# Measurement: Reliability, construct validation, and scale construction

William Revelle and Kayla M. Garner
Northwestern University

**Abstract**

Adequate measurement of psychological phenomena is a fundamental aspect of theory construction and validation. Forming composites scales from individual items has a long and honored tradition, although for predictive purposes, the power of using individual items should be considered.

To construct a scale or set of scales to measure one or more constructs seems straightforward: write items to measure what you are interested in and then administer these items to samples of the population of interest. Unfortunately, reality is not quite so simple. Although it is seductively easy to write items and form them into scales, we want to be able to evaluate how well we are measuring our constructs. These are questions of the reliability and validity of the scales[1]. Throughout this chapter, we will refer to examples taken from open source data sets most of which are available in the *psychTools* package (Revelle, 2022b) in the R statistical system (R Core Team, 2022). The analyses are done using the *psych* package[2]. (Revelle, 2022a) in R.

[1] We emphasize the plural, scales, rather than the singular, scale, because part of the process of validation will be showing what our scales are not. In addition, most research programs are multivariate and tend to emphasize multiple constructs.

[2] When ever referring to data sets or functions, we will use `boldfaced` fixed pitch font.

contact: William Revelle (0000-0003-4880-9610) revelle@northwestern.edu
Kayla M. Garner (0000-0003-0245-5415) kaylagarner@u.northwestern.edu:
Draft of chapter to appear in Harry T. Reis, Tessa Wesrt and Charles M. Judd (eds) *Handbook of Research Methods in Social and Personality Psychology* Slightly revised draft version of September 3

Box 1: Steps towards constructing scales

1. Define the question and the goals of the assessment.
   (a) Review the relevant literature to define the domain.
   (b) Write items (or choose already written items) that tap this domain.
2. Administer the items to a representative sample.
3. Convert the data into a machine readable form .
   (a) Find basic descriptive statistics.
   (b) Examine for bad data.
4. Form a matrix of correlations or covariances from the data. Allow for missing data if missing is at random.
5. Apply a dimensional reduction technique (e.g. factor analysis).
   (a) Consider multiple solutions with fewer or more factors.
   (b) Consider higher order (e.g., bifactor) models
6. Select those items
   (a) With highest loadings on the factors (if trying to maximize internal consistency) (unsupervised learning)
   (b) With highest predictive validity for a criterion (supervised learning
7. Score the resulting scales and examine the item statistics. Drop those items that are missfitting.
8. Validate the scales on an independent sample.
9. Consider why the scales do not work as expected. Think about ways of improving the items.
10. Redefine the question and go back to step 1

A scale is typically formed by combining two or more items thought to measure the same construct into a single composite scale. This is done because it is thought to increase the reliability and generality of the resulting scale. This involves several steps: *a*) Choosing between prediction and explanation; *b*) Specifying the construct(s) to measure; *c*) choosing items thought to measure these constructs; *d*) administering the items; *e*) examining the

properties of composite of items (scales), $f$) validating the resulting scales.

## Prediction versus Explantation: Theory First (and last)

Some of the most used tests and their associated scales were developed to predict real world outcomes. Thus, the Strong Vocational Interest test (Strong Jr., 1927) and the Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1943) were used to predict vocational choices and psychopathology not from some factor based set of latent constructs, but merely by choosing items that could separate known groups. Similarly, the US military used neuropsychiatric screening measures to aid selection of recruits. The utility of these instruments was a straightforward evaluation of their success in discriminating known groups. The emphasis was upon *predictive* validity. Unfortunately, every criteria required a separate validation study. Although a source for many a dissertation, this did not seem to produce any cumulative wisdom.

A drastic change from an emphasis upon predictive validity came with the emergence of *construct validity* as a way of furthering psychology theory (Cronbach & Meehl, 1955; Loevinger, 1957). However, this has not met with universal agreement (see e.g., Borsboom et al., 2004; Embretson, 2007; Sartori & Pasini, 2007, for refreshing restatements of validity). We will discuss the advantages and disadvantages of both approaches: those that emphasize predictive validity and those that emphasize construct validity. Thus, the first question the investigator needs to address is what is the purpose of the scale to be developed.

Before delving into scale construction, it is important to consider the elements of scales: items.

## Writing or choosing items to present

Items form the basic unit of scales. They need to be written clearly and unambiguously and have a number of response alternatives.[3] Typical items may be single words, phrases, sentences, or even paragraphs. Usually a set of items will have some instructions and context for the items. Excellent suggestions for writing good items are discussed by Clark & Watson (1995, 2019). It is very important to have subject matter experts (that is, people familiar with the constructs you are trying measure) as well as the targeted end users to screen the items for relevance, understandability, and subtle forms of bias (Condon et al., 2021) Participants, particularly college freshman, will respond to items even though they will admit independently not to not actually understand the meaning of the words (Graziano et al., 1998).

[3] Even items in a check list for endorsement of symptoms has an alternative, don't endorse it.

*Example Items*

Among the many possibilities, items can be single words to measure affect, sentence fragments asking about typical thoughts and feelings, preferences for activities, or assessments of cognitive ability. (See box 2).

Box 2: Examples of types of items.

A Items can be measures of emotional states such as energetic arousal or tension (Thayer, 1989) e.g., I typically feel ...

1 Active
2 Afraid
3 Alert
4 ...

(Adapted from the Motivational State Questionnaire (Revelle & Anderson, 1998). See the `msq` and `msqR` data sets in *psychTools*, N=3032.

B More conventional personality trait items such as Extraversion and Neuroticism. e.g., How much do you agree with the following statements:

1 I like going out a lot?
2 I want to be left alone ?
3 I laugh a lot?
4 ...

(Adapted from the `spi` (Condon, 2018). See `spi` dataset, N=4,000.

C Can ask about normal activities: e.g. How often have you:

1 Shot a gun
2 Meditated
3 Took antacids
4 ...

(Adapted from the Oregon Avocational Interest Scales (ORAIS) (Goldberg, 2010)

D Or can be measures of cognitive ability :

1 What number is one fifth of one fourth of one ninth of 900?
2 In the following alphanumeric series, what letter comes next? I J L O S?
3 If the day after tomorrow is two days before Thursday, then what day is today?
4 ...

(Adapted from the International Cognitive Ability Resource (ICAR) (Condon & Revelle, 2014). See `ability` data set for the "ICAR 16", N = 1,525

*Open source items*

A very useful compendium of 3,320 "Personality" items was released to the public as the International Personality Item Pool (Goldberg, 1999) and is available for download from `ipip.ori.org`. A larger set of sources for more than 12,000 items and the actual items for about 5,000 public domain items is available as the Database of Individual Differences (Condon, 2019). Also available from Condon (2019) are the 17,950 adjectives in the original data set from Allport & Odbert (1936) that formed the core data set of what has become known as the "lexical hypothesis".

Indeed, when examining 2,084 items included in the IPIP which represented 403 separate scales, Condon (2018) reported that 696 items recovered 100% of 168 of these scales and between 60% and 85% of 235 additional scales. The data for 126,884 respondents to these 696 items are openly available at DataVerse (Condon & Revelle, 2015; Condon et al., 2017a,b) and have been discussed in multiple publications (e.g., Revelle et al., 2021). Of these 696 items, 135 items show a very clean factor structure of 27 short 5 item scales, and 5 higher order scales similar to the Big 5 of other inventories. These 135 items form the SAPA Personality Inventory (SPI) (Condon, 2018). Data for 4,000 participants from the SPI are included in the `spi` data set in *psychTools* as are the actual items presented. We will use these items and scales as examples in the following pages. The 696 were chosen from a large number of other inventories (Table 1) and the entire correlation matrix of these items as well as their content is easy to find from the open source data at Dataverse (Condon et al., 2017b). The structure of facets based upon 2,889 items with data from the Eugene-Springfield data set (Goldberg, 2008; Goldberg & Saucier, 2016) has been reported by (Schwaba et al., 2020).

*Response Alternatives*

Although many older scales (e.g., the EPI, Eysenck & Eysenck, 1964) were formed from dichotomous items (Agree-Disagree or Yes-No) more typically items are given with multiple response alternatives. This is in order to increase the discriminability of each item. Some argue for an odd number of alternatives, to allow for a middle or neutral response; others suggest that an even number of alternatives forces people to make a decision as to direction and does not include a neutral response. Increasing the number of alternatives increases the internal consistency of the scale up to about 5-6 alternatives, gains beyond that are at the expense of increased effort on the part of the participant (Simms et al., 2019).

As a way of simplifying scale construction, Likert (1932); Likert et al. (1934) introduced the "Likert item" (merely an item with symmetric alternatives representing agreement with or against an attitude). Combining multiple such items produced a "Likert scale". Standard usage now is to refer to "Likert-like items" and the resulting "Likert-like scales". A single item should not be confused with a scale: e.g., a " 7-point Likert scale" is incorrect usage. The original paper (Likert, 1932) is a masterpiece of scale construction. It

Table 1: Sources of personality items used by (Condon, 2018) and (Schwaba et al., 2020). Of the 2689 items in the 21 inventories, Condon identified 696 unique items that captured most of the content of multiple dimensions and scales. These 696 formed the basis of the SAPA inventory items and data that are in the public domain. For the actual items see Condon (2018, 2019). The analysis by Schwaba and colleagues (2020) was done at the facet level. For the references to the tests listed in the table, please see Condon (2018); Schwaba et al. (2020). For the full list of the 696 items, see the ItemInfo.tab in (Condon et al., 2017b).

