Statistical analyses and computer programming in personality
A chapter for the Cambridge University Press Handbook of Personality Psychology

William Revelle, Lorien Elleman and Andrew Hall
Department of Psychology, Northwestern University

Abstract

The use of open source software for statistical analysis provides personality researchers new tools for exploring large scale data sets and allows developing and testing new psychological theories. We review the development of these techniques and consider some of the major data analytic strategies. We provide example code for doing these analyses in R.

---

contact: William Revelle revelle@northwestern.edu
We greatly appreciate the thoughtful comments from David Condon. This is the final version as submitted to CUP.
The official version is at The Cambridge Handbook of Personality Psychology, pp. 495-534.
DOI: https://doi.org/10.1017/9781108264822.045. Published 18 September 2020.

Prologue: A brief history of Open Source Statistical Software

It is hard to imagine early in the 21st century that many of the statistical techniques we think of as modern statistics were developed in the late 19th and early 20th centuries. What we now call regression was the degree of "reversion to mediocrity" as introduced by Francis Galton (1886). A refinement of regression was the measure of "co-relations" (Galton, 1888) which was specified as the relationship of deviations as expressed in units of the probable error. Galton's insight of standardization of units became known as 'Galton's coefficient' which, when elaborated by Pearson (1895,8,9) became what we now know as the Pearson Product Moment Coefficient. Because the Pearson derivations were seen as too complicated for psychologists, Charles Spearman (1904b) explained the developments of Pearson and his associates to psychologists where "the meaning and working of the various formulae have been explained sufficiently, it is hoped, to render them readily usable even by those whose knowledge of mathematics is elementary" (p 73 Spearman, 1904b). Later in that same paper, he then developed reliability theory and the correction for attenuation. In his second amazing publication that year he developed the basic principles of factor analysis and laid out his theory of general intelligence (Spearman, 1904a).

Fundamental theorems of factor analysis (Thurstone, 1933; Eckart and Young, 1936) and estimates of reliability (Brown, 1910; Spearman, 1910; Kuder and Richardson, 1937; Guttman, 1945) came soon after and were all developed when computation was done by "computers" who were humans operating desk calculators. If not difficult, computation was tedious in that it required repeatedly finding sums of squares and sums of cross products. When examining the correlation structure of multiple items, the number of correlations went up by the square of the number of items, and thus so did the computational load. Test theory as developed in the 1930s and 1940s led to a number of convenient short cuts by which the sums of test and item variances could be used to estimate what the correlations of composites of items would be if certain assumptions were met. Thus, the coefficients of Cronbach (1951); Guttman (1945); Kuder and Richardson (1937) (α, λ_3 and KR20) were meant to be estimates based upon the structure of covariances without actually finding the covariances.

Another short cut was to dichotomize continuous data in order to find correlations. For example, in an ambitious factor analysis of individual differences in behavior among psychiatric patients (Eysenck, 1944) made use of this shortcut by dichotomizing continuous variables and then finding the Yule (1912) coefficient. Unfortunately, such a procedure produced a non-positive definite matrix which makes reanalysis somewhat complicated.

With the need to calculate flight paths for large artillery shells and rockets, the ideas developed by Babbage for his "Analytical Engine" in 1838 (Bromley, 1982) and algorithms for programming (Lovelace, 1842) were converted from punched cards and automatic looms into the electrical tubes and circuit boards of von Neuman (Isaacson, 2014). The age of computation had arrived. It was now possible to properly analyze covariance structures.

For personality psychologists, these were exciting times, for it allowed for the cal-

culation of larger correlation matrices and the use of factor analysis with more elegant algorithms than the centroid approach. Indeed, Raymond Cattell moved from Harvard to the University of Illiinois because he was promised access to the new 'ILLIAC' computer (and would not have to teach undergraduates).

Main frame computers and proprietary software

At first, software for these new devices was tailor made for the particular machine, and there was a plethora of programming languages and operating systems. In the late 1950s the programming language FORTRAN (later renamed to be Fortran when computers learned about lower case) was developed at IBM for the numeric computation necessary for scientific computing and then translated for other operating systems. While some programs would run on IBM 709 and 7090s, others would only work on the super computers of the time, the Control Data Corporation's 1604 and 6400. An early package of statistical programs, developed for the IBM 7090 in 1961 at UCLA was the BioMedical Package (BMDP). At first BMDP was distributed for free to other universities but BMDP (Dixon and Brown, 1979) eventually became a commercial package which has since disappeared. Two other statistical systems (SAS® and SPSS), originally developed at other universities (North Carolina State and Stanford) and shared with others, were also developed.

These three major software systems for doing statistics came from three somewhat different traditions. BMDP was developed for biomedical research, SAS® for agriculture research, and SPSS for statistics in the social sciences (SPSS, 2008). Although all three systems were originally developed at universities and were freely distributed to colleagues, all three soon became incorporated as for profit corporations. All were developed for main frame computers where instructions were originally given in stacks of Hollerith Cards (known to many as IBM cards), containing 80 characters per card. All of the programs made use of the FORTRAN programming language and still, many years later, have some of their main frame geneaology embedded in their systems. Older researchers still shudder at the memories of needing to wait for 24 hours after turning in a box of cards only to discover one typo negated the entire exercise.

S and R: interactive statistics

In contrast to the statistical analyses done on mainframes, S and subsequently R were developed for interactive statistics. The S computing 'environment' was developed for Unix in the 1970's at Bell labs by John Chambers and his associates (Becker, Chambers, and Wilks, 1988). It was meant to take advantage of interactive computing where the user could work with his/her data to better display it and understand it. After several iterations it became the defacto statistical package for those using the Unix operating system. In 1992, two statisticians, Ross Ihaka and Robert Gentleman, at the University of Otago, in New Zealand started adapting S to run on their Mac computers. It incorporated the list oriented language Scheme and emphasized object-oriented programming. Most importantly, they

shared the design specifications with other interested developers around the world and intentionally did not copyright the code. Rather, they, with the help of John Chambers and the rest of the R Development Core Team, deliberately licensed R under the GNU General Public License of the Free Software Foundation which allows users to copy, use, and modify the code as long as the product remains under the GPL (R Core Team, 2018).

Perhaps the real power of R is that because it is open source, it is extensible. That means that anyone can contribute packages to the overall system. That, and the power of the GPL and open source software movement has led to an amazing effect. From the original functions in R and the ones written by the R Core Team, more than 12,600 packages have been contributed to the CRAN, the Comprehensive R Archive Network, and at least (as of this writing) more than 34,000 packages are available on GitHub. R has become the lingua franca of statistics and many new developments in psychological statistics are released as R functions or packages. When writing methodology chapters or articles, the associated R code to do the operations may also be given (as we do in this chapter).

With the growing recognition of the importance of replicable research, the publication of the R scripts and functions to do the analysis, as well as the release of R readable data sets is an essential step in allowing others to understand and repeat what we have done. Because the source code of all of the R packages is available, users can check the accuracy of the code and report bugs to the developers of the particular package. This is the ultimate peer review system, in that users use, review, and add to the entire code.

Given its open source nature and growing importance in scientific computing, much of the rest of this chapter will be devoted to discussing how particular analyses can be done in R. This is not to deny that commercial packages exist, but to encourage the readers of this handbook to adopt modern statistics. The actual code used for the tables and figures is included in the Appendix.

Finally, little appreciated by many users of R is that it is not just a statistical environment, it is also a very high level programing language. Although some of the packages in R are written in Fortran or C++, many packages are written in R itself. R allows operations at the higher matrix level and allows for object-oriented programming. Each function operates on 'objects' and returns another 'object'. That is, functions can be chained together to add value to previous operations. This allows users to include standard functions in their own functions with the output available for even more functions. Actual programming in R is beyond the scope of this chapter, but is worth learning for the serious quantitative researcher. Without developing packages, a willingness to write more and more complicated scripts is a positive benefit.

Getting and using R

R may be found at https://www.r-project.org and the current release is distributed through the Comprehensive R Archive Network https://cran.r-project.org. Popular interfaces to R include Rstudio (https://www.rstudio.com) which is particularly useful for PCs (the Mac version comes with its own quite adequate interface). Once R is

downloaded and installed, it is useful to install some of the powerful packages that have been added to it. We will make use of some of these packages, particularly the *psych* package which has been specifically developed for personality and psychological research (Revelle, 2018). See the appendix for detailed instructions.

Data, Models, and Residuals

The basic equation in statistics is that:

$$Data = Model + Residual \iff Residual = Data - Model. \tag{1}$$

That is, the study of statistics is the process of modeling our data. Our models are approximations and simplifications of the data (Rodgers, 2010). Our challenge as researchers is to find models that are good approximations of the data but that are not overly complicated. There is a tradeoff between the two goals of providing simple descriptions of our data and providing accurate descriptions of the data. Consider the model that the sun rises in the East. This is a good model on average and as a first approximation, but is actually correct only twice a year (the equinoxes). A more precise model will consider seasonal variation and recognize that in the northern hemisphere, the sun rises progressively further north of east from the spring equinox until the summer solstice and then becomes more easterly until the fall equinox. An even more precise model will consider the latitude of the observer.

If we think of degrees of freedom in models as money, we want to be frugal but not stingy. Typically we evaluate the quality of our models in terms of some measure of *goodness of fit*. Conceptually, fit is some function of the size of the residuals as contrasted to the data. Because almost all models will produce mean residuals of zero, we typically invoke a cost function such as ordinary least squares to try to find a model that minimizes our squared residual. As an example, the algebraic mean is that model of the data that minimizes the squared deviations around it (the variance).

The following pages will consider a number of statistical models, how to estimate them, and how to evaluate their fit. All of what follows can be derived from a serious consideration of Equation 1.

In what follows we discuss two types of variables and three kinds of relationships. In a distinction reminiscent of the prisoners in the cave discussed by Plato in the *The Republic* (Plato, 1892), we consider two classes of variables: those which we can observe, and those which we can not observe but are the latent causes of the observed variables. Our observations are of the movement of the shadows on the cave's wall; we need to infer the latent causes of these shadows. Many of our tools of data analysis (e.g., factor analysis, reliability theory, structural equation modeling, and item response theory) are just methods for estimating latent variables and their inter-relationships from the pattern of relationships between observed variables. Traditionally we make this distinction by using Greek letters for unobserved (latent) population values and Roman letters for observed

values. When we portray patterns of relationships graphically (e.g., Figure 1) we show observed variables as boxes and latent variables as circles. As may be seen in Figure 1, there are three kinds of relationships between our variables: relations between observed variables, relations between latent and observed variables, and relations between latent variables. Theories are organizations of latent variables as they represent the relationships between observed variables.

Basic Descriptive Statistics

Before any data analysis may be done, the data must be collected. This is more complicated than it seems, for it involves consideration of the latent variables of interest; presumed observed markers of these latent variables; the choice of subjects (are they selected randomly, systematically, are they volunteers, are they WEIRD (Henrich, Heine, and Norenzayan, 2010)), the means of data collection (self report, observer ratings, life narratives, computerized measurement, web based measures, etc.); the number of times measures are taken (e.g., once, twice for test-retest measures or measures of change, multiple times in studies of growth changes or of emotions over time); the lags between repeated measures (minutes, hours, days, or years) and whether there are experimental manipulations to distinguish particular conditions (Revelle, 2007).

Once the data are collected, it is of course necessary to prepare them for data analysis. That is to say, to transfer the original data into a form suitable for computer analysis. If hand coding must be done (e.g., scoring life narratives, Guo, Klevan, and McAdams, 2016) the separate ratings must be entered in a way that allows for computer based analysis (e.g., a reliability calculation) to be made.

Typically the data are organized as a two dimensional table (e.g., a spreadsheet in EXCEL or OpenOffice) with subjects as rows and variables as columns. If there are repeated measures per subject, the data might have a separate row for each occasion, but with one column identifying the subject and another the occasion. Consider the data in Table 1 which are taken from an example data set `msqR` in the *psych* package (Revelle, 2018) from R[1]. The `msqR` data set was collected over about 10 years as part of a long term series of studies of the interactive effect of personality and situational stressors on cognitive performance.

An under-appreciated part of data analysis is the basic data cleaning necessary to work with real data. Mistakes are made at data entry, participants fall asleep, other participants drop out, some do not answer every question, some participants are intentionally deceptive in their responses. It is important before doing any analysis to find basic descriptive statistics to look for impossible responses, to examine the distribution of responses, and to attempt to detect outliers. However, as discussed by Wilcox (2001), merely examining the shape of the distribution is not enough to detect outliers and it is useful to apply *robust estimators* of central tendency and relationships. The *WRS2* package (Mair, Schoenbrodt,

[1]In the following pages, we use boldfaced text for `functions` and *italics* for packages.

Figure 1. The basic set of statistical relationships may be seen as relations among and between observed variables (boxes) and latent variables (circles).

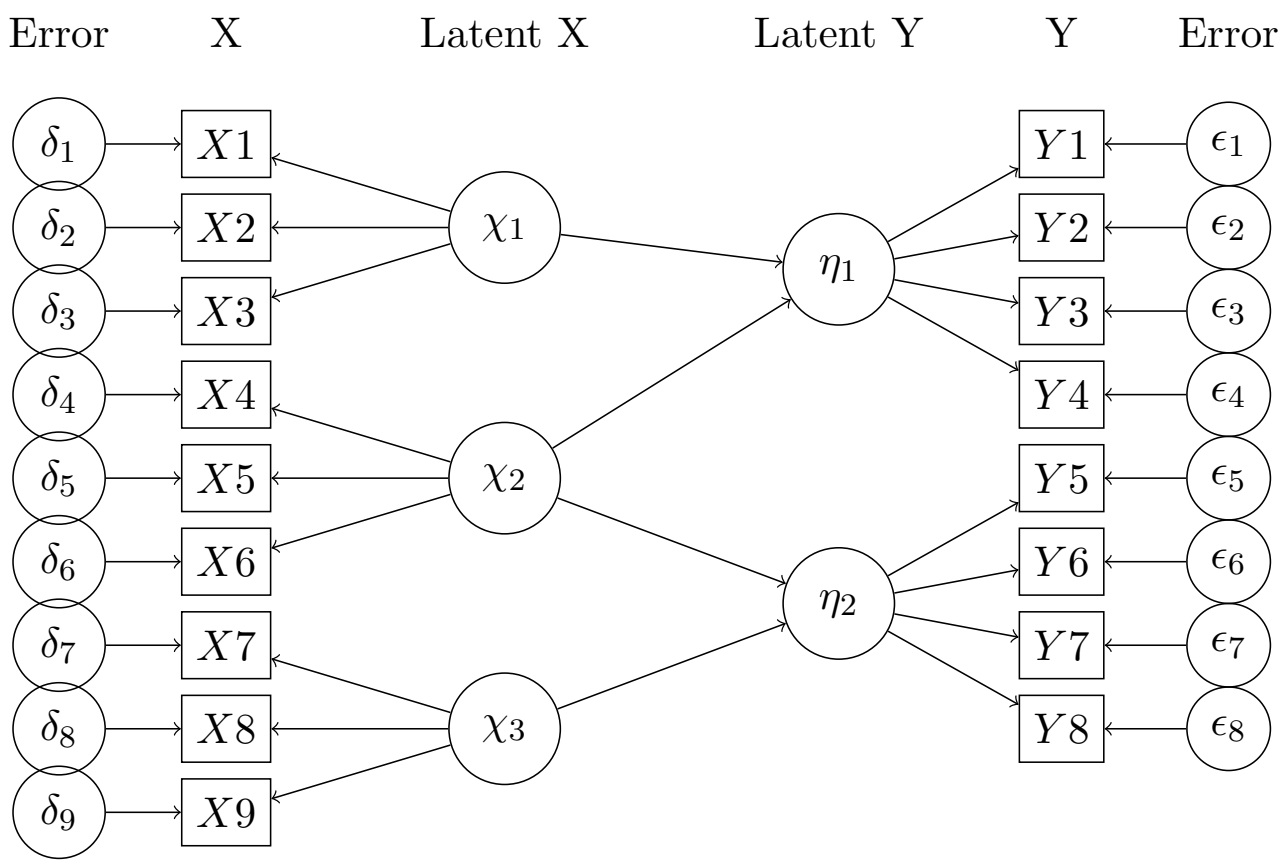

Table 1: A representative sample of eight subjects with time 1 and time 2 observations on 10 emotion terms. The data are from the `msqR` data set (N=3,032) which has repeated measures for 2,086 participants. The full data set is used for many of following examples. It is included in the *psych* package. The data are shown in 'long' format with repeated measures 'stacked' on top of each other to represent multiple time points. 'Wide' format would represent the different time points as separate columns for each subject.

Line #	id	time	anxis	at.es	calm	cnfdn	cntnt	jttry	nervs	relxd	tense	upset
1	1	1	1	2	2	2	2	1	0	2	0	0
2	2	1	1	2	2	1	2	0	1	2	1	1
3	3	1	2	2	2	2	2	0	1	2	1	0
4	4	1	0	2	2	2	3	0	0	2	0	0
5	5	1	0	3	3	2	2	1	0	2	0	0
6	6	1	1	3	2	3	3	0	0	3	0	0
7	7	1	0	1	1	1	1	0	0	2	0	0
8	8	1	0	2	3	1	2	0	0	2	0	0
69	1	2	1	2	2	2	2	1	0	1	0	0
70	2	2	1	2	2	1	2	1	1	2	1	1
71	3	2	1	2	1	2	2	0	1	2	1	0
72	4	2	1	1	0	2	3	1	1	1	1	0
73	5	2	0	2	3	2	1	0	0	2	0	0
74	6	2	1	2	2	3	3	1	0	3	0	0
75	7	2	0	1	1	1	1	1	0	1	0	0
76	8	2	0	2	2	1	2	0	0	2	0	0

and Wilcox, 2017) implements many of the robust statistics discussed by Wilcox and his colleagues (Wilcox and Keselman, 2003; Wilcox, 2005). For example, the algebraic mean is just the sum of the observations divided by the number of observations. The trimmed mean is the same after a percentage (e.g., 10%) are removed from the top and bottom of the distribution. The trimmed mean is more robust to outliers than is the algebraic mean. The median is the middle observation (the 50th percentile) and is an extreme example of a trimmed mean (with trim=.5). The minimum and maximum observations, and the resulting range are most useful for detecting improper observations. Skew and Kurtosis are functions of the third and fourth power of the deviations from the mean (Mardia, 1970). The `describe` function will also report various percentiles of the distribution including the Inter Quartile Range (25th to 75th percentiles) (Table 2).

Tests of statistical significance: Normal theory and the bootstrap

Those brought up in the Fisherian tradition of Null Hypothesis Significance Testing (NHST) traditionally compare fit statistics to their expected value given normal theory. Fits are converted into standardized scores (z scores) and then probabilities are found from the normal distribution. This works well with large samples where errors are in fact random. For smaller samples, the variation of estimates of mean differences compared to the sample based standard error are larger than expected given the normal. This problem led to the

Table 2: Descriptive statistics for the data in Table 1. An important step before doing any more advanced analysis is to search for outliers by examing the data for impossible values and comparing the values of the algebraic versus trimmed mean versus the median.

Variable	vars	n	mean	sd	medin	trmmd	mad	min	max	range	skew	kurtosis	se	IQR
id	1	16	4.50	2.37	4.5	4.50	2.97	1	8	7	0.00	-1.45	0.59	3.50
time	2	16	1.50	0.52	1.5	1.50	0.74	1	2	1	0.00	-2.12	0.13	1.00
anxious	3	16	0.62	0.62	1.0	0.57	0.74	0	2	2	0.35	-0.96	0.15	1.00
at.ease	4	16	1.94	0.57	2.0	1.93	0.00	1	3	2	-0.02	-0.19	0.14	0.00
calm	5	16	1.88	0.81	2.0	1.93	0.00	0	3	3	-0.51	-0.20	0.20	0.25
confident	6	16	1.75	0.68	2.0	1.71	0.74	1	3	2	0.29	-1.04	0.17	1.00
content	7	16	2.06	0.68	2.0	2.07	0.00	1	3	2	-0.06	-0.98	0.17	0.25
jittery	8	16	0.44	0.51	0.0	0.43	0.00	0	1	1	0.23	-2.07	0.13	1.00
nervous	9	16	0.31	0.48	0.0	0.29	0.00	0	1	1	0.73	-1.55	0.12	1.00
relaxed	10	16	1.94	0.57	2.0	1.93	0.00	1	3	2	-0.02	-0.19	0.14	0.00
tense	11	16	0.31	0.48	0.0	0.29	0.00	0	1	1	0.73	-1.55	0.12	1.00
upset	12	16	0.12	0.34	0.0	0.07	0.00	0	1	1	2.06	2.40	0.09	0.00

introduction of the t statistic for comparing the means of smaller groups (Student, 1908) (`t.test`) and the r to z transformation (`r2z`) for tests of correlations (Fisher, 1921). (Use `cor.test` for one correlation, `corr.test` for many correlations). Most functions in R will return both the statistic and the probability of that statistic. Many will also return a confidence interval for the statistic. But the probabilities (and therefore the confidence intervals) are a function of the *effect size*, the sample size, and the distribution of the parameter being estimated. Unfortunately, not all tests can be assumed to be normally distributed and it is unclear how to find the distribution of arbitrary parameters of a distribution (e.g., the median). Extending ideas such as the 'Jackknife' proposed by Tukey (1958), Efron (1979) proposed the 'bootstrap' (having considered such names as the 'shotgun'), a powerful use of random sampling (Efron and Gong, 1983).

The basic concept of the bootstrap is to treat the observed sample as the entire population, and then to sample repeatedly from this 'population' with replacement; find the desired estimate (e.g, the mean, the median, the regression weight) and then do this again, and again, and many times again. Each sample, although the same size as the original sample, will contain (on average) 63.2% of the subjects in the original sample, with 36.8% being repeated at least once[2] The resulting distribution of the estimated value can be used to find confidence intervals without any appeal to normal theory.

For those who use NHST, it is important to understand the probability that a real effect is detected, the power, is not the same as the probability of rejecting the 'nil' hypothesis that an effect is 0 (Cohen, 1988,9,9; Streiner, 2003). A number of R packages (e.g.

[2]This perhaps unintuitive amount is $1 - \frac{1}{e}$ and is the limit of the probability of an item not being repeated as the number of cases increases ($p = 1 - (1 - \frac{1}{n})^n$).

pwr, Champely, 2018) include easy to use power calculators to find the the power of a design given an effect size, sample size, and desired α level.

One of the great advances of modern statistics is the use of the bootstrap and other randomization tests. In the space of seconds, 1,000 to 100,000 bootstrap resamples can be calculated for almost any statistic. We will use this procedure when we find the confidence intervals for correlation coefficients (Table 5) and in particular for the effect of mediation in a regression model.

Correlation and Regression

Originally proposed by Galton (1888) and refined by Pearson (1895) and Spearman (1904b), the linear regression coefficient and its standardized version, the correlation coefficient are the fundamental statistics of research. In terms of deviation scores ($x = X - \bar{X}$ and $y = Y - \bar{Y}$)

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

Depending upon the characteristics of the data, the correlation as defined in Equation 2 has many different names. If both X and Y are continuous, then the resulting correlation is known as the Pearson Product Moment Correlation Coefficient (or just Pearson r). If converted to ranks, as Spearman's ρ, and if both X and Y are dichotomous as the ϕ coefficient (Table 3). Three of the correlations shown are estimates of what the latent continuous correlation would be if two continuous latent variables (χ and ψ) were artificially dichotomized (the tetrachoric), or split into multiple levels (the polychoric) correlation.

Because the first four of these correlations are just different forms of the Pearson r (albeit it in different forms), the same estimation function can be used. In core R this is just the `cor` function or to find covariances the `cov` function. The last three require specialized functions written for polytomous (or dichotomous) data (i.e., the *psych* package functions `polyserial`, `tetrachoric` and `polychoric`. All of these functions are combined in `mixed.cor`.)

The tetrachoric correlations of ability data (answers are right or wrong) and the polychoric correlations of self report temperament scales (typically on a 1-4, 1-5, or 1-6 scale), being the modeled correlation of continuous latent scores, will be larger in absolute value than the Pearson correlations of the same data. In addition, these estimates of the latent correlations are not affected by differences in distributions the way the Pearson r on the observed variables is. An example of the difference between a Tetracoric and a Pearson ϕ is seen in Table 7 where $\phi = .32$ but the inferred relationship between two continuous variables was .54.

The ubiquitous correlation coefficient

The correlation is also a convenient measure of the size of an effect (Ozer, 2007). It has long been been known that the difference in means compared to the within group

Table 3: A number of correlations are Pearson r in different forms, or with particular assumptions. The first four use $r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$, The last three are based upon assumptions of normality of a latent X and Y, with an artificial dichotomization or categorization into discrete (but ordered) groups.

Coefficient	symbol	X	Y	Assumptions
Pearson	r	continuous	continuous	
Spearman	rho (ρ)	ranks	ranks	
Point bi-serial	r_{pb}	dichotomous	continuous	
Phi	ϕ	dichotomous	dichotomous	
Bi-serial	r_{bis}	dichotomous	continuous	normality
Tetrachoric	r_{tet}	dichotomous	dichotomous	bivariate normality
Polychoric	r_{pc}	categorical	categorical	bivariate normality

standard deviation: the d statistic of Cohen (1962,9,9) is a better way to compare the difference between two groups than Student's t statistic. For it is the size of the difference that is important, not the significance. An undue reliance of "statistical significance" has ignored the basic observation that the test of significance = size of effect x size of study (Rosenthal, 1994) and that the resulting p value is a non-linear function of the size of the effect. To remedy this problem, Cohen (1962) developed the d statistic for the comparison of two groups (use `cohen.d`). Generalizations for multiple groups or continuous variables allow the translation of many alternative indices of effect size into units of the correlation coefficient (see Table 4). Robust alternatives to d (found by `d.robust`) express differences in terms of trimmed means and Winsorized variances (Algina, Keselman, and Penfield, 2005; Erceg-Hurn and Mirosevich, 2008). Basic principles in reporting effect sizes are available in a recent tutorial (Pek and Flora, 2018).

Multiple Regression and the general linear model

Just as the t-test and the F-test may be translated into correlations units, so they can be thought of in terms of the general linear model (Judd and McClelland, 1989):

$$\hat{\mathbf{Y}} = \mu + \beta \mathbf{X} + \epsilon. \tag{3}$$

$\mathbf{X}$ can be an experimental *design matrix* with one or more independent grouping variables but it can also include a set of person variables. In the case of just one dichotomous grouping variable, then Equation 3 is just the regression of the two levels of $\mathbf{X}$ with the dependent variable and is similar to the comparison of the means of Student's t (Student, 1908). t is typically expressed as difference of means compared to the standard error of that difference but is better expressed as an effect size multiplied by one half the square root of the degrees of freedom (df) or the ratio of the correlation times the square root of

Table 4: Alternative Estimates of effect size. Using the correlation as a scale free estimate of effect size allows for combining experimental and correlational data in a metric that is directly interpretable as the effect of a standardized unit change in x leads to r change in standardized y.

Statistic	Estimate	r equivalent	as a function of r
Pearson correlation	$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y}$	r_{xy}	
Regression	$b_{y.x} = \frac{Cxy}{\sigma_x^2}$	$r = b_{y.x}\frac{\sigma_y}{\sigma_x}$	$b_{y.x} = r\frac{\sigma_x}{\sigma_y}$
Cohen's d	$d = \frac{X_1 - X_2}{\sigma_x}$	$r = \frac{d}{\sqrt{d^2+4}}$	$d = \frac{2r}{\sqrt{1-r^2}}$
Hedge's g	$g = \frac{X_1 - X_2}{s_x}$	$r = \frac{g}{\sqrt{g^2+4(df/N)}}$	$g = \frac{2r\sqrt{df/N}}{\sqrt{1-r^2}}$
t - test	$t = \frac{d\sqrt{df}}{2}$	$r = \sqrt{t^2/(t^2+df)}$	$t = \sqrt{\frac{r^2 df}{1-r^2}}$
F-test	$F = \frac{d^2 df}{4}$	$r = \sqrt{F/(F+df)}$	$F = \frac{r^2 df}{1-r^2}$
Chi Square		$r = \sqrt{\chi^2/n}$	$\chi^2 = r^2 n$
Odds ratio	$d = \frac{ln(OR)}{1.81}$	$r = \frac{ln(OR)}{1.81\sqrt{(ln(OR)/1.81)^2+4}}$	$ln(OR) = \frac{3.62r}{\sqrt{1-r^2}}$
$r_{equivalent}$	r with probability p	$r = r_{equivalent}$	

the degrees of freedom to the coefficient of alienation ($\sqrt{1-r^2}$, Brogden, 1946).

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{d\sqrt{df}}{2} = \frac{r}{\sqrt{1-r^2}}\sqrt{df} \tag{4}$$

where the degrees of freedom are $n_1 + n_2 - 2$ (Rosnow and Rosenthal, 2003). The slope of the regression is the effect size, dividing this by the coefficient of alieniation and multiplying by the square root of df converts the regression to a t. It is found in R with the `t.test` function.

If $\mathbf{X}$ has two categorical grouping variables (e.g., x_1 and x_2), then we have

$$\hat{y} = \mu + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \tag{5}$$

which for categorical values of $\mathbf{X}$ is just the traditional analysis of variance of two main effects and an interaction (Fisher, 1925). This may be found using the `aov` function which acts on categorical variables and returns the traditional ANOVA output. With unbalanced repeated measures designs, the `lme` function included in the *lme4* package (Bates, Maechler, Bolker, and Walker, 2015) allows a specification of random and fixed effects.

The advantage of the general linear model for psychologists interested in individual differences is that continuous person variables can be included in the same model as experimental variables. This is a great improvement from prior approaches which would artificially dichotomize the person variable into high and low groups in order to use an

ANOVA approach. By retaining the continuous nature of the predictor, we improve the power over the ANOVA test.

As an example of using the general linear model, we use a data set from Tal-Or, Cohen, Tsfati, and Gunther (2010) that is discussed by Hayes (2013). Tal-Or et al. (2010) measured the effect of an experimental manipulation of salience of a news article (cond) on presumed media influence (PMI), perceived importance of the issue (import), and reported willingness to change one's behavior (reaction)[3]. The observed correlations are found by using the `lowerCor` function and are given in Table 5.

Table 5: Correlations of the conditions with Perceived Media Influence, Importance of the message, and Reaction to the message (Tal-Or et al., 2010). As is traditional in NHST, correlations that are larger than would be expected by chance are marked with 'magic astericks'. Confidence intervals for these correlations are shown given normal theory (upper and lower normal) as well as estimated by 1,000 bootstrap resamplings of the data (lower and upper empirical).

The Tal-Or et al. correlation matrix from `lowerCor`

Variable	cond	pmi	imprt	rectn
cond	1.00			
pmi	0.18*	1.00		
import	0.18*	0.28**	1.00	
reaction	0.16	0.45***	0.46***	1.00

Note: $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$.

Empirical and normal theory based confidence intervals from `cor.ci`.

Variable	lwr.m	lwr.n	estmt	uppr.n	uppr.m
cond-pmi	0.01	0.01	0.18	0.36	0.36
cond-imprt	0.02	0.00	0.18	0.35	0.35
cond-rectn	-0.01	-0.01	0.16	0.33	0.33
pmi-imprt	0.09	0.09	0.28	0.45	0.45
pmi-rectn	0.31	0.31	0.45	0.57	0.56
imprt-rectn	0.32	0.32	0.46	0.60	0.61

There are multiple ways to analyze these data. We could naively do three t-tests of the experimental manipulation, find all of the intercorrelations, or do a regression predicting reaction from the condition, perception of media influence (PMI) and perceived importance of the message (Importance). All of these alternatives are shown in the appendix. The `setCor` and `mediate` functions will also draw the regressions as path diagrams (Figure 3).

[3]With the kind permission of Nurit Tal-Or, Jonathan Cohen, Yariv Tsfati, and Albert C. Gunther, these data were added to the *psych* package as the `Tal_Or` data set.

Mediation and Moderation

In the Tal-Or et al. (2010) data set, the experimental manipulation affected the dependent variable of interest (reaction) but also two other variables (perception of media influence and perceived importance of the message). There is a direct effect of condition on reaction, as well as indirect effects through PMI and Importance. Conventional regression shows the direct effect of reaction controlling for the indirect effects that go through PMI and import. The *total effect* of condition on reaction is their covariance divided by the condition variance and is known as the c effect ($c = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{.125}{.25} = .5$). If we label the paths from cond to PMI ($a_1 = .48$) and from condition to import ($a_2 = .63$), then the *indirect effects* are the sum of the products through the two mediators ($b_1 = .40, b_2 = .32 \rightarrow a_1b_1 + a_2b_2 = .48 * .40 + .63 * .32 = .4$) then the *direct effect* is the total less the indirect effect (c' = c - ab = .1). We say that the effect of the experimental manipulation is *mediated* through its effect on perceived importance and perceived media influence. The error associated with the mediating term (ab) or the sum of product terms ($a_1b_1 + a_2b_2$) needs to be found by bootstrapping the model multiple times (Preacher, 2015; Hayes, 2013; Preacher, Rucker, and Hayes, 2007; MacKinnon, 2008). By default, the `mediate` function in *psych* does 5,000 bootstrap iterations. See the Appendix for sample output. Other packages in R that are specifically designed to test mediation hypotheses include the *mediation* (Tingley, Yamamoto, Hirose, Keele, and Imai, 2014) and *MBESS* (Kelley, 2017) packages.

When doing regressions, we sometimes are interested in the interactions of two of the predictor variables. For instance, when examining how women react to discriminatory treatment of a hypothetical other Garcia, Schmitt, Branscombe, and Ellemers (2010)[4] considered the interactive effects of beliefs about inequality and type of protest (individual vs. collective vs. none) as they affected the appraisal of the other person. This example of a moderated regression is discussed by Hayes (2013).

Interactions (also known as moderation, or moderated regression) are found by entering the product term of the two interacting variables. There are several questions to ask in this analysis that will change the interpretability of the results. For example, should the data be mean-centered before finding the product term, and should the path models be done using standardized or unstandardized regressions? The recommendation from Aiken and West (1991) and Cohen, Cohen, West, and Aiken (2003) is to mean center. However, Hayes (2013) rejects this advice. In both cases, the interaction terms will be identical, but the main effects will differ depending upon centering or not centering. The argument for mean centering is to remove the artificial correlation between the main effects and the interaction term. For with positive numbers X and Y, their product XY will be highly correlated with X and Y. This means that the linear effects of X and Y will be underestimated. The `setCor` and `mediate` functions will by default mean center the data before

[4]With the kind permission of Donna M. Garcia, Michael T. Schmitt, Nyla R. Branscombe, and Naomi Ellemers, the data are included as the `Garcia` data set in the *psych* package

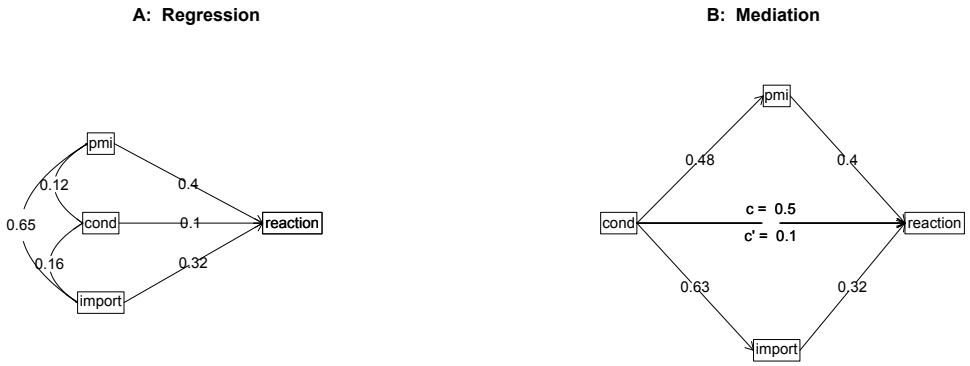

Figure 2. Regression and mediation approaches to the Tal-Or et al. (2010) data set. The curved lines represent covariances, the straight lines, regressions. Panel A shows the full regression model, Panel B shows the total effect (c=.5) and the direct effect (c' = .1) removing the indirect effect (ab) through PMI (.19) and through Import (.20).

finding the product term, however this option can be modified. The `lm` does not and so we need to take an extra-step to do so. The function `scale` will mean center (and by default standardize). The second question, whether to standardize or not, is one of interpretability. Unstandardized coefficients are in the units of the predictors and the criteria and show how much the DV changes per unit change in each IV. The standardized coefficients, on the other hand are unit free and show how much change occurs per standard deviation change in the predictors. Standardization allows for easier comparison across studies but at the cost of losing the direct meaning of the regression slope. In the Appendix we show the code for mean centering using `scale` and then using the `lm` function to do the regression with the interaction term. We also show how the `setCor` function combines both operations.

Correlation, regression and decision making

When reporting standardized regression weights (β_i) the amount of variance in the dependent variable accounted for by the regression model is $R^2 = \Sigma\beta_i r_i$. However it is important to recognize that the slopes (β_i) are the optimal fit for the observed data and that the fit will probably not be as good in another sample. This problem of overfitting is particularly problematic in machine learning (see below) when the number of variables used in the regression is very large. Thus, regression functions will report the R^2 as well as shrunken or adjusted R^2 which estimate what the fit would be in another sample. For

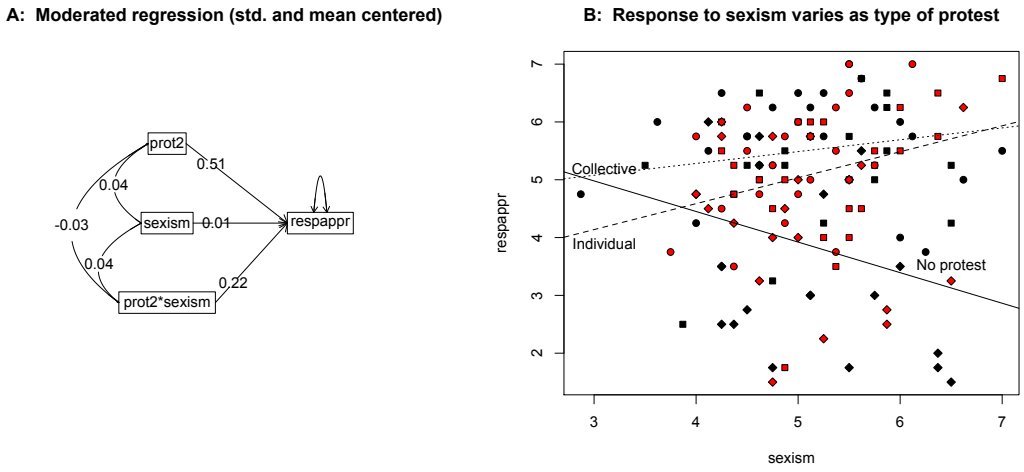

Figure 3. Two ways of showing moderation effects: Panel A, as a path diagram with the product term or Panel B: as a plot of the continuous variable (sexism) showing the individual regression slopes for the three protest conditions. Data from Garcia et al. (2010).

n subjects and k variables, the adjusted $\tilde{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$ (Cohen et al., 2003), that is, there will be more shrinkage for small sample sizes and a large number of predictors.

The R^2 for a particular model is maximized by using the regression weights, but because of what is known as the "Robust beauty of improper linear models" (Dawes, 1979) or the principal that "it don't make no nevermind" (Wainer, 1976), as long as the predictors are moderately correlated with the criterion, using unit weights (1, 0, -1) works almost as well. Weights are said to be 'fungible' (Waller, 2008; Waller and Jones, 2010) in that an infinite set of weights will do almost as good a job as the optimal weights.

Although the variance in the criterion accounted for by the predictors is R^2, it is better to report the actual R, which reflects the amount of change in the criterion for unit changes in the predictors (Ozer, 2007). Change is linear with R, not R^2. This is particularly important when discussing the correlation of a dichotomous predictor with a dichotomous outcome (e.g., applicants are selected or not selected for a job, they succeed or they fail). Consider the four outcomes shown in Table 6 applied to a decision study by Danielson and Clark (1954) and elaborated on by Wiggins (1973). Of 504 military inductees, 89 were later diagnosed as having psychiatric problems requiring their discharge. How well could this future diagnosis be predicted? Using a screening test given to all of the inductees, 55% of the future psychiatric diagnoses could be predicted, with a false alarm (false positive) rate of 19%. This leads to an accuracy of classification (Valid Positives + Valid Negatives) of .76 and a *sensitivity* of .55 and a *specificity* of .81 (Table 7). In this kind of binary

decision, the ϕ coefficient is a linear function of the difference between the percent of Valid Positives and the number expected due to the base rates (BR) times the selection ratio (SR):

$$\phi = \frac{VP - BR * SR}{\sqrt{(BR)(1-BR)(SR)(1-SR)}} \tag{6}$$

In the case of BR = SR =.5, 50% accuracy means a 0 correlation, 60 % a correlation of .2, 70% a correlation of .4, etc. That is, the number of correct predictions is a linear function of the correlation (Ozer, 2007; Rosenthal and Rubin, 1982; Wiggins, 1973).

An alternative approach when considering accuracy in decision making is known as 'signal detection theory' which was developed to model the detection of a signal in a background of noise (Green and Swets, 1966). *d'* (d-prime) relects the sensitivity of the observer and β the criterion the observer was using to make the decision. Similar ideas are seen in the NHST approach to significance testing, where effect size is equivalent to d' and the criterion used (.05, .01) is the decision criterion. The relationship predicted accuracy as a function of the selection ratio, the base rates, and the size of the correlation was discussed by Taylor and Russell (1939) who present tables for different values. The equivalance of these various procedures in seen in (Figure 4) which presents graphically the cell entries in Table 7. The `AUC` (area under the curve) function will take the two by two table of decision theory and report $d', \phi, r_{tetrachoric}$ as well as total accuracy, sensitivity, and specificity.

Table 6: The four outcomes of a decision. Subjects above a particular score on the decision axes are accepted, those below are rejected. Similarly, the criterion of success is such that those above a particular value are deemed to have succeed, those below that value to have failed. All numbers are converted into percentages of the total.

		Decision = Predicted Outcome		
		Accept	Reject	
Outcome	Success	Valid Positive (VP)	False Negative (FN)	Base Rate (BR)
	Failure	False Positive (FP)	Valid Negative (VN)	1 - Base Rate (1-BR)
		Selection Rate (SR)	1-Selection Rate (1-SR)	

Accuracy = Valid Positive + Valid Negative
Sensitivity = Valid Positive /(Valid Positive + False Negative)
Specificity = Valid Negative / (Valid Negative + False Positive)
Phi = $\frac{VP-BR*SR}{\sqrt{BR(1-BR)*SR*(1-SR)}}$

Table 7: Applying decision theory to a prediction problem: the case of predicting future psychiatric diagnoses from military inductees. (Data from Danielson and Clark (1954) as discussed by Wiggins (1973).

Raw Data

	Predicted Positive	Predicted Negative	Row Totals
True Positive	49	40	99
True Negative	79	336	406
Column Totals	118	376	505

Fraction of Total

	Predicted Positive	Predicted Negative	Row Totals
True Positive	.097	.079	.196
True Negative	.157	.667	.804
Column Totals	.234	.746	1.00

$$\begin{aligned}
\text{Accuracy} &= .097 + .667 = .76 \\
\text{Sensitivity} &= .097/(.097 + .079) = .55 \\
\text{Specificity} &= .667 \ / \ (.667 + .157) = .81 \\
\text{Phi} &= \frac{.097 - .196 * .234}{\sqrt{.196 * .804 * .234 * .747}} = .32
\end{aligned}$$

Latent Variable Modeling: EFA, CFA and SEM

There has long been tension in psychological research between understanding and causal explanation versus empirical prediction and control. The concept of the underlying but unobservable cause that accounts for the patterns of observed correlations was implicit in the measurement models of Spearman (1904a) and considered explicitly by Borsboom, Mellenbergh, and van Heerden (2003). This is the logic of the reflective latent variable indicate in the paths of Figure 1 from the latent variables χ_1 or χ_2 to the observed variables $X_1 ... X_6$. Figure 1 represents multiple causal paths: from the latent $\chi_{1..3}$ to the observed $X_{1..9}$ and from the latent $\eta_{1..2}$ to the observed $Y_{1..6}$ as well as from the latent predictors ($\chi_{1..3}$) to the latent criteria ($\eta_{1..2}$). In this perspective, items are reflective measures of the latent trait (Loevinger, 1957; Bollen, 2002) and can be thought to be caused by the latent trait. The contrasting approach of prediction and control was traditionally the domain of behaviorists emphasizing the power of environmental stimuli upon particular response patterns. Stimulus-response theory had no need for latent variables; outcomes were perfectly predictable from the stimulus conditions. In personality research this was the appeal of empirical keys for the MMPI (Hathaway and McKinley, 1943; Butcher, Dahlstrom, Graham, Tellegen, and Kaemmer, 1989), or Strong's Vocational Interest Test (Strong, 1927), and continues now with the statistical learning procedures we will discuss

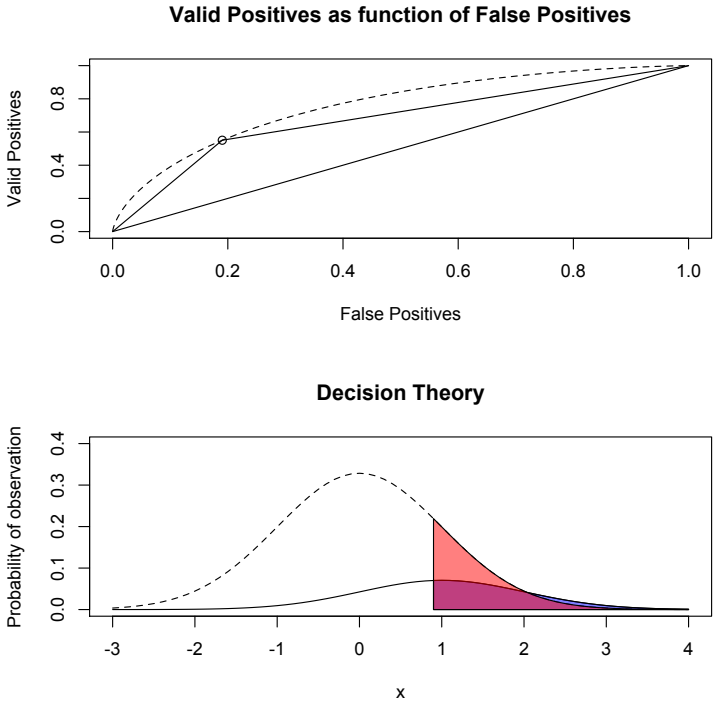

Figure 4. Signal detection theory converts the frequencies of a 2 x 2 table into normal equivalents and shows the relative risks of false positives and false negatives. The number of valid positives will increase at a cost of increasing false positives. Figure from the `AUC` function with input from Table 7.

later. In this empirical approach, scales are composites formed of not necessarily related items. The items are said to be formative indicators that "cause" the latent variable.

Exploratory Factor Analysis

The original concept for factor analysis was Spearman's recognition that the correlations between a number of cognitive ability tests was attenuated due to poor measurement. When correcting for measurement error (see the reliability section where we discuss such corrections for attenuation) all of the cognitive domains were correlated almost perfectly. The underlying latent factor of these tests was thought to be a measure of general intelligence.

Although the initial calculations were done on tables of correlations, when a kindly mathematician told Thurstone in 1931 that his generalization of Spearman's procedure was just the taking the square root of a matrix (Bock, 2007), Thurstone immediately applied

this new matrix algebra to his ability measures and produced his *Vectors of the Mind* (Thurstone, 1933). Correlations were no longer arranged in tables, they were now elements of "correlation matrices". Factor analysis was seen as the approximation of a matrix with one of lesser rank. In modern terminology, factor analysis is just an *eigen decomposition* problem and is a very straight forward procedure.

For any symmetric matrix, $\mathbf{R}$ of rank n there is a set of *eigen vectors* that solve the equation $\mathbf{x_i R} = \lambda_i \mathbf{x_i}$ and the set of n eigenvectors are solutions to the equation

$$\mathbf{XR} = \lambda \mathbf{X}$$

where $\mathbf{X}$ is a matrix of orthogonal eigenvectors and λ is a diagonal matrix of the *eigenvalues*, λ_i. Finding the eigenvectors and eigenvalues is computationally tedious, but may be done using the `eigen` function. That the vectors making up $\mathbf{X}$ are orthogonal means that $\mathbf{XX'} = \mathbf{I}$ and they form the *basis space* for $\mathbf{R}$ that is: $\mathbf{R} = \mathbf{X}\lambda\mathbf{X'}$. In plain terms, it is possible to recreate the correlation matrix $\mathbf{R}$ in terms of an orthogonal set of vectors (the *eigenvectors*) scaled by their associated *eigenvalues*.

We can find the *principal components* of $\mathbf{R}$ by letting

$$\mathbf{C} = \mathbf{X}\sqrt{\lambda}$$

and therefore

$$\mathbf{R} = \mathbf{CC'}. \tag{7}$$

But such a decomposition is not very useful, because the size (rank) of the $\mathbf{X}$ matrix is the same as the original $\mathbf{R}$ matrix. However, if the components are in rank order of their eigenvalues, the first k ($k < n$) components will provide the best fit to the $\mathbf{R}$ matrix when compared to any other set of vectors. Such a principal components analysis (PCA) is useful for optimally describing the observed variables. The components are merely weighted sums of the variables and may be used in applied prediction settings. The components are the k orthgonal sums that best summarize the the total variability of the correlation matrix. The `pca` function will do this analysis.

An alternative model, the *common factor* model attempts to fit the variance that the n variables have in common and ignores that variance which is unique to each variable.

$$\mathbf{R} \approx \mathbf{FF'} + \mathbf{U}^2. \tag{8}$$

where $\mathbf{F}$ is of rank k, and $\mathbf{U}^2$ is a diagonal matrix of rank n. The $\mathbf{U}^2$ matrix may be thought of as the residual variance when we subtract the model ($\mathbf{FF'}$) from the data ($\mathbf{R}$) $\mathbf{U}^2 = \mathbf{R} - \mathbf{FF'}$. Although it would seem that these two equations (7, 8) are quite similar, they are not. For in the first case, the components are formed from linear sums of the variables, while in the second, the variables reflect the linear sums of the factors.

Equation 7 can be solved directly for $\mathbf{C}$, but equation 8 has different solutions for $\mathbf{F}$ depending upon the values in the $\mathbf{U}^2$ matrix which in turn depend upon the value of k.

If we know the amount of variance each variable shares in common with all of the other variables (this is known as the *communality* and is $h_i^2 = 1 - \mathbf{U}_i^2$) then we can solve for the factors. But, unfortunately, we do not know $\mathbf{U}^2$ unless we know $\mathbf{F}$. The solution to this conundrum takes advantage of the power of computers to do *iterative* solutions. Make an initial guess of $\mathbf{U}^2$, solve equation 8 for $\mathbf{F}$ and take the resulting $\mathbf{U}^2$ as the input for the next iteration. Repeat these steps until the change in $\mathbf{U}^2$, from one step to the next is very small and then quit (Spearman, 1927; Thurstone, 1933,9,9).

Consider the correlation matrix in Table 8. A conventional initial estimate for the communalities (diag($\mathbf{I}$-$\mathbf{U}^2$)) might be the Squared Multiple Correlation (SMC) of each variable with all the others (the last line of Table 8 shows these values). Enter these in the diagonal of the matrix and solve for $\mathbf{F}$. Unfortunately exploratory factor analysis is not quite as simple as this for there are at least four decisions that need to be made: what kind of correlation to use, which factor extraction algorithm to use, how many factors to extract, and what rotation or transformation should be applied?

Table 8: The `Thurstone` correlation matrix is a classic data set discussed in detail by R. P. McDonald (McDonald, 1985,9) and and is used as example in the *sem* package as well as in the PROC CALIS manual for SAS. These nine tests were grouped by Thurstone and Thurstone (1941) (based on other data) into three factors: Verbal Comprehension, Word Fluency, and Reasoning). The original data came from Thurstone and Thurstone (1941) but were reanalyzed by Bechtoldt (1961) who broke the data set into two. McDonald, in turn, selected these nine variables from the larger set of 17 found in `Bechtoldt.2`. The sample size is 213.

Variable	Sntnc	Vcblr	Snt.C	Frs.L	F.L.W	Sffxs	Ltt.S	Pdgrs	Ltt.G
Sentences	1.00								
Vocabulary	0.83	1.00							
Sent.Completion	0.78	0.78	1.00						
First.Letters	0.44	0.49	0.46	1.00					
Four.Letter.Words	0.43	0.46	0.42	0.67	1.00				
Suffixes	0.45	0.49	0.44	0.59	0.54	1.00			
Letter.Series	0.45	0.43	0.40	0.38	0.40	0.29	1.00		
Pedigrees	0.54	0.54	0.53	0.35	0.37	0.32	0.56	1.00	
Letter.Group	0.38	0.36	0.36	0.42	0.45	0.32	0.60	0.45	1.00
SMC	0.74	0.75	0.67	0.55	0.52	0.43	0.48	0.45	0.43

Which correlation?

If the data are continuous (or have at least 8-10 response levels), then the normal Pearson r is the appropriate measure of relationship. But if the data are dichotomous (as would be the case for items if scoring correct/correct on an ability test) or polytomous (as is normally the case when scoring personality questionnaires with a 1-5 or 1-6 rating scale), then it is better to use the `tetrachoric` correlation (for dichotomous items) or its generalization to polytomous items, the `polychoric` correlation. The principal reason for

doing so is that the Pearson correlation for items that differ in their mean endorsement rate can not have a high correlation value and is attenuated. As discussed earlier, the tetrachoric is the modeled correlation of the latent traits affecting the scores on the items not the observed scores of the items themselves.

Unfortunately, using tetrachoric correlations will frequently produce correlation matrices which are said to be non-positive-definite, which means some of the eigen values of the matrix are negative. With appropriate assumptions, such matrices can be corrected (*smoothed*) by adding a small number to any negative eigen value, adjusting the positive ones to keep the same total, and then recreating the matrix from the original eigen vectors and the adjusted eigen values (Wothke, 1993). This is done in the `cor.smooth` function.

Factor extraction. Factors are approximate solutions to Equation 8 and have a degree of misfit. Each factoring method attempts to minimize this misfit. The basic fitting equation is

$$E = \frac{1}{2} tr[(\mathbf{R} - \mathbf{F}\mathbf{F}')\mathbf{W}]^2 \tag{9}$$

where *tr* means the trace (sum of the diagonals) of a matrix. If $\mathbf{W}$ is the identity matrix, minimizing E is equivalent to ordinary least squares (OLS); if $\mathbf{W} = \mathbf{R}^{-1}$, it is generalized least squares (GLS), and if $\mathbf{W} = \mathbf{F}\mathbf{F}'^{-1}$ it is maximum likelihood (ML) (Loehlin, 2004). Maximum likelihood (Lawley and Maxwell, 1962,9) has the advantage that under normal theory it finds the model that maximizes the likelihood of the data given the model, but with the disadvantage that it requires taking the inverse of the model. GLS is a close approximation of ML, but requires that the original correlation matrix be invertible. OLS does not require taking inverses but will not produce 'optimal' solutions (in the ML sense). OLS (and the variant known as minimum residual (Harman and Jones, 1966) has the advantage that it is more robust to violations of the model and will produce meaningful solutions even in the presence of many, minor, 'nuisance' factors (MacCallum, Browne, and Cai, 2007). Empirically, although not minimizing the ML criterion, *minres* solutions are very close to it. All of these factor extraction techniques are available in the `fa` function in the *psych* package as are alpha factoring (Kaiser and Caffrey, 1965) and minimum rank factoring (Shapiro and ten Berge, 2002).

Number of factors. An unsolved problem in EFA is how many factors to extract. Henry Kaiser is said to have solved the problem every day before breakfast, but the challenge is to find *the* solution (Horn and Engstrom, 1979). Perhaps the best known solution (Kaiser, 1970) is also the worst: extract as many factors as the number of principal components with eigen values larger than 1. This procedure, although the default for many commercial packages, routinely will extract too many factors (Revelle and Rocklin, 1979). Statistical criteria (e.g., extract factors as long as the χ^2 of the residual matrix is significant) suffer from the problem of being dependent upon sample size: the larger the sample, the more factors are extracted. An appealing technique is to plot the successive eigen values and look for a sharp break. Where the *scree* of trivial factors suddenly jumps to

larger values, stop factoring (Cattell, 1966b). Another useful technique involving the plot of the eigen values is to compare observed values versus those from random data (Horn, 1965). When the observed eigen values are less than those from random data, too many factors have been extracted. This is a useful rule of thumb, but seems to break down with more than about 500-1000 subjects, at which point the random eigen values are all essentially 1.0. Yet another approach is to plot the size of the average minimum partial correlation of the residual matrix. Where this achieves a minimum is an appropriate place to stop (Velicer, 1976). For factoring items, a comparison of the goodness of fit of models which zero out all except the largest loading for each item seems to produce a reasonable estimate (Revelle and Rocklin, 1979). Finally, continuing the the extraction of factors as long as they are interpretable is not unreasonable advice, although those of us unable to interpret many factors will tend to be biased towards extracting fewer. The `nfactors` function applies all of these tests, but unfortunately the typical result is that none of them agree.

Rotations and Transformation. Given a factor solution $\mathbf{F}$ with elements (loadings) of f_{ij} what is the best way to interpret it? The loadings reflect the correlation of the factors with the items and differ by item and by factor. The sum of the squared loadings for each item (row wise) is the amount of variance in that item accounted for by all of the factors. This is known as the communality ($h_i^2 = \Sigma f_{ij}^2$). Items with high communality are well explained by all of the factors, those with low communality are badly explained. For the same value of communality, a variable is said to be more complex if several variables are needed to explain its variance (have high loadings) and less complex if just one variable has a high loading. An index of item complexity is $c_i = \frac{(\Sigma f_{ij}^2)^2}{\Sigma f_{ij}^4}$ which will achieve a minimum of 1 if all of the explained variance in an item is due to one factor (Hofmann, 1978). A similar measure of factor complexity is to do the operation column wise. Multiplying the $\mathbf{F}$ matrix by a orthogonal transformation matrix($\mathbf{T}$) will not change the communalities but can change the item and factor complexities. In the orthogonal case, this is known as rotation, if the resulting solution has correlated factors, we should refer to this as an oblique transformation. We want to chose a transformation that provides a more 'simple structure' (Thurstone, 1947) than the original $\mathbf{F}$ matrix. A number of different solutions to this problem take advantage of the *GPArotation* package (Bernaards and Jennrich, 2005) and are included in the `fa` function. Browne (2001) discusses how many of these are part of the (Crawford and Ferguson, 1970) family of rotations. Some of the most frequently used include *Varimax* (Kaiser, 1958,9) and Quartimax (Neuhaus and Wrigley, 1954) for orthogonal rotations and *oblimin* (Harman, 1976; Jennrich, 1979), *Promax* (Hendrickson and White, 1964) *Bifactor* (Holzinger and Swineford, 1937; Reise, 2012) and *Geomin* (Yates, 1988) for oblique solutions. Unfortunately, some of these rotation procedures achieve local minima in their fitting functions and it is recommended to do multiple random restarts to confirm solutions. The net result of an oblique tranformation is the factor *pattern* matrix

($\mathbf{F}$) and the factor *structure* matrix ($\mathbf{S} = \mathbf{F}\phi$) where ϕ is the correlation between the factors. When reporting an oblique transformation, it is important to show both the pattern ($\mathbf{F}$) and the correlation (ϕ).

Factor score indeterminancy. A problem with factor analysis is that although the model is well defined at the structure level (modeling the covariances of the variables) it is indeterminate at the individual score level (Grice, 2001). Factor scores are best estimates of an individual's score but should not be equated with the factor. Factor score estimates correlate with the latent factors, but this correlation may be far from unity. The `fa` function returns the correlation of the factor scores with the factor. If the correlation is less than .707 (an R^2 of .5), then two estimates of the factor score vector may actually be negatively correlated with each other. Correlations of the factors with the factor score estimates are a function of the number of variables marking the factor as well as the communality of the variables.

Table 9: The *minres* solution using the `fa` function of the `Thurstone` data set. The factor solution was then transformed to simple structure using the `oblimin` transformation. h^2 is the communality estimate, u^2 is the unique variance associated with the variable, *com* is the degree of item complexity. The *pattern* coefficients are shown as well as the correlation (ϕ) between the factors. Because this is an oblique solution, the correlation matrix) is reproduced by $\mathbf{F}\phi\mathbf{F}' + \mathbf{U}^2$. The sums of squares for an oblique solution are the diagonal elements of $\phi\mathbf{F}'\mathbf{F}$.

Variable	MR1	MR2	MR3	h2	u2	com
Sentences	**0.90**	-0.03	0.04	0.82	0.18	1.01
Vocabulary	**0.89**	0.06	-0.03	0.84	0.16	1.01
Sent.Completion	**0.84**	0.03	0.00	0.74	0.26	1.00
First.Letters	0.00	**0.85**	0.00	0.73	0.27	1.00
Four.Letter.Words	-0.02	**0.75**	0.10	0.63	0.37	1.04
Suffixes	0.18	**0.63**	-0.08	0.50	0.50	1.20
Letter.Series	0.03	-0.01	**0.84**	0.73	0.27	1.00
Pedigrees	**0.38**	-0.05	**0.46**	0.51	0.49	1.96
Letter.Group	-0.06	0.21	**0.63**	0.52	0.48	1.25
SS loadings	2.65	1.87	1.49			
MR1	1.00	0.59	0.53			
MR2	0.59	1.00	0.52			
MR3	0.53	0.52	1.00			

Confirmatory Factor Analysis

With the introduction of statistical measures of fit, it is now possible to fit and then test particular factor models (Jöreskog, 1978; Rindskopf and Rose, 1988). Because most models do not fit in an absolute sense, model comparison is recommended. Fit

Table 10: Comparing the Varimax orthogonally rotated PCA ($RC_{1..3}$) Minimum Residual ($MR_{1..3}$), obliquely transformed PCA ($TC_{1..3}$) and oblique Minimum Residual ($MR_{1..3}$) solutions. In order to show the structure more clearly, loadings > .30 are boldfaced.

Variable	RC1	RC2	RC3	MR1	MR2	MR3	TC1	TC2	TC3	MR1	MR2	MR3
Sentences	**0.86**	0.24	0.23	**0.90**	0.01	0.03	**0.83**	0.25	0.26	**0.90**	-0.03	0.04
Vocabulary	**0.85**	**0.31**	0.19	**0.88**	0.10	-0.02	**0.83**	**0.32**	0.22	**0.89**	0.06	-0.03
Sent.Completion	**0.85**	0.26	0.19	**0.89**	0.04	-0.01	**0.78**	0.28	0.23	**0.84**	0.03	0.00
First.Letters	0.23	**0.82**	0.23	0.03	**0.84**	0.07	0.23	**0.79**	0.23	0.00	**0.85**	0.00
Four.Letter.Words	0.18	**0.79**	**0.30**	-0.03	**0.81**	0.16	0.21	**0.71**	0.29	-0.02	**0.75**	0.10
Suffixes	**0.31**	**0.77**	0.06	0.17	**0.79**	-0.14	**0.31**	**0.62**	0.13	0.18	**0.63**	-0.08
Letter.Series	0.25	0.16	**0.83**	0.10	-0.01	**0.84**	0.23	0.18	**0.80**	0.03	-0.01	**0.84**
Pedigrees	**0.53**	0.08	**0.61**	**0.49**	-0.14	**0.55**	**0.45**	0.17	**0.52**	**0.38**	-0.05	**0.46**
Letter.Group	0.10	**0.31**	**0.80**	-0.11	0.21	**0.82**	0.16	**0.31**	**0.63**	-0.06	0.21	**0.63**

statistics include χ^2, the Root Mean Square Error of Approximation (RMSEA) which adjusts the χ^2 for the degrees of freedom and sample size (N) (RMSEA = $\sqrt{\frac{\chi^2 - df}{df(N-1)}}$), the standard deviation of the residuals (RMSR), the Akaike Information Criterion ($AIC = \chi^2 + k(k-1) - 2df$) which also considers the number of variables in the model (k), and the Baysian Information Criterion ($BIC = \chi^2 + ln(N)(k(k+1)/2 + df)$). These fits are actually estimates of misfit. That is, the larger the χ^2, the less well the model fits the data. The question then becomes how bad is bad? Barrett (2007) gives very strict interpretation of what makes a good model; Marsh, Hau, and Wen (2004) suggests there is no golden rule of fit, and Loehlin and Beaujean (2017) give a very useful discussion of how to report fit statistics. The most important thing to remember is that this is a model comparison procedure where we compare multiple models to see which is better, not which is correct.

EFA is seen as a hypothesis generation procedure and CFA as a hypothesis confirmation procedure: the initial model might be derived from an EFA and then tested using CFA on a different data set. A powerful but easy to use package to do this is *lavaan* (Rosseel, 2012). Other CFA packages include *sem* (Fox, Nie, and Byrnes, 2016) and *OpenMX* (Neale, Hunter, Pritikin, Zahery, Brick, Kickpatrick, Estabrook, Bates, Maes, and Boker, 2016). *lavaan* syntax is very straightforward, and allows one to specify and test any particular model.

An important use of CFA is evaluating whether the factor structure of a set of variables is the same across time, or across groups. These are important questions when comparing people across groups or people over time, for it is not possible to make comparisons if the measures are different. Three tests of *factor invariance* are typically considered: configural, metric, and scