteria.

Analysis

1. Multiple correlations of each of 10 criteria predicted by the Big 5.
2. Multiple correlations of each of 10 criteria predicted by the Little 27.
3. Short scales (up to 10 items) formed from the items that best predict the criteria using the `bestScales` function.
4. All results derived on sample 1, cross validated on sample 2.


Cross validated correlations
R
0.1
0.2
0.3
0.4
0.5
p1edu
p2edu
ER
wllnss
smoke
exer
eductn
age
sex
health
B5.items
L27.items
best.items

Best items results are very interpretable: Smoking

Table: Best Items predicting smoking $\omega_h = .29, \alpha = .76, r = .28$

Variable	smoke	Item	B5	L27
q_1461	-0.22	Never spend more than I can afford.		SelfControl
q_1867	-0.18	Try to follow the rules.	Consc	Authoritarianism
q_1609	0.18	Rebel against authority.	Open	Authoritarianism
q_598	0.16	Do crazy things.		SensationSeeking
q_1624	-0.15	Respect authority.		Authoritarianism
q_1173	0.14	Jump into things without thinking.		Impulsivity
q_736	-0.14	Easily resist temptations.		SelfControl
q_1590	-0.13	Rarely overindulge.		SelfControl
q_1462	-0.13	Never splurge.		SelfControl
q_56	-0.13	Am able to control my cravings.		SelfControl

$$R_{b5} = .18 \quad R_{L27} = .28 \quad R_{best10} = .28$$

$$\text{N } R_{b5} = 70 \quad N_{L27}\ 135 \quad N_{best10}\ 10$$

NORTHWESTERN UNIVERSITY

Exercise

Table: Best Items predicting exercise $\omega_h = .48, \alpha = .79, r = .29$

Variable	exer	Item	B5	L27
q_1024	-0.27	Hang around doing nothing.		EasyGoingness
q_1052	-0.24	Have a slow pace to my life.		EasyGoingness
q_1452	-0.22	Neglect my duties.	Consc	Industry
q_1444	-0.21	Need a push to get started.	Consc	Industry
q_1979	0.20	Work hard.	Consc	Industry
q_1371	0.20	Love life.		WellBeing
q_1505	-0.20	Panic easily.	Neuro	Anxiety
q_1662	0.19	Seek adventure.		SensationSeeking
q_808	-0.19	Fear for the worst.	Neuro	Anxiety
q_578	-0.19	Dislike myself.	Neuro	WellBeing

$R_{b5} = .22 \quad R_{L27} = .29 \quad R_{best10} = .29$

N $R_{b5} = 70 \quad N_{L27}$ 135 $\quad N_{best10}$ 10

Health

Table: Best Items predicting health $\omega_h = .57, \alpha = .86, r = .44$

A table from the psych package in R

Variable	helth	Item	B5	L27
q_820	0.34	Feel comfortable with myself.		WellBeing
q_2765	0.34	Am happy with my life.		WellBeing
q_578	-0.33	Dislike myself.	Neuro	WellBeing
q_811	-0.31	Feel a sense of worthlessness or hopelessness.	Neuro	WellBeing
q_1371	0.31	Love life.		WellBeing
q_56	0.29	Am able to control my cravings.		SelfControl
q_1505	-0.27	Panic easily.	Neuro	Anxiety
q_1452	-0.26	Neglect my duties.	Consc	Industry
q_808	-0.25	Fear for the worst.	Neuro	Anxiety
q_1024	-0.25	Hang around doing nothing.		EasyGoingness

$R_{b5} = .40 \quad R_{L27} = .45 \quad R_{best10} = .45$

N $R_{b5} = 70 \quad N_{L27}$ 135 $\quad N_{best10}$ 10

Conclusions

1. We have known for years but seem to have forgotten that:
2. Validity and Internal consistency (redundancy) trade off with other. Increasing internal consistency/redundancy reduces external validity.
3. Highly internally consistent scales (spears) might make more theoretical sense but
4. Diffuse scales containing unrelated items that all relate to the criteria (nets) have higher predictive validity.
5. Cross validated correlations are higher with diffuse items that can be examined for interpretation than highly intenally consistent scales.
6. One catches more fish with nets than spears.

Condon, D. M. (2018). The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model. *PsyArXiv*.

Condon, D. M., Doebler, P., Holling, H., Gühne, D., Rust, J., Stillwell, D., Sun, L., Chan, F., Loe, A., and Revelle, W. (2014). International Cognitive Ability Resource. *https://icar-project.com*.

Condon, D. M. and Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence*, 43:52–64.

Condon, D. M., Wood, D., Mõttus, R., Booth, T., Costantini, G., Greiff, S., Johnson, W., Lukaszewski, A., Murray, A., Revelle, W., Wright, A. G. C., Ziegler, M., and Zimmermann, J. (2020). Bottom Up Construction of a Personality Taxonomy. *European Journal of Psychological Assessment*, 36(6):923–934.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334.

NORTHWESTERN UNIVERSITY

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302.

Eagly, A. H. and 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(5):1339–1358.

Elleman, L. G., McDougald, S., Revelle, W., and 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(6):948–958.

Garner, K. M. (2024). 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. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In Mervielde, I., Deary, I., De Fruyt, F., and Ostendorf, F., editors, *Personality psychology in Europe*, volume 7, pages 7–28. Tilburg University Press, Tilburg, The Netherlands.

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

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10(4):255–282.

Hase, H. D. and Goldberg, L. R. (1967). Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin*, 67(4):231–248.

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

Helmes, E. and Reddon, J. R. (1993). A perspective on developments in assessing psychopathology: A critical review of the mmpi and mmpi-2. *Psychological bulletin.*, 113(3):453–471.

Hogan, R. (1982). A socioanalytic theory of personality. In *Nebraska Symposium on Motivation*, pages 55–89. University of Nebraska Press.

Hogan, R. (2024). Personality. In *Personality, Leadership, and Organizational Effectiveness (in prep)*, chapter 1.

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

Humphreys, L. G. (1994). Intelligence from the standpoint of a (pragmatic) behaviorist. *Psychological Inquiry*, 5(3):179–192.

Kuder, G. and Richardson, M. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2(3):151–160.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports Monograph Supplement 9*, 3:635–694.

NORTHWESTERN UNIVERSITY

Möttus, R., Wood, D., Condon, D. M., Back, M., Baumert, A., Costani, G., Epskamp, S., Greiff, S., Johnson, W., Lukaszeksi, A., Murray, A., Revelle, W., Wright, A. G., Yarkoni, T., Ziegler, M., and Zimmerman, 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(6).

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement*, 28(2):99–117.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Revelle, W. (2024a). *psych:Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston,

NORTHWESTERN UNIVERSITY

https://CRAN.r-project.org/package=psych, 2.4.7 edition. R package version 2.4.7.

Revelle, W. (2024b). *psychTools Tools to Accompany the Psych Package for Psychological Research*. Northwestern University, Evanston, psychTools. R package version 2.4.3.

Revelle, W. (2024c). The seductive beauty of latent variable models: Or why i don't believe in the easter bunny. *Personality and Individual Differences*, 221:112552.

Revelle, W., Amaral, P., and Turriff, S. (1976). Introversion-extraversion, time stress, and caffeine: effect on verbal performance. *Science*, 192:149–150.

Revelle, W. and Anderson, K. J. (1992). Models for the testing of theory. In Gale, A. and Eysenck, M., editors, *Handbook of Individual Differences: Biological Perspectives*, chapter 4, pages 81–113. John Wiley and Sons, Chichester, England.

Revelle, W., Anderson, K. J., and Humphreys, M. S. (1987). Empirical tests and theoretical extensions of arousal-based

theories of personality. In Strelau, J. and Eysenck, H., editors, *Personality Dimensions and Arousal*, pages 17–36. Plenum, New York.

Revelle, W., Condon, D. M., Wilt, J., French, J. A., Brown, A., and Elleman, L. G. (2017). Web and phone based data collection using planned missing designs. In Fielding, N. G., Lee, R. M., and Blank, G., editors, *Sage Handbook of Online Research Methods*, chapter 37, pages 578–595. Sage Publications, Inc., 2nd edition.

Revelle, W., Dworak, E. M., and Condon, D. M. (2021). Exploring the persome: The power of the item in understanding personality structure. *Personality and Individual Differences*, 169.

Revelle, W. and Garner, K. M. (2024). Alternative metaphors for validity: Spear fishing versus fishing nets. In *World Conference on Personality*.

Revelle, W., Humphreys, M. S., Simon, L., and Gilliland, K. (1980). Interactive effect of personality, time of day, and

caffeine: A test of the arousal model. *Journal of Experimental Psychology General*, 109(1):1–31.

Revelle, W., Wilt, J., and Rosenthal, A. (2010). Individual differences in cognition: New methods for examining the personality-cognition link. In Gruszka, A., Matthews, G., and Szymura, B., editors, *Handbook of Individual Differences in Cognition: Attention, Memory and Executive Control*, chapter 2, pages 27–49. Springer, New York, N.Y.

Stewart, R. D., Mõttus, R., Seeboth, A., Soto, C. J., and Johnson, W. (2022). The finer details? the predictability of life outcomes from big five domains, facets, and nuances. *Journal of Personality*, 90(2):167–182.

Strong Jr., E. K. (1927). Vocational interest test. *Educational Record*, 8(2):107–121.

Tellegen, A. and Waller, N. G. (2008). Exploring personality through test construction: Development of the multidimensional

personality questionnaire. *The Sage handbook of personality theory and assessment*, 2:261–292.

Yarkoni, T. and Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122.

Zola, A., Condon, D. M., and Revelle, W. (2017). The convergence of observer ratings and self reports from SAPA. In *The bienniel meeting of the Association of Research in Personality*, Sacromento.