| **Measures and authors (see Condon (2018); Schwaba et al. (2020) for the references)** | **SAPA N Items** | Schwaba (2022) N Facets | Schwaba (2022) N items |
|---|---|---|---|
| Abridged Big Five Circumplex (Hofstee, de Raad & Goldberg, 1992) | **278** | 45 | 485 |
| Big-Five Aspects (DeYoung, Quilty, & Peterson, 2007) | **100** | 10 | 98 |
| Big-Five Factor Markers (100 items) (Goldberg, 1992) | **100** | NA | NA |
| Big-Five Factor Markers ("Mini-IPIP") (Donnellan et al., 2006) | **20** | NA | NA |
| Big Five Inventory (Soto & John, 2009) | **NA** | 10 | 35 |
| Behavioral Inhibition System/Behavioral Approach System Scales (Carver & White, 1994) | **NA** | 4 | 20 |
| California Psychological Inventory (Gough, 1996) | **201** | 20 | 462 |
| Eysenck Personality Questionnaire - Revised (Eysenck, 1985) | **79** | NA | NA |
| HEXACO-PI (Lee & Ashton, 2004) | **240** | 24 | 192 |
| Hogan Personality Inventory High Level Scales (Hogan & Hogan, 2007) | **111** | NA | NA |
| Hogan Personality Inventory Homogenous Item Clusters (Hogan & Hogan, 2007) | **179** | NA | NA |
| Hogan Personality Inventory (Hogan & Hogan, 1992) | **NA** | 35 | 193 |
| IPIP Interpersonal Circumplex Scales (Markey & Markey, 2009) | **19** | NA | NA |
| IPIP-Multidimensional Personality Questionnaire (Tellegen & Waller, 2008) | **127** | NA | NA |
| IPIP NEO-PI-R $^{TM}$ Facets (Johnson, 2014; Maples et al., 2014) | **120** | NA | NA |
| Jackson Personality Inventory (Jackson, 2004) | **111** | NA | NA |
| Jackson Personality Inventory-Revised (Jackson, 1994) | **NA** | 15 | 300 |
| Multidimensional Personality Questionnaire (Tellegen, 2003) | **127** | 11 | 276 |
| NEO-PI-R $^{TM}$ (Costa & McCrae, 1992) | **300** | 30 | 240 |
| Plasticity & Stability scales (DeYoung, 2010) | **40** | NA | NA |
| Questionnaire Big-Six (48 items) (Thalmayer, Saucier, & Eigenhuis, 2011) | **48** | NA | NA |
| 7-factor scales (Saucier, 1997) | **54** | NA | NA |
| 6-Factor Personality Questionnaire (Jackson, Paunonen, & Tremblay, 2000) | **137** | 18 | 108 |
| 16 Personality Factor Questionnaire (Cattell, 2002) | **109** | NA | NA |
| Sixteen Personality Factor Questionnaire, Fifth Edition (Conn & Rieke, 1994) | **NA** | 16 | 185 |
| Temperament and Character Inventory (Cloninger et al., 1994) | **189** | 30 | 295 |
| Total number of items/facets included | 2689 | 268 | 2889 |
| Total number of unique items | 696 | NA | NA |

considers multiple scaling procedures ranging from the complex to the simple sum score as well as considering the meaning of such scores when considering national and international issues. A critique of Likert like items is that the assumption that the response differences between strongly agree and agree are equivalent to those between agree and disagree is unreasonable (Loevinger, 1957). Indeed, to Loevinger (1957) by cofounding intensity with direction, Likert-like items should be avoided.

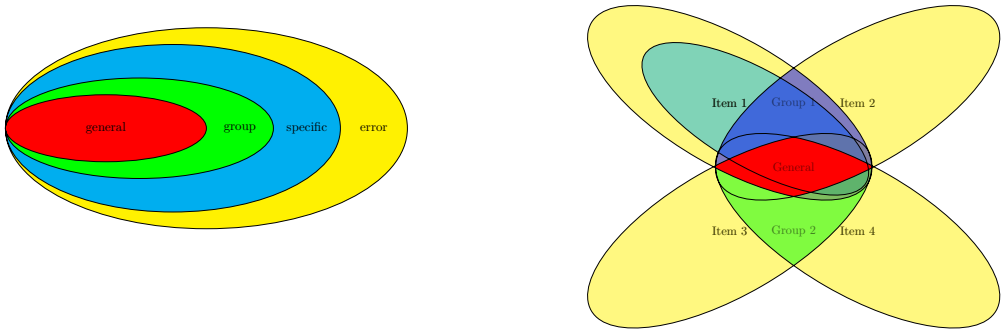
An alternative to the simple choices in Likert-like items Zhang & Savalei (2016) show that using phrases as response alternatives improves factor structure (but a cost of greater time taken for each item).

### *Decomposing Item Variance*

Traditionally, items are thought to reflect mostly error with just a small fraction of the item variance reflecting some underlying score. This reflects that even good items rarely have correlations with other items $> .3$. This is why items are typically aggregated with other items into composite scales. A more careful analysis suggests that the variance of a single item may be thought of in terms of four separate parts: *general* variance which is associated with all items in a scale, *group* variance which is that which is associated with some but not all items, *specific* variance which is associated just with that item, *error* reflects instability of the construct from moment to moment (Figure 1). The reliable (e.g., non-error) variance is the sum of the first three of these components and will contribute to the items stability over short periods of time. When examining the stability of items over 15 minutes with 143 intervening items Condon (2022) found values between .6 and .7 for most items. This suggests that the reliable item variance ($1 - \sigma^2_{error}$) is about 60-70% of the total item variance. An unknown part of the specific item variance is due to the linguistic competence of participants to understand the item wording. While the specific variance can be estimated by item stability, the general and group components of variance can not be identified for a single item, for an item is known by the company it keeps.

Items can share variance associated with all other items (general variance) and can also share variance with just some subset (group variance) of items. Determining the amount of variance in each of these requires some structural analysis of the items (Figure 1).

As an example of items that strongly share group factors associated with stereotypic Feminine gender-role behaviors and interests and stereotypic Masculine gender-role behaviors and interests but that do not share a general factor of gender identity, consider six items taken from Ursala Athenstaedt (2003) who examined stereotypical gender-role behavior. “Sewing on a button” and “Making a bed” are both highly correlated with each other (.64) and with reported gender (.63 and .53) as are “Doing Repairs” and “Changeing Fuses” (.70) which also correlate strongly with reported gender (-.61 and -.65). However, these two groups of items do not share a general factor of gender identity and in fact just correlate -.12 with each other.




A) B)

*Figure 1.* Items may be decomposed into shared general variance, shared group variance, unique specific but reliable variance, and error variance (Panel A. An item's reliable variance can be estimated by test retest correlations of the items. However, general and group are only defined in terms of other items formed into a scale (Panel B).

Table 2: Six example items from Athenstaedt (2003) show clear independent cluster structure of stereotypic interests for Feminine and Masculine gender roles which both have high correlations with gender, but do not share a general "gender" factor.

Raw correlations

| Variable | Sew on B | Make Bed | Do Iron | Do Reps | Chng Fu | Shvl Snow | Gender |
|---|---|---|---|---|---|---|---|
| Sew on Button | 1.00 | | | | | | |
| Make Bed | 0.64 | 1.00 | | | | | |
| Do Ironing | 0.68 | 0.59 | 1.00 | | | | |
| Do Repairs | -0.06 | -0.09 | -0.17 | 1.00 | | | |
| Change Fuses | -0.11 | -0.08 | -0.16 | 0.70 | 1.00 | | |
| Shovel Snow | -0.01 | 0.02 | -0.09 | 0.51 | 0.49 | 1.00 | |
| Gender | 0.63 | 0.53 | 0.67 | -0.61 | -0.65 | -0.49 | 1.00 |

Average within and between group correlations, corrected for item overlap.

| Variable | F | M | MF | gender |
|---|---|---|---|---|
| F | 0.64 | | | |
| M | -0.08 | 0.57 | | |
| MF | 0.33 | -0.30 | 0.29 | |
| gender | 0.61 | -0.59 | 0.60 | 1.00 |

*Composite scales as sum of items*

Ever since Spearman (1904b) aggregated classroom performance and Binet averaged responses to various simple tasks to form a composite measures of intellectual ability, (Binet & Simon, 1905, 1916) psychologists have been constructing composite scales. Although started with measures of ability, the construction of psychological scales quickly spread to studies of interests (Strong Jr., 1927), values (Allport & Vernon, 1933) and temperament (Bernreuter, 1931). This explosion of interest in measurement coincided with the beginning of the systematic study of individual differences initiated by Galton (Galton, 1865, 1884). (For a discussion of the early leaders of the study of individual differences, see Dawis, 1992).

Composite scales are formed by giving multiple items thought to measure the same construct or to have predictive validity for a particular criterion. How to form such scales is discussed below.

## Administering and scoring the items

An unappreciated part of scale construction is the actual administration of the items to be examined. This involves considering the target population for the scale(s). What is the presumed limit of generalizability of the scale? Are there age, educational, or occupational constraints? E.g., if targeting college students, items developed for elementary or high school students are probably not appropriate. Similarly, items designed for college educated participants might not be appropriate for those with less education.

An important question is how many participants should be in the original test development? This is difficult to answer because it partly depends upon the number of items/factor. A simple simulation suggests that about 400 participants are needed to adequately recover a five factor structure of the 25 items of the `bfi` data set, but only 300 are needed to recover five factors from the 135 items of the `spi` and even fewer (200) to get a clean two factor solutions of the 72 items of the `msqR` data set.[4] Thus we agree with the suggestion (Clark & Watson, 1995) that 300 participants is a good goal, but in the era of web based data collection, while more is always better, the increase in benefit is a function of the square root of the sample size. A useful way to increase the number of items given but with each participant taking far fewer items each is to randomly sample items across subjects (Revelle et al., 2017, 2021).

When administering the items, it is important that the participants are fully engaged in taking the items. If the items are given using paper and pencil individually, or in small groups, having the test administrator appear interested is important. Many tests are now given using the web, and it is useful to know why the participants are taking the test. Are "bots" trying to defraud the researcher, are the answers given by MTurk workers paid by the item, or are they actual volunteer participants giving real answers. Screening for too

[4] A solution for a subset of subjects was considered to match the entire sample if the corresponding factors had congruence coefficients $> .95$.

rapid responding, inconsistent responding, failure to answer *validity* questions (e.g., answer 3 to this item), looking for patterning of responses (e.g., "straight lining" , "ski sloping", long strings of identical response), is recommended (Meade & Craig, 2012; Reyes, 2020; Ward & Meade, 2018).

A recurring problem when giving tests either on the web or in person is that some unknown percentage of subjects give dishonest answers, either intentionally to deceive, or due to a lack of involvement (careless responding) (Nichols & Greene, 1997; Johnson, 2005; Meade & Craig, 2012). (In addition to carelessness and deception, Johnson also includes linguistic incompetence as a threat to item validity). One solution is to include psychological antonyms (Goldberg & Kilkowski, 1985). People who say they are both wide awake and sleepy are presumably not paying attention. The threat of careless or deceptive responses is real and can have serious effects on surveys (Arias et al., 2020).

If answered by pencil and paper, the data need to be transcribed to a machine readable form by hand entry. This is a surprising source of error, and the data should be entered twice by independent workers and the results compared (automatically). If entered electronically, the accuracy of the software collection system needs to be validated.

Once entered into a machine readable format, basic descriptive statistics allows for the detection of impossible responses (the range of responses should not exceed the maximum - minimum possible response) (`describe`). Automatic outlier detection techniques include examining the Mahalobinis distances for the responses for each respondent (`outlier`).

### Determining the structure of the resulting items/scales

If a large number of items have been written to measure the construct of interest, some are better than others. We choose the best items to form a scale and reject those items that do not improve the quality of the scale. Even if all of the items are measures of the construct, by choosing the best ones we can develop shorter measures of the same construct. We can choose the best ones in two different ways that meet two different needs. Borrowing for the terminology of *Machine Learning* we can refer to these two approaches as *unsupervised* and *supervised* learning. Unsupervised learning procedures are dimension reduction techniques that examine the internal structure of multiple groups of items, these include principal components, factor analysis, or cluster analysis. Supervised learning merely means that we have a criterion against which to choose the best items. Supervised learning is a generalization of basic regression approaches to choose items based upon the correlation with the criterion. We will first discuss unsupervised procedures that help us determine the structure and the construct validity of a set of scales, and then supervised procedures in discussions of predictive validity.

If interested in measures of specific constructs, it is important to evaluate the internal consistency and structure of the scales to be formed. If interested in developing scales that are used to predict particular criteria, internal consistency is less important, but should be examined anyway.

*The correlation or covariance matrix*

The first step in evaluating structure is to find the inter-item correlation or covariance matrix for all of the items. This is typically just the Pearson Product Moment Correlation Coefficient (PPMCC) which for matrices of deviation scores $\mathbf{X}$ and $\mathbf{Y}$ with elements x and y is just

$$r_{xy} = \frac{\Sigma\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}.^{5} \tag{1}$$

A correlation is merely the covariance of standard scores. Matrices of correlations may be found by multiplying the covariance matrix by the square root of the inverse of the variances. The variances are just the diagonal of the covariance matrices of the $\mathbf{X}$ and $\mathbf{Y}$ matrices:

$$\begin{aligned} \mathbf{C}_{xy} &= \mathbf{X}\mathbf{Y'}\mathbf{N}^{-1} \\ \mathbf{V}_x &= diag(\mathbf{X}\mathbf{X'}\mathbf{N}^{-1}) \\ \mathbf{V}_y &= diag(\mathbf{Y}\mathbf{Y'}\mathbf{N}^{-1}) \\ \mathbf{R}_{xy} &= \mathbf{V}_x^{-.5}\mathbf{C}_{xy}\mathbf{V}_y^{-.5}. \end{aligned} \tag{2}$$

Correlations may be found by using the `cor` function, covariances by the `cov` function (both in base R). Equation 1 works with deviations of raw scores as well as of ranks (Spearman $\rho$). For dichotomous data, Formula 1 is known as the $\phi$ coefficient. For arbitrarily dichotomized data taken from a bivariate normal distribution, $\phi$ underestimates the observed r for the undichotomized data. The `tetrachoric` correlation estimates the latent Pearson from the two by two table associated with $\phi$. This is not a closed form calculation but rather is done by iteratively trying different values of $\rho_t$ that with the assumption of underlying normally distributed x and y will best fit the observed two x two probabilities. Similarly, the `polychoric` correlation will estimate the bivariate normal correlation of x and y for categorical data where the categories are assumed to be taken from the bivariate normal. If the number of alternative responses is less than 6 or 7, the polychoric will be noticeably larger than the observed Pearson. As the number of categories increases, these differences vanish. The tetrachoric, $\rho_t$, and polychoric, $\rho_p$ correlations are thus estimate of what the Pearson correlation would be between latent variables represented by the artificially (dichotomized for tetrachoric) grouped observed variables. Although useful for determining structure, these coefficients should not be used for estimating reliability for they inflate the estimate (Revelle & Condon, 2019).

A useful, but probably not intuitive, understanding of the correlation coefficient is as the cosine of the angle between two vectors (X and Y) in the N dimensional space defined by the N participants. Other ways of understanding the correlation include the square root of the amount of variance in Y accounted for by the liner regression of Y on X, or

[5] This formula ignores sample size because it is entered in both the numerator and denominator, and thus cancels out.

as the slope of the regression line between two standardized variables (see Rodgers & Nicewander, 1988, for a more thorough treatment of the correlation coefficient).

Matrices of correlations or covariances are the basis for dimension reduction procedures such as factor analysis, and are also the basis of multiple regression. The standardized coefficients for predicting $\mathbf{Y}$ from $\mathbf{X}$ are just

$$\begin{aligned} \beta_{y.x} &= \mathbf{R}_{xx}^{-1}\mathbf{R}_{xy} \\ \hat{\mathbf{Y}} &= \beta_{y.x}\mathbf{X}. \end{aligned} \tag{3}$$

In R these coefficients may be found by the `lm` (core R) or `setCor` (*psych* package) functions.

Correlations may be displayed in tabular form (Table 2) or as "heatmaps" showing the strength of the relationship (Figure 2). It should be clear by inspection that the 10 neuroticism items shown in panel A all correlate strongly with each other, although the last three correlate negatively with the first seven. Similarly, by inspection, the 10 sex role items seem to show two different clusters of items. Thus, we would probably want to find one neuroticism score, but two sex role scores. But how do we do this organization of structure if we don't trust our ability to inspect correlation matrices? This is a question of the dimensionality of the items and is solved through conventional applications of linear algebra known as principal components and factor analysis.

*Number of dimensions*

A data matrix, $\mathbf{X}$ with N subjects and n variables is of rank (min(N,n)) which is typically n (the number of variables). If the variables are correlated at all, this space may be approximated by a lower rank space of dimension k. That is, the n dimensional space is projected onto a lower dimensional space. The problem becomes what is the the value of k that summarizes the data with the least loss of information. (This notion of a lower level space providing a useful model of a higher dimensional space is most easily seen in the use of the mean as a summary statistic. For the mean is merely a projection of all the data points to one value, the mean. The goodness of fit of the mean to the data is a function of the average squared deviation of the actual data from the mean, that is the variance.) The advantage of dimension reduction is while few people can intuit information from more than a two or three dimensional space, the data for individuals given just 10 items needs to be represented in a 10 dimensional space and the curse of dimensionality overwhelms our intuition (Del Giudice, 2021).

*Dimension reduction through factor analysis or principal components analysis*

Any correlation matrix, $\mathbf{R}$ can be represented by the product of orthogonal vectors. $\mathbf{X}$ (the eigen vectors) scaled by a vector of relative importance $\boldsymbol{\lambda}$ (the eigen values)

$$\mathbf{R} = \mathbf{X}\boldsymbol{\lambda}\mathbf{X}' \tag{4}$$

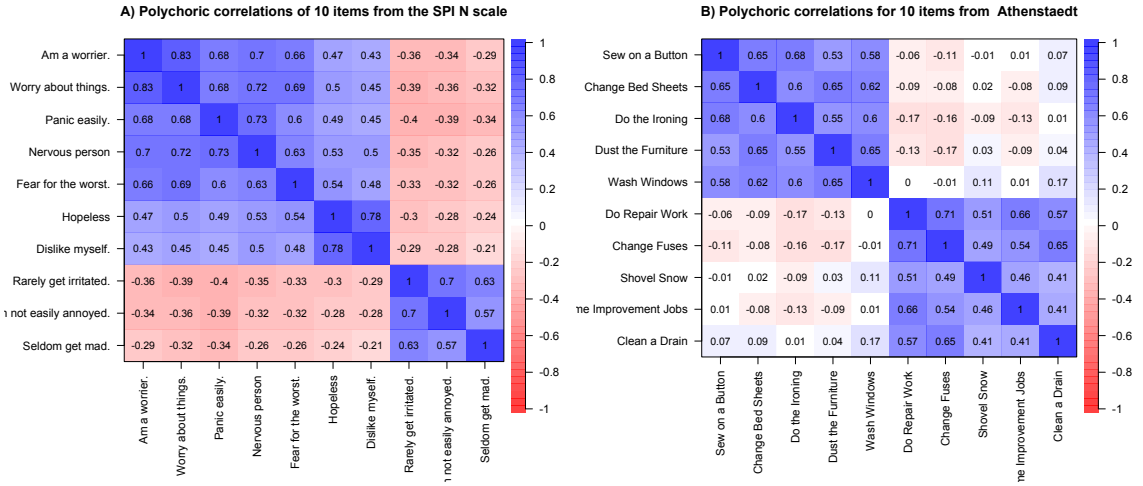



*Figure 2.* The importance of examining item structure. Panel A shows the correlations of 10 items from the SPI Neuroticism scale (Condon, 2018); Panel B 10 items from the Athenstaedt (2003) study of gender role behaviors and interests. Both scales show high $\alpha$ values but differ in their general factor saturation ($\omega_g$). The Neuroticism items show a clear general factor ($\omega_g = .57, \alpha = .88, \omega_t = .91, \bar{r} = .43, r_{median} = .39$) with two minor group factors; the Athenstaedt items show two clear group factors with no evidence for a general factor ($\omega_g = .15, \alpha = .77, \omega_t = .85, \bar{r} = .25, r_{median} = .14$). check these numbers for polychoric or raw correlations

This is interesting, but not very useful, for it does not reduce the dimensionality. But if we define $\mathbf{C} = \mathbf{X}\sqrt{\boldsymbol{\lambda}}$ (the principal components) and choose just the first k component vectors associated with the largest values of $\boldsymbol{\lambda}$ we find

$$\mathbf{R} \approx \mathbf{C}\mathbf{C'}. \tag{5}$$

The values of $\mathbf{C}$ are found that best estimate $\mathbf{R}$ for the rank k. But $\mathbf{R}$ is made up of diagonal of 1s and the off diagonal correlations. If we want to just find the best estimates of the correlations, we use the factor model

$$\mathbf{R} \approx \mathbf{F}\mathbf{F'} + \mathbf{U}^2 \tag{6}$$

where $\mathbf{U^2}$ is a a diagonal matrix of *uniquenesses*. While equation 5 can be solved analytically (e.g., `pca`), equation 6 needs to be solved by iteratively trying various values of $\mathbf{U}^2$ (e.g., `fa`). Essentially *factor analysis* is doing an eigen value decomposition using k factors, not of the original $\mathbf{R}$ matrix but of a reduced matrix where the diagonal is formed from $h^2 = 1 - u^2$.

Two vocabulary items used in these procedures are the *communalities*, $h^2$, and the *eigen values*. The former is just the amount variance in a variable accounted for by all

of the components or factors and is equal to the diagonal of the approximated correlation matrix without the addition of the uniquenesses, that is, the $diag(\mathbf{CC'})$ or $diag(\mathbf{FF'})$. The other is the size of the eigen values or the sum of squares of the loadings on each factor.

These two procedures are conceptually very different in that components represent sums of items, while factors are models of the items. Items are said to be reflective indicators of the factors while they are formative indicators of the components. Factors are thus seen as causing the items, while components are just (weighted) sums of the items. As the number of items increase, the importance of the diagonal lessens and effectively the results of the two models converge. Factor loadings (the correlation of the factors with the items) can be thought of as the asymptotic limit of component loadings as the number of variables defining each component increases.

More important than the components versus factors question is what is the optimal number of factors to extract and to what extent should the solutions be transformed into structures that appear more simple. The number of factors question is very difficult, and there is no agreed upon answer. Among the many ways to answer what is the optimal number, unfortunately probably the worst one has been implemented in several commercial statistical packages. Asking how many eigen values are greater than one (the so called "little jiffy criterion" ) is probably the worst answer, and there are several appealing alternatives. Parallel analysis (Horn, 1965) is simply a comparison of the eigen values of the observed solution to eigen values of random data of the same number of variables and subjects. Multiple answers to the number of factors question are given by the `nfactors` function in *psych*.

Yet another unsolved problem is the question of what is the optimal transformation/rotation of the observed factors.[6] Factors as extracted are difficult to interpret as the factors are extracted to maximize the variance explained by each successive factor. This results in the situation that most variables will share high correlations (loadings) with the first factor and then half will tend to have positive and the half negative loadings with the second factor. Although optimal in prediction, these solutions are very hard to interpret. A solution is to transform the original factor solution into one that has "simple structure" which is typically taken to mean an independent cluster model (i.e., most variables have high correlations with one just one factor, with the remaining correlations being small). These transformations will not affect the communality of the variables, but will affect the amount of variance accounted for by each factor. A number of such transformations or rotations are available in the *GPArotation* package (Bernaards & Jennrich, 2005). A popular othogonal rotation is `varimax`, popular oblique tranformations include include `geomin` and `oblimin`. Different software packages have different defaults and it is important to know which one is used.

Factor analysis is typically used in scale construction to help identify the items that

[6] The term rotation should be reserved for orthogonal transformations of the factors, the more general term transformations allow for correlated or oblique solutions. Unfortunately, common usage seems to have lost this distinction and rotations may be said to be orthogonal or oblique.

should be formed into a scale. Factor analytic techniques may be divided into Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). Both procedures offer goodness of fit statistics based upon the residuals of the correlations and differ primarily in the number of parameters estimated. EFA finds loadings for all variables on all factors and then rotates these solutions to try to maximize some rotational criterion based upon the dispersion of the factor loadings. CFA, on the other hand, typically tries to estimate a cluster like model in which each variable is defined by just one factor. The items loading on the factor must be specified ahead of time to conduct a CFA. When comparing the dimensions identified from our set of 10 gender role items, both EFA and PC produce identical sets of items (Table 3).

Table 3: Comparing factor analysis and components analysis of 10 items from Athenstaedt (2003) (Figure 2 panel B). The loadings (RC1 and RC2 are slightly larger for the PCA solution than the factor solution (MR1 and MR2( but the residuals are larger (Figure 3). Salient loadings (those > .4) are boldfaced. h2 (the communality) is the amount of common variance modeled by the factors or components and is just the sum of squares of the loadings. Not shown, but the correlation between MR1 and MR2 = -.08. The factor congruence coefficients for these two solutions are .998 and .998.

| MR1 | MR2 | h2 | RC1 | RC2 | pc h2 | Item |
|---|---|---|---|---|---|---|
| **0.81** | 0.00 | **0.66** | **0.85** | -0.02 | **0.72** | Change Bed Sheets |
| **0.80** | 0.11 | **0.64** | **0.84** | 0.10 | **0.71** | Wash Windows |
| **0.78** | 0.01 | **0.60** | **0.83** | -0.01 | **0.68** | Sew on a Button |
| **0.77** | -0.10 | **0.62** | **0.82** | -0.13 | **0.69** | Do the Ironing |
| **0.75** | -0.05 | **0.58** | **0.81** | -0.07 | **0.67** | Dust the Furniture |
| -0.04 | **0.86** | **0.75** | -0.10 | **0.87** | **0.77** | Do Repair Work |
| -0.07 | **0.84** | **0.71** | -0.12 | **0.86** | **0.75** | Change Fuses |
| 0.15 | **0.70** | **0.50** | 0.13 | **0.77** | **0.61** | Clean a Drain |
| -0.01 | **0.69** | **0.48** | -0.05 | **0.77** | **0.59** | Do Home Improvement Jobs |
| 0.07 | **0.61** | 0.37 | 0.04 | **0.70** | **0.50** | Shovel Snow |
| 3.10 | 2.81 | | 3.48 | 3.21 | | Sums of squares |
| .21 | .28 | | .35 | .32 | | Proportion of total variance |

When evaluating goodness of fit of any model, it is important to examine the residuals. In this case the residuals are just the original correlations - the modeled correlations. Because factor analysis is modeling just the correlations, the residual correlations are smaller than they are for the principal components (Figure 3). In the case of such clear structure, the salient loadings are identical in both models.

*Bifactor and higher order representations*

Alternative to simple factor models where items are thought to load on just one primary factor are bifactor models or higher order models. The bifactor model (Holzinger & Swineford, 1937; Reise et al., 2007; Reise, 2012) allows two major loadings for each item, one on general factor, the second on a group factor. Bifactor solutions can be found using

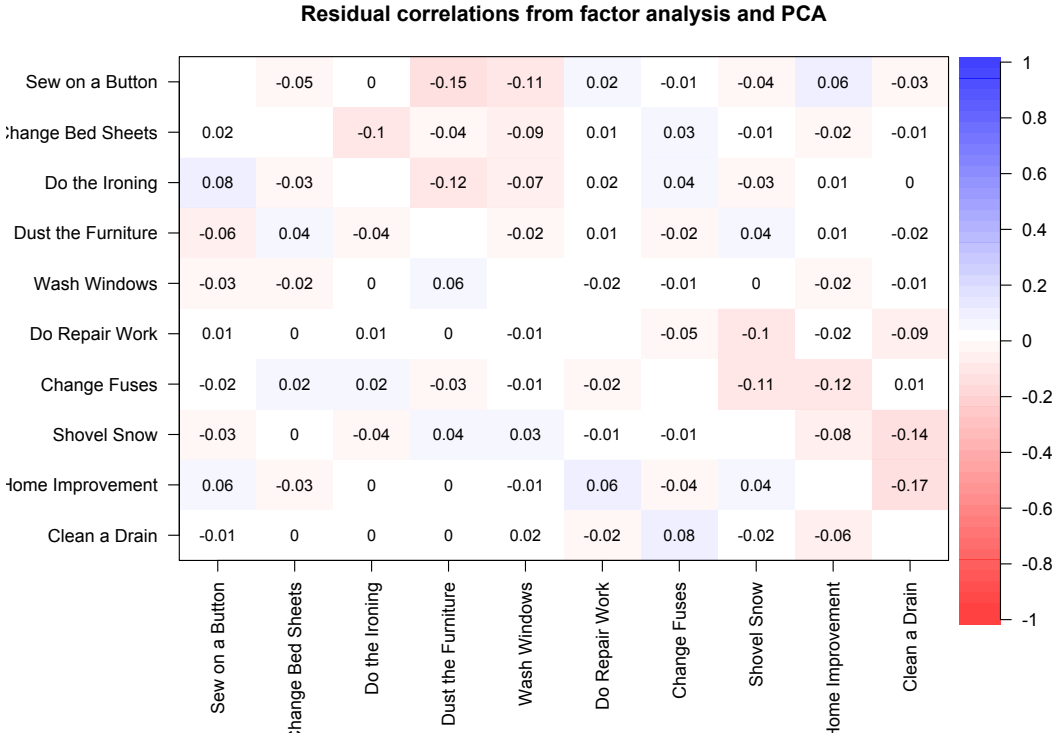



*Figure 3.* Residual correlations for 10 items taken from the Athenstaedt (2003) data. The lower off diagonal elements are from the factor solution, the upper off diagonal from the principal components solution. The factor and components solutions are shown in Table 3.

the `bifactor` rotation applied to the EFA solution, as can a transformation of the loadings (Schmid & Leiman, 1957) from a correlated factors model to a bifactor solution.(Figure 4).

*Cliuster analytic approaches to dimension reduction*

Not as mathematically as well defined as *principal components* or *factor analysis* a somewhat more intuitive approach to forming composites from items is *hierarchical cluster analysis* (e.g., `iclust`, Revelle, 1979). The basic algorithm is straightforward: combine the most similar pair of variables to form a new variable to replace the original pair and then repeat this process until some criterion is reached. One such criterion is to find the maximum value of $\alpha$, another, which has shown to be more useful is to find the maximum value of $\beta$. Where $\beta$ is defined as the worst split half reliability. `iclust` solutions for the ten anxiety and the 10 MF items are shown in Figure 5.

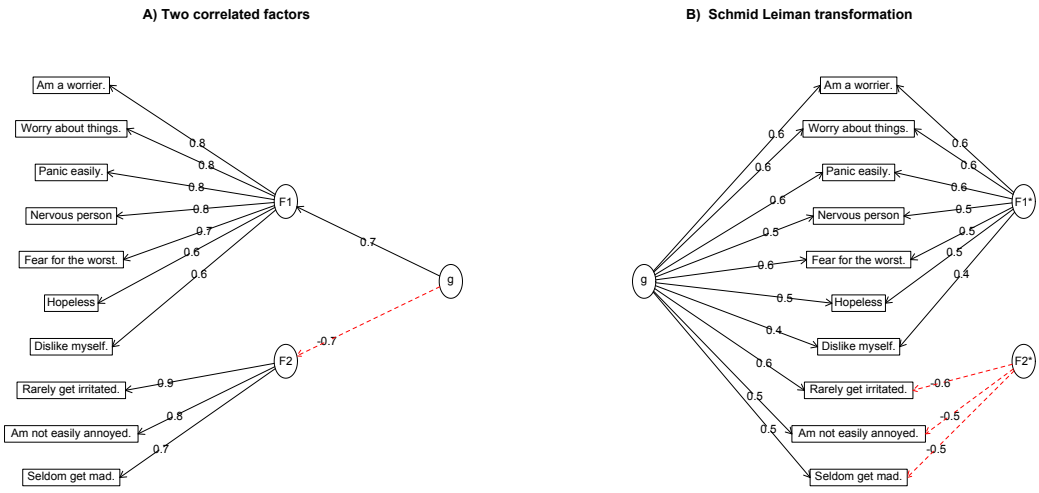



*Figure 4.* An example of the Schmid-Leiman transformation applied to two correlated factors of anxiety. Panel A displays two correlated factors with a higher order general factor. Panel B is the result of a Schmid-Leiman transformation of this solution and shows a bifactor like solution. See Table 4 for the loadings.

Table 4: A Schmid-Leiman transformation applied to two factors of the Neuroticism items (Figure 2, panel A). All of the items share a general factor loading (g), and the two group factors show a clear structure. h2 = communalities, u2 = uniquenesses, p2 is the amount of general factor variance that is common for each item. $\omega_g$ is the squared sum of the general factor variance divided by the total variance = .59. $\omega_t$ is the total amount of variance accounted for by the model = .93.

| g | F1. | F2. | h2 | u2 | p2 | item |
|---|---|---|---|---|---|---|
| **0.58** | **0.60** | 0.03 | **0.70** | 0.30 | **0.48** | Would call myself a nervous person. |
| **0.61** | **0.60** | -0.01 | **0.73** | 0.27 | **0.51** | Worry about things. |
| **0.59** | **0.59** | 0.01 | **0.69** | 0.31 | **0.50** | Am a worrier. |
| **0.55** | **0.55** | 0.01 | **0.61** | 0.39 | **0.49** | Fear for the worst. |
| **0.60** | **0.52** | -0.07 | **0.64** | 0.36 | **0.56** | Panic easily. |
| **0.48** | **0.48** | 0.01 | **0.47** | **0.53** | **0.50** | Feel a sense of worthlessness or hopelessness. |
| **0.45** | **0.45** | 0.00 | **0.40** | **0.60** | **0.50** | Dislike myself. |
| **-0.63** | 0.00 | **0.62** | **0.78** | 0.22 | **0.51** | Rarely get irritated. |
| **-0.57** | 0.01 | **0.55** | **0.63** | 0.37 | **0.52** | Am not easily annoyed. |
| **-0.51** | -0.01 | **0.51** | **0.52** | **0.48** | **0.50** | Seldom get mad. |

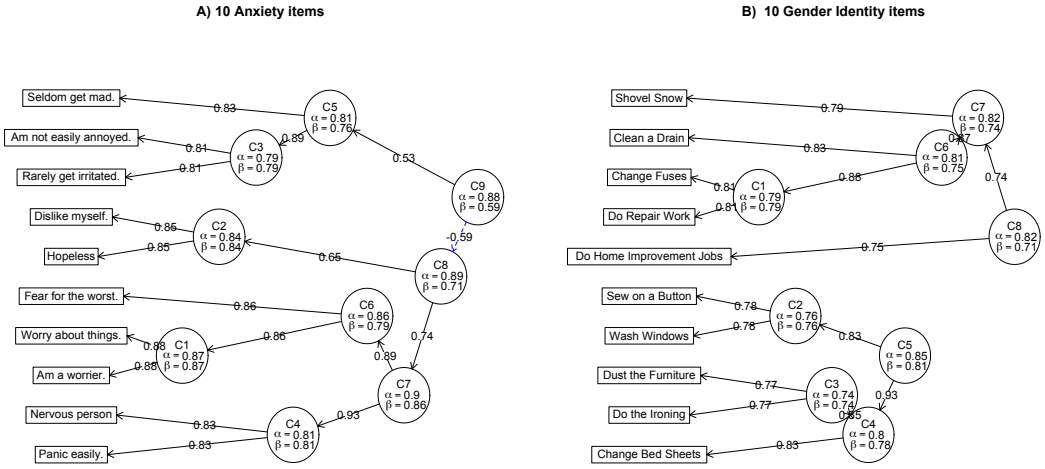



*Figure 5.* Hierarchical cluster analysis using the `iclust` algorithm for the 10 anxiety items from the `spi` (panel A) and 10 gender-role behavior items from (Athenstaedt, 2003) (panel B). Values inside each cluster (circle) are the cluster name and the values of $\alpha$ and $\beta$. Path coefficients show the correlations of the clusters with items and with each other. The cluster structure shows more information about the similarity of the items than is found in the factor analytic output.

## Forming, scoring, and evaluating scales

### *Selecting items*

Even the best item writer will have some ideas that work better than others. Item discriminations may be seen by rank ordering the factor loadings. Items with low factor loadings should be dropped. Other candidates are completely redundant items (e.g.,"Like to attract attention" and "dislike being the center of attention".) Although useful for tests of response validity, two such items are very redundant and one would be a candidate for elimination. Items will also differ in their endorsement frequency, and items with the same level of endorsement frequency are candidates for elimination.

### *Forming composite scales*

Once the items to use for a scale have been defined based upon the results of an EFA or cluster analysis, the next step is to find scores from them. Although some have proposed factor scores as the appropriate way to do this, simple sum scores have been shown to be just as good, if not better. This is primarily because unit weighted scores are actually sample independent (they are all weighted equally), while factor scores are based upon weights derived from from a particular sample (Widaman & Revelle, 2022). Further more, as the number of items increases, the importance of differential weights becomes

less. Fields differ in the use of sum scores (just the sum of the responses) versus mean scores (the mean of the item responses). Clearly the item sums are just a linear function of the item means. We recommend mean scores as they are in the unit of the items and are independent of the number of items (thus, if items range from 1 to 6, then scale scores will as well). To interpret sum scores, on the other hand, one needs to know how many items there are and what is the range of each item.

When items have negative correlations with the overall scale they need to be reversed. Reverse scoring an item is equivalent to subtracting the item from the highest option + the lowest option (i.e, for items that range from 1-6, a reversed item is subtracted from 7. This is the same as the laborious procedure of recoding 1s as 6s, 2 as 5, 3s as 4, etc.) Some prefer to do this before starting, we recommend keeping the items in their original direction and then allowing the scoring program to do the reversals. Thus, we form lists of items to be scored and the direction in which to score them.

In the appendix we show how use functions from the *psych* package to form composite scales with some basic information about the quality of the scale. Scoring multiple scales at one time allows for an examination of the correlations between scales, as well as the internal structure of each scale. If scales include overlapping items, `scoreOverlap` gives statistics that correct for the artificial inflation of the correlations for overlapping scales.

*Reliability*

It has been known since Spearman (1904a) that tests are "befuddled with error" (McNemar, 1946, p 294) and that observed correlations are attenuated estimates of latent correlations. Estimates of reliability are used to correct this attenuation to find the underlying correlation. Unfortunately, there are many ways to estimate reliability that differ depending upon whether they use just one measure (e.g., $\alpha$, the range of split halfs, $\omega_g$, $\omega_t$, or two measures at one time (alternate form) or one measure at two time points (test-retest) (Revelle & Condon, 2019). Of the measures based at one time point, one ($\alpha$) requires the least calculation but is also much less informative than "model based" estimates such as $\omega_g$ or $\omega_t$. McDonald (1999) introduced two model based estimates based upon the factor structure of the items, unfortunately, he used the same name, $\omega$ for both of these. Zinbarg et al. (2005) relabeled these as $\omega_g$ for the general factor saturation and $\omega_t$ for the total reliable variance[7]. $\alpha$ is a generalization of the Spearman (1910); Brown (1910) correction for test length and reflects the number of items and the average item-intercorrelation. As we saw when comparing panels A and B in Figure 2, that items have high average correlations does not necessarily reflect a general construct. Guttman (1945) reviewed six ways of estimating reliability from one test, but concluded that the better procedure was to examine the test-retest correlation.

[7] Although Zinbarg et al. (2005) referred to the general factor estimate as $\omega_h$ for the hierarchical way in which it is found, we prefer to use the $\omega_g$ notation to reflect that it estimates the *general* factor of the test (Revelle & Condon, 2019).

That these various estimates provide different results may be seen when comparing $\alpha$, the model based estimates ($\omega_g$ and $\omega_t$) and the test retest correlation over several weeks for five scales from the Eysenck Personality Inventory (Eysenck & Eysenck, 1964) (Table 5). It may be seen that $\alpha$ tends to under-estimate the test-retest reliability and is a large overestimate of the general factor saturation ($\omega_g$) of the test. We do not encourage the use of $\alpha$ and include it for historical reasons only. Except for the test-retest statistics, the `reliability` function in *psych* reports all of the values seen in Table 5. If repeated measures are available, the `testRetest` function will report the item level and scale level correlations. We do not encourage the use of tetrachoric correlations for determining scale reliabilities, but include the average tetrachoric over time to replicate the finding from Condon (2022) that items have much more reliable variance than normally thought.

Table 5: Comparing $\alpha$ to model based ($\omega_g$ and $\omega_t$) consistency measures to test-retest correlations and the minimum and maximum split half reliabilities for the `epiR` data set in the *psychTools* package. 472 participants took the Eysenck Personality Inventory (Eysenck & Eysenck, 1964) twice with a several week delay. $\bar{r}$ is the average within test interitem correlation, $\bar{r}_{retest}$ is the average test-retest correlation for identical items in each scale. $\bar{r}_{retest}{}^*$ is the average tetrachoric correlation between identical items over time.

| Variable | $\alpha$ | $\omega_g$ | $\omega_t$ | Retest | min | max | $\bar{r}$ | $r_{med}$ | $\bar{r}_{retest}$ | $\bar{r}_{retest}{}^*$ | N items |
|---|---|---|---|---|---|---|---|---|---|---|---|
| E | 0.76 | 0.37 | 0.78 | 0.82 | 0.58 | 0.84 | 0.11 | 0.10 | 0.54 | 0.77 | 24 |
| N | 0.81 | 0.43 | 0.82 | 0.79 | 0.69 | 0.86 | 0.15 | 0.13 | 0.52 | 0.75 | 24 |
| L | 0.40 | 0.14 | 0.43 | 0.66 | 0.30 | 0.45 | 0.07 | 0.07 | 0.54 | 0.77 | 9 |
| Imp | 0.52 | 0.31 | 0.59 | 0.72 | 0.41 | 0.61 | 0.11 | 0.10 | 0.52 | 0.74 | 9 |
| Soc | 0.76 | 0.53 | 0.79 | 0.81 | 0.66 | 0.83 | 0.20 | 0.17 | 0.56 | 0.79 | 13 |

*Number of items*

Increasing the number of items in a scale will increase $\alpha$ and $\omega_t$ but not necessarily $\omega_g$. In addition to increasing the total reliability ($\omega_t$), increasing the number of items improves the breadth of the construct. Thus, it is more convincing when measuring a trait such as extraversion to ask more than just whether you enjoy going to parties, but to include a broader definition of the construct (laughing loudly, enjoying being the center of attention). However it is possible to measure some constructs with single items (e.g.,"I have high self esteem", Robins et al., 2001). The power of single item measures (SIMPs) is also discussed by Woods & Hampson (2005) who show that single item scales perform as well as more conventional longer scales in terms of their convergent and discriminant correlations.

## Validity

For perhaps the first 50 years of modern psychological assessment, the concept of test validity was very simple. Did the test predict what it was supposed to predict? Thus, the validity of the MMPI or the Strong was how well they discriminated known groups. In the 1950's however, with a move away from an emphasis upon behaviorism and with a move to theory construction using latent constructs, this approach was strongly criticized (Loevinger, 1957) and the idea of construct validity was introduced (Cronbach & Meehl, 1955; D. T. Campbell & Fiske, 1959). More recently, others have pushed back on the supposed advantages of construct validity and have emphasized that the validity of a test reflects a casual statement: "A test is valid for measuring an attribute if variation in the attribute causes variation in the test scores." (p 1067, Borsboom et al., 2004). Unfortunately, this strong definition is problematic if looking at predictive validity over extended time periods.

Another attack on construct validity is the emphasis on prediction rather than understanding (Yarkoni & Westfall, 2017). Using the example of machine learning where models can be accurate but not transparent as to why particular items work in the model, Yarkoni and his colleagues have challenged the conventional emphasis upon latent variables as necessary for psychological research. Condon & Möttus (2021) have further argued that over-reliance on parsimonious latent variable modeling – while good for those who own and administer psychological tests – has impeded the pace of psychological research, as it necessarily results in the discarding of item-level information. Möttus et al. (2020) emphasize the power of the item rather of the scale to predict outcomes. Unfortunately, the use of items does not further parsimonious theory construction. But nature is not simple and recognizing this complexity is probably advantageous.

### *Construct Validity*

Loevinger (1957) pointed out that because it changes for every criteria, the simple notion of validity as predictive accuracy is inadequate and suggested that there are "two contexts for defining validity be recognized, administrative and scientific. There are essentialy two kinds of administrative validity, content and predictive-concurrent. There is only one kind of validity that exhibits the property of transposibility or invariance under changes in administrative setting which is the touchstone of scientific usefulness: that is construct validity." p 641. To (Loevinger, 1957) constructs are to traits as statistics are to parameters.

The standard treatment of construct validity is to consider what a test measures by its pattern of correlations: does it have high correlations with what theory says it should (convergent validity) and does it have low correlations with theoretically unrelated constructs (divergent validity). Stronger evidence for validity may be found by comparing measures across methods. Methods typically are means of assessment, e.g., self report, peer reports, behavioral ovbservations or physiological measures. Each method is thought

to have contaminating variance such that two traits measured with the same method might correlate because of the shared method. Trait measures across methods should correlate just the extent they both are measures of the same trait (D. T. Campbell & Fiske, 1959).

Thus, Zola et al. (2021) reports the correlations between self reports and peer reports for web based data for ≈ 900 web participants who had nominated peers to rate them on eight traits. We show the correlations of five of these traits in a *Multi-Trait, Multi-Method* matrix (D. T. Campbell & Fiske, 1959) where the main diagonal shows internal consistency estimates, the sub-diagonal between self and peer reports the convergent validities, and the other correlations show discriminative validity (Table 6). All correlations are corrected for unreliability. The strong agreement between self reports and peer reports of the same traits (average convergent correlations of .72) and low mean correlations with the self report measures (.18) or the other peer reports (.09) shows the appropriate pattern of convergent and discriminative validity. That the peer reports correlate amongst themselves with an average correlation of .41 shows less discriminative validity for peer reports than self reports and suggests the presence of a *methods factor*.[8]

Table 6: Multi-Trait, Multi-Method Correlation matrix adapted from (Zola et al., 2021). Data were collected as part of the SAPA project using a massively missing data structure. 158,631 participants responded to random subsets of items in the SAPA personality inventory (Condon, 2018). Each anonymous participant was asked to nominate anonymous raters. Raters provided ratings for 908 participants on 8 measures. For simplicity, we show the MTMM correlations for five of these. The diagonal shows internal consistency estimates of reliability. All correlations are corrected for reliability.

| | Self reports | | | | | Peer reports | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | Agrbl | Cnscn | Nrtcs | Extrv | Opnnn | Agrbl | Cnscn | Stblt | Extrv | IntlO |
| Agreeableness | **0.88** | | | | | | | | | |
| Conscientiousness | **0.32** | **0.87** | | | | | | | | |
| Neuroticism | -0.14 | -0.20 | **0.90** | | | | | | | |
| Extraversion | 0.28 | 0.13 | -0.28 | **0.90** | | | | | | |
| Opennness | 0.09 | 0.06 | -0.10 | 0.14 | **0.86** | | | | | |
| Peer-Agreeableness | **0.78** | 0.15 | 0.00 | 0.00 | -0.14 | **0.45** | | | | |
| Peer-Conscientiousness | 0.21 | **0.79** | -0.16 | -0.01 | -0.06 | **0.37** | **0.61** | | | |
| Peer-Stability | 0.18 | 0.23 | **-0.81** | 0.07 | 0.10 | **0.49** | **0.41** | **0.61** | | |
| Peer- Extraversion | **0.36** | **0.45** | **-0.43** | **0.76** | 0.18 | 0.18 | **0.44** | **0.39** | **0.51** | |
| Peer-IntellectOpenness | 0.21 | 0.12 | -0.26 | 0.14 | **0.49** | **0.49** | **0.48** | **0.52** | **0.31** | **0.47** |

*Predictive Validity*

An alternative to constructing scales based upon principles of internal consistency with an emphasis upon construct validity is to make up scales with items selected to predict

[8] The polychoric correlation matrix of the items is available as the `zola` data set in *psychTools*.

a certain criterion. This of course was the model for inventories such as the MMPI or the Strong (D. P. Campbell & Borgen, 1999). The procedure is extremely simple: choose items with the highest correlation with a criterion and ideally, the lowest correlation with each other.

*Cross validation.* A well known danger of such simple empiricism is the tendency to capitalize on chance. Correlations will fluctuate from sample to sample and the best items from any particular sample will not necessarily be the best in another sample (Cureton, 1950). The solution is surprisingly simple and has been done for decades: cross validation. That is, develop scales on one sample and then validate the scales on a different sample. One traditional way was to split a sample into two parts, the derivation sample and the validation sample. Simple generalizations of this procedure are *k fold resampling* and *bagging* (boot strap resampling with aggregation). K-fold resampling removes 1/k (a fold) from the sample for the derivation sample and validates on the remaining data. This is done k times. Traditional cross validation would thus be a k-fold with k = 2. More modern applications seem to prefer k = 10 to have 90% derivation sample and a 10% cross validation sample. Bagging repeatedly forms bootstrap samples (which because they are sampling with replacement will typically extract 1- 1/e = 63.2% of the subjects) for the derivation sample (the bag) and then validate it on the remaining 1/e of the sample (out of bag). In boot strap resampling, a number (e.g., 100, 1000) of samples of the original sample are formed with each participant being sampled with replacement. The resulting samples will be the same size as the original, but with an average of 62.3% of the original subjects in each random sample. The remaining 37.7% of the sample can be used for cross validation. Values from each sample are then aggregated. Both of these procedures may be done using the `bestScale` function in *psych* (Elleman et al., 2020).

As an example comparing factorially derived scales with an empirical keyed scale, consider the 50 items from the (`GERAS`) Gender Related Attributes Survey from Gruber et al. (2020). These items were chosen to reflect gender differences in three different domains: personality, cognition and behavior. An empirically keyed set of 10 items was found using `bestScales` (Table 7). The validity of this scale can be compared favorably to scales associated with male and female preferences formed for each domain, and then total domain scores, and then total M and F scores. Within each domain, the scales showed high internal consistency, but low correlations between scales (Table 8).

A recent review (Eagly & Revelle, 2022) used this same data set to show that aggregating items improves prediction particularly if the aggregated items share little variance. The advantages of aggregating into short scales is that it reduces the overfitting found when using many predictors. These scales do not need to be unidimensional (high values of $\omega_g$) to be useful (Tables 8,9) but relevant content makes them more understandable (D. P. Campbell & Borgen, 1999). Eagly & Revelle (2022) showed that the predicted validity is a function of the the number of predictors, k, the average item validity ($\bar{r_y}$) and

Table 7: 10 gender related activities most correlated with gender from the `bestScales` function. Items are chosen based upon correlations with the criterion and are not chosen to maximize internal consistency. ($\omega_h$ = .46). Data are from Gruber et al. (2020) and are included in the `GERAS` data set.

| Mean cross validated r | Item |
|---|---|
| 0.45 | Watching a romantic movie |
| 0.43 | Dancing (classic standard dances, ballet, Latin, free dance, etc.) |
| 0.42 | Talking on the phone with a friend |
| 0.41 | Rhythmic gymnastics |
| 0.36 | Gossiping |
| 0.36 | Shopping |
| -0.34 | Watching action movies |
| -0.34 | Watching sports on TV |
| 0.34 | Yoga |
| -0.33 | Writing a computer program |

Table 8: Correlations (corrected for item overlap) and internal consistencies (on the diagonal) of the GERAS domain and total scores. Also shown are three measures of internal consistency ($\alpha$, $\omega_h$ and $\omega_t$ ). The Best 10 items were selected using the `bestScales` function to maximize validity (69.) and ignores internal consistency. The validity of .69 agrees with the results from 100 bootstraps shown in Table 9)

| Variable | MF.all | M | F | Pers | Cog | Act | M.P | F.P | M.C | F.C | M.A | F.A | Best 10 | gendr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MF.all | 0.85 | | | | | | | | | | | | | |
| M | 0.67 | 0.81 | | | | | | | | | | | | |
| F | -0.70 | -0.28 | 0.83 | | | | | | | | | | | |
| Pers | 0.73 | 0.58 | -0.59 | 0.77 | | | | | | | | | | |
| Cog | 0.56 | 0.42 | -0.48 | 0.36 | 0.67 | | | | | | | | | |
| Act | 0.71 | 0.57 | -0.57 | 0.54 | 0.36 | 0.75 | | | | | | | | |
| M.Pers | 0.51 | 0.65 | -0.18 | 0.54 | 0.24 | 0.38 | 0.66 | | | | | | | |
| F.Pers | -0.63 | -0.30 | 0.71 | -0.67 | -0.31 | -0.46 | -0.23 | 0.80 | | | | | | |
| M.Cog | 0.48 | 0.57 | -0.21 | 0.35 | 0.51 | 0.32 | 0.37 | -0.19 | 0.73 | | | | | |
| F.Cog | -0.33 | -0.03 | 0.49 | -0.16 | -0.48 | -0.19 | 0.03 | 0.26 | -0.02 | 0.70 | | | | |
| M.Act | 0.52 | 0.62 | -0.23 | 0.41 | 0.22 | 0.59 | 0.41 | -0.24 | 0.24 | -0.08 | 0.75 | | | |
| F.Act | -0.56 | -0.25 | 0.65 | -0.42 | -0.33 | -0.59 | -0.17 | 0.46 | -0.24 | 0.22 | -0.17 | 0.75 | | |
| Best10 | 0.63 | 0.39 | -0.62 | 0.46 | 0.36 | 0.66 | 0.27 | -0.44 | 0.32 | -0.19 | 0.30 | -0.73 | 0.74 | |
| gender | 0.61 | 0.43 | -0.55 | 0.43 | 0.35 | 0.66 | 0.27 | -0.39 | 0.31 | -0.19 | 0.38 | -0.62 | 0.69 | 1.00 |
| $\alpha$ | 0.85 | 0.81 | 0.83 | 0.77 | 0.67 | 0.75 | 0.66 | 0.80 | 0.73 | 0.70 | 0.75 | 0.75 | 0.74 | |
| $\omega_h$ | 0.26 | 0.25 | 0.23 | 0.09 | 0.05 | 0.09 | 0.01 | 0.25 | 0.30 | 0.42 | 0.48 | 0.48 | 0.46 | |
| $\omega_t$ | 0.86 | 0.83 | 0.85 | 0.81 | 0.74 | 0.79 | 0.65 | 0.84 | 0.84 | 0.76 | 0.78 | 0.79 | 0.77 | |
| n.items | 50 | 25 | 25 | 20 | 14 | 16 | 10 | 10 | 7 | 7 | 8 | 8 | 10 | 1 |

Table 9: Predicting Gender by using one, two, three, or six composite scales, all 50 items as well as the 10 best items. Data from the Gender Related Attributes Survey (Gruber et al., 2020). The derivation and cross validation samples were based upon multiple regression using 100 bootstrap resamplings of the original data with the same model applied to each hold out sample. Although the derivation samples regressions improve with the number of predictors (1-50), this leads to overfitting, particularly in the case of using 50 items not formed into scales (Compare the derivation and cross validation columns). The first five models are based upon using all 50 items, aggregated into scales of different lengths.The last model is the result of the `bestScales` solution, also bootstrapped 100 times. Individual scale lengths are shown in Table 8. For item content, see the help pages for `GERAS`.

100 bootstrap cross validations

| Model # predictors | N items used | Scales used | Derivation | Cross validation |
|---|---|---|---|---|
| 1 | 50 | MF.all | 0.64 | 0.62 |
| 2 | 50 | M + F | 0.64 | 0.64 |
| 3 | 50 | Pers + Cog + Act | 0.67 | 0.67 |
| 6 | 50 | M.pers + F.pers + M.Cog + F.Cog + M.act + F.act | 0.70 | 0.69 |
| 50 | 50 | All 50 items | 0.78 | 0.64 |
| 10 | 10 | 10 best items | 0.69 | 0.69 |

the average within scale correlation, $\bar{r_x}$ :

$$r_{yx_k} = \frac{k\bar{r_y}}{\sqrt{k + k(k-1)\bar{r_x}}}. \tag{7}$$

Thus, although internal consistency helps make a scale understandable from a construct validity point of view, we see from Equation 7 that it reduces the predictive validity of a scale.

In a comparison of empirical and homogenous keying (e.g. factor analytic) procedures, Goldberg (1972) concluded that emphasizing homogeneous keys is probably better for easily predicted criteria but that empirical keying was superior for harder to predict criteria. With the advent of larger samples, the advantageous of empirical keying over the traditional "Big 5" constructs becomes much more obvious (Revelle et al., 2021).

## Summary and Conclusions

Modern scale construction procedures owe a great deal to the developments over the past 100 years. Probably the most important step in scale construction is deciding what items are most useful to measure the constructs at hand. Giving these items to the target population and then choosing the best items using either factorially homogenous keying techniques (unsupervised learning) or straight empirical correlations with a criterion (supervised learning) are the next steps in the process. Subsequent validation of the scales given patterns of correlations with other measures (construct validity) or cross validated

correlations with criterion is essential. For the readers convenience, we summarized these steps in Box 1.

## References

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, *47*(211). doi: 10.1037/h0093360

Allport, G. W., & Vernon, P. E. (1933). *Studies in expressive movement.* New York, N.Y.: Macmillan.

Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods*, *52*(6), 2489–2505. doi: 10.3758/s13428-020-01401-8

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*(4), 309-318. doi: 10.1111/1471-6402.00111

Bernaards, C., & Jennrich, R. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement*, *65*(5), 676-696. doi: 10.1177/0013164404272507

Bernreuter, R. (1931). *Bernreuter personality inventory.* Stanford University Press.

Binet, A., & Simon, T. (1905). New methods for the diagnosis of the intellectual level of subnormals. *L'annee Psychologique*, *12*, 191–244 Binet, 1905 A. Binet, New methods for the diagnosis of the intellectual level of subnormals, L'Annee Psychologique 12 (1905), pp. 191–244 (Translated in 1916 by E. S. Kite in The Development of Intelligence in Children. Vineland, NJ: Publications of the Training School at Vineland.).

Binet, A., & Simon, T. (1916). *The development of intelligence in children: Translated by Elizabeth S. Kite* (H. H. Goddard, Ed.). Baltimore, Md.: William and Wilkens Company.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061-1071. doi: 10.1037/0033-295X.111.4.1061

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*(3), 296-322. doi: 10.1111/j.2044-8295.1910.tb00207.x

Campbell, D. P., & Borgen, F. H. (1999). Holland's theory and the development of interest inventories. *Journal of Vocational Behavior*, *55*(1), 86-101. doi: 10.1006/jvbe.1999.1699

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(8), 81-105. doi: 10.1037/h0046016

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309 - 319. doi: 10.1037/1040-3590.7.3.309

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological assessment*, *31*(12), 1412-1427. doi: 10.1037/pas0000626

Condon, D. M. (2018). *The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model*. PsyArXiv /sc4p9/. 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, June). $RetestReliability = f(Stability, Memory, Personality) + \epsilon$. (Presented at symposium in honor of Sarah Dubrow)

Condon, D. M., & Möttus, R. (2021). A role for information theory in personality modeling, assessment, and judgment. In D. Wood, S. Read, & P. D. H. . A. Slaughter (Eds.), *Measuring and modeling persons and situations* (p. 1-31). doi: 10.1016/B978-0-12-819200-9.00018-1

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

Condon, D. M., & Revelle, W. (2015). Selected personality data from the SAPA-Project: 08dec2013 to 26jul2014. *Harvard Dataverse*. doi: 10.7910/DVN/SD7SVE

Condon, D. M., Roney, E., & Revelle, W. (2017a). Selected personality data from the sapa-project: 22dec2015 to 07feb2017. [48,350 participant data file and codebook]. *Harvard Dataverse*. doi: 10.7910/DVN/TZJGAT

Condon, D. M., Roney, E., & Revelle, W. (2017b). Selected personality data from the sapa-project: 26jul2014 to 22dec2015. [54,855 participant data file and codebook]. *Harvard Dataverse*. doi: 10.7910/DVN/GU70EV

Condon, D. M., Wood, D., Möttus, R., Booth, T., Costani, G., Greiff, S., ... Zimmerman, J. (2021). Bottom Up Construction of a Personality Taxonomy. *European Journal of Psychological Assessment*. doi: 10.1027/1015-5759/a000626

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302. doi: 10.1037/h0040957

Cureton, E. E. (1950). Validity, reliability, and baloney. *Educational and Psychological Measurement*, *10*(1), 94–96. doi: 10.1177/001316445001000107

Dawis, R. V. (1992). The individual differences tradition in counseling psychology. *Journal of counseling psychology*, *39*(1), 7-19. doi: 10.1037/0022-0167.39.1.7

Del Giudice, M. (2021). Individual and group differeneces in multivariate domains: what happens with the number of traits increases? *PsyArXiv*. doi: 10.31234/osf.io/rgzd2

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*, *0*(0). 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*(6), 948-958. doi: /10.1027/1015-5759/a00059

Embretson, S. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher*, *36*(8), 449–455. doi: 10.3102/0013189X07311600

Eysenck, H. J., & Eysenck, S. B. G. (1964). *Eysenck Personality Inventory.* San Diego, California: Educational and Industrial Testing Service.

Galton, F. (1865). Hereditary talent and character. *Macmillan's Magazine*, *12*, 157-166.

Galton, F. (1884). Measurement of character. *Fortnightly Review*, *36*, 179-185.

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 I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, p. 7-28). Tilburg, The Netherlands: Tilburg University Press.

Goldberg, L. R. (2008). *The Eugene-Springfield Community Sample: Information available from the research participants* (Technical Report No. 48-1). Eugene, Oregon: Oregon Research Institute.

Goldberg, L. R. (2010). Personality, demographics and self reported acts: the development of avocational interest scales from estimates of the mount time spent in interest related activities. In C. Agnew, D. Carlston, W. Graziano, & J. Kelly (Eds.), *Then a miracle occurs: Focusing on the behavior in social psychological theory and research* (p. 205-226). New York, NY: Oxford University Press.

Goldberg, L. R., & Kilkowski, J. M. (1985). The prediction of semantic consistency in self-descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of personality and social psychology*, *48*(1), 82-98. doi: 10.1037//0022-3514.48.1.82

Goldberg, L. R., & Saucier, G. (2016). *The Eugene-Springfield Community Sample: Information available from the research participants* (Technical Report No. 56-1). Eugene, Oregon: Oregon Research Institute.

Graziano, W. G., Jensen-Campbell, L. A., Steele, R. G., & Hair, E. C. (1998). Unknown words in self-reported personality: Lethargic and provincial in Texas. *Personality and Social Psychology Bulletin*, *24*(8), 893–905. doi: 10.1177/0146167298248008

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*(4), 612-623. doi: 10.1027/1015-5759/a000528

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282. doi: 10.1007/BF02288892

Hathaway, S., & McKinley, J. (1943). *Manual for administering and scoring the MMPI.* Minneapolis: University of Minnesota Press.

Holzinger, K., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54. doi: 10.1007/BF02287965

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. doi: 10.1007/BF02289447

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103-129. doi: 10.1016/j.jrp.2004.09.009

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*(140), 1-53.

Likert, R., Roslow, S., & Murphy, G. (1934). A simple and reliable method of scoring the Thurstone attitude scales. *The Journal of Social Psychology*, *5*(2), 228–238. doi: 10.1080/00224545.1934.9919450

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

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

McNemar, Q. (1946). Opinion-attitude methodology. *Psychological Bulletin*, *43*(4), 289-374. doi: 10.1037/h0060985

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological methods*, *17*(3), 437-455. doi: 10.1037/a0028085

Möttus, R., Wood, D., Condon, D. M., Back, M., Baumert, A., Costani, G., ... 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). doi: 10.1002/per.2311

Nichols, D. S., & Greene, R. L. (1997). Dimensions of deception in personality assessment: The example of the MMPI-2. *Journal of Personality Assessment*, *68*(2), 251–266. doi: 10.1207/s15327752jpa6802_3

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667-696. doi: 10.1080/00273171.2012.715555

Reise, S. P., Morizot, J., & Hays, R. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*(0), 19-31. doi: 10.1007/s11136-007-9183-7

Revelle, W. (1979). Hierarchical cluster-analysis and the internal structure of tests. *Multivariate Behavioral Research*, *14*(1), 57-74. doi: 10.1207/s15327906mbr1401_4

Revelle, W. (2022a). psych: Procedures for psychological, psychometric, and personality research (2.2.9 ed.) [Computer software manual]. psych. (R package version 2.2.9)

Revelle, W. (2022b). psychTools tools to accompany the psych package for psychological research [Computer software manual]. psychTools. (R package version 2.2.9)

Revelle, W., & Anderson, K. J. (1998). *Personality, motivation and cognitive performance: Final report to the army research institute on contract MDA 903-93-K-0008* (Tech. Rep.). Evanston, Illinois, USA.: Northwestern University.

Revelle, W., & Condon, D. M. (2019). Reliability from $\alpha$ to $\omega$: A tutorial. *Psychological Assessment.*, *31*(12), 1395-1411. doi: 10.1037/pas0000754

Revelle, W., Condon, D. M., Wilt, J., French, J. A., Brown, A., & Elleman, L. G. (2017). Web and phone based data collection using planned missing designs. In N. G. Fielding, R. M. Lee, & G. Blank (Eds.), *Sage handbook of online research methods* (2nd ed., p. 578-595). Sage Publications, Inc.

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

Reyes, D. L. (2020). Combatting carelessness: Can placement of quality check items help reduce careless responses? *Current Psychology*. doi: 10.1007/s12144-020-01183-4

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and social psychology bulletin*, *27*(2), 151–161. doi: 10.1177/0146167201272002

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, *42*(1), 59–66. doi: 10.2307/2685263

Sartori, R., & Pasini, M. (2007). Quality and quantity in test validity: How can we be sure that psychological tests measure what they have to? *Quality & Quantity*, *41*(3), 359–374. doi: 10.1007/s11135-006-9006-x

Schmid, J. J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 83-90. doi: 10.1007/BF02289209

Schwaba, T., Rhemtulla, M., Hopwood, C. J., & Bleidorn, W. (2020, 07). A facet atlas: Visualizing networks that describe the blends, cores, and peripheries of personality structure. *PLOS ONE*, *15*(7), 1-21. doi: 10.1371/journal.pone.0236893

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? psychometric perspectives using personality questionnaire data. *Psychological Assessment*, *31*(4), 557–566. doi: 10.1037/pas0000648

Spearman, C. (1904a). "General Intelligence," objectively determined and measured. *American Journal of Psychology*, *15*(2), 201-292. doi: 10.2307/1412107

Spearman, C. (1904b). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72-101. doi: 10.2307/1412159

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271-295. doi: 10.1111/j.2044-8295.1910.tb00206.x

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

Thayer, R. E. (1989). *The biopsychology of mood and arousal*. The biopsychology of mood and arousal. xi, 234 pp. New York, NY: Oxford University Press.

Ward, M., & Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Applied Psychology*, *67*(2), 231-263. doi: https://doi.org/10.1111/apps.12118

Widaman, K. F., & Revelle, W. (2022). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*. doi: 10.3758/s13428-022-01849-w

Woods, S. A., & Hampson, S. E. (2005). Measuring the big five with single items using a bipolar response scale. *European Journal of Personality*, *19*(5), 373-390. doi: 10.1002/per.542

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122. doi: 10.1177/1745691617693393

Zhang, X., & Savalei, V. (2016). Improving the factor structure of psychological scales: The expanded format as an alternative to the likert scale format. *Educational and Psychological Measurement*, *76*(3), 357-386. doi: 10.1177/0013164415596421

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's $\alpha$, Revelle's $\beta$, and McDonald's $\omega_H$: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123-133. doi: 10.1007/s11336-003-0974-7

Zola, A., Condon, D. M., & Revelle, W. (2021, 08). The Convergence of Self and Informant Reports in a Large Online Sample. *Collabra: Psychology*, *7*(1). (25983) doi: 10.1525/collabra.25983

## Appendix

Here we show some limited commands using R to do the various operations discussed in our chapter. The assumption is that the most recent release of R (R Core Team, 2022) has been installed and that two *packages* have also been installed from CRAN. The *psych* package (Revelle, 2022a) was tailored made for scale construction and for basic psychometrics, the *psychTools* (Revelle, 2022b) includes many example data sets. Both packages come with *Vignettes* which go into much more detail about the specific uses. All functions in R packages downloaded from CRAN (the comprehensive R archive) include help pages which can be consulted as well.

R code to generate tables

R code

```
#first make sure that psych and psychTools are active
library(psych)
library(psychTools)

#show  items associated with various inventories discussed in the chapter
#for each set, examine the help pages and then look at the items

# The msqR contains various state emotion terms
?msqR  #show the help file for the msqR data set for example items

# The spi (SAPA personality Inventory) has 135 personality items with
# Five large factors and 27 lower level factors
?spi     #show the help file for the spi items
lookupFromKeys(spi.keys, spi.dictionary)  #show the spi items  organized by scales

#The Athenstatdt data set asks about gender related activities
?Athenstaedt  #help file for the Athenstaedt data
lookupFromKeys(Athenstaedt.keys, Athenstaedt.dictionary[2])  # show the items by keys

#The GERAS is another set of gender related activiites
?GERAS
lookupFromKeys(GERAS.keys,GERAS.dictionary[4])

#Table 2 shows the correlations of 6 variables from Athenstaedt
mf <- cs(V46,V45,V72,V32,V29,V54,gender)  #specify the variables
R <- mixedCor(Athenstaedt[mf])$rho    #find the polychoric and polyserial correlations

#create scoring keys for three scales.  Specify reverse keying with a - sign
 mf3.keys <- list(F=cs(V46,V45,V72), M = cs(V32,V29,V54),
    MF= cs(V46,V45,V72,   -V32,-V29,-V54), gender="gender")

sc3 <- scoreOverlap(mf3.keys,R)  #find the appropriate statistics

labels=c("Sew on Button","Make Bed", "Do ironing","Do Repairs","Change Fuses",
           "Shovel Snow","Gender")
rownames(R) <- colnames(R) <- labels
lowerMat(R)  #print the lower off diagonal with pretty spacing
lowerMat(sc3$MIMS)  #show the MIMS  (Mulit-Item Multi-Scale) matrix from the scores
```

```
#Show heat maps of the SPI N items and the MF items from Athenstaedt

anx <- selectFromKeys(spi.keys$Neuro)  #the spi$Neuro scale is 14 items
anx <- anx[-c(1,7,8,13)]  #cut it down to 10 7 pos 3 neg
anx[8] <- anx[7]     #rearrange for pretty output
anx[7] <- "q_578"

lab <- lookupFromKeys(list(anx),spi.dictionary)    #get the labels
lab1 <- lab[[1]][2]
 levels(lab1$item)[144] <- "Nervous person"   #abbreviate for pretty figure
 levels(lab1$item)[55] <- "Hopeless"
 lab.anx <- lab1$item

  R.anx <- polychoric(spi[anx])$rho    #find the polychoric correations and show them
  corPlot(R.anx,labels=lab.anx,xlas=3,
      main="A) Polychoric correlations of 10 items from the SPI Neuroticism")

#now do it for the Athenstaedt 10
  mf10 <- selectFromKeys(Athenstaedt.keys$MF10)
 mf.lab <- lookupFromKeys(list(Athenstaedt.keys$MF10),Athenstaedt.dictionary)
 mf.lab1 <- mf.lab[[1]][2]
 mf.lab <- mf.lab1$Item
 R <- polychoric(Athenstaedt[mf10])$rho
 corPlot(R, labels=mf.lab, xlas =3,
      main="B) Polychoric correlations for 10 items from  Athenstaed")
```

*Dimension reduction with EFA/PCA/clustering*

R code

```
#Do a factor analysis and a PCA of the mf items
f2 <- fa(R,2)    #use fa function and ask for 2 factors
fa.lookup(f2, dictionary = Athenstaedt.dictionary)  # show it with labels
pc2 <- pca(R,2)
#show the pca without variable names

#combine into 1 data frame for pretty output
temp <- fa.lookup(f2,dictionary=Athenstaedt.dictionary[2])
temp.pc <- fa.lookup(pc2,dictionary=Athenstaedt.dictionary[2])
 df.10 <- data.frame(temp[1:3],temp.pc[c(1:3,5)])
df2latex(df.10,rowlabels=FALSE)   #For those who use Latex

temp.r.poly<- resid(f2)
temp.pc.poly <- resid(pc2)
residuals <- lowerUpper(temp.r.poly,temp.pc.poly)
corPlot(residuals,labels=mf.lab ,xlas=3,
   main="Residual correlations from factor analysis and PCA")

#Omega analysis to show the hierarchical and bifactor structure

om <- omega(R.anx ,2)  #just extract two factors
fa.lookup(om,dictionary = spi.dictionary[2])   #show the abbreviated results
omega.diagram(om, labels=lab.anx)   #draw the figures -default is show the SL  figure
omega.diagram(om, labels = lab.anx, sl=FALSE)   #don't draw the SL figure
```

```
# Yet another alternative is hierarchical cluster analysis of items (iclust)

ic.anx <- iclust(R.anx)
iclust.diagram(ic.anx, labels = lab.anx)

ic.mf <- iclust(R)
iclust.diagram(ic.mf, labels = mf.lab)
```

*Reliability measures*

Once we have identified our scales, we can find various estimates of reliability. We can do this for multiple scales from the same data set.

R code

```
#Finding reliability for a number of scales at one time
#We use the example of the epi.keys and the epiR (for repeated measures)

reliability(epi.keys,epiR)

epi.keys  #show the keying information
#The retest values are a bit more tedious and require one scale at time
E.retest <- testRetest(epiR, keys = epi.keys$E) #repeat this for each key



# Finding scales to find measures of construct validity, the Zola example
?zola

#the correlation of 135 spi items and 30 peer reports  are in the zola data set
#We combione the item level correlations to find higher level scale correlations
lookupFromKeys(zola.keys,zola.dictionary)  #show all the items for all the scales
scores <- psych::scoreOverlap(zola.keys[c(1:5,33:37)],zola)
scores   #shows the complete output
lowerMat(t(scores$corrected))   #we transpose it to get the corrected correlations
```

*Scoring scales*

Empirically based keys can be found using the `bestScales` function. These keys can then be combined with theory based keys and scored together using `scoreitems`. The `scoreOverlap` function will not find scale scores, but will find the correlations of the scales and report useful statistics.

R code

```
#The bestScale function applied to the GERAS data set
bs <- bestScales(GERAS.items, "gender", dictionary=GERAS.dictionary[4], folds=10)
bs   #show the summary statistics
```

```
bs$best.keys   #show the best keys

#combine these best keys with the normal GERAS.keys
#rearrange keys to match the table
geras.keys <- GERAS.keys
geras.key <- list(MF.all = geras.keys$MF.all, M= geras.keys$M,  F=geras.keys$F,
    Pers=geras.keys$Pers, Cog=geras.keys$Cog,Act=geras.keys$Act,
   M.Pers = geras.keys$M.pers,F.Pers = geras.keys$F.pers,
   M.Cog = geras.keys$M.cog,F.Cog = geras.keys$F.cog,
 M.Act = geras.keys$M.act,F.Act = geras.keys$F.act,
  Best10 = bs$best.keys, gender= geras.keys$gender)

#find the actual scores
scales <- scoreItems(geras.key, GERAS.items)   #We score at the data level
scales   #show the summary statistics
scores <- scales$scores   #scores are an object in the scales output

#or find other statistics
sc <- scoreOverlap(geras.key,GERAS.items)
sc
reliability(geras.key ,GERAS.items)
```

Boot strap resampling allows us to cross validate our prediction models

R code

```
#Now, do the cross validations of these various scales
#First find all the scale scores
scales <- scoreItems(geras.key, GERAS.items)
scores <-  scales$scores
mod1 <- crossValidationBoot(gender ~ MF.all, data = scores)
mod2 <-   crossValidationBoot(gender ~ M + F, data = scores)
mod3 <-   crossValidationBoot(gender ~ Pers + Cog + Act , data = scores)
mod6<-   crossValidationBoot(gender ~ M.Pers + M.Cog +  M.Act + F.Pers + F.Cog +  F.Act
              , data = scores)
mod50 <- crossValidationBoot( y=51, x =1:50, data = GERAS.items)
mod10  <- crossValidationBoot(gender ~ Best10, data=scores)
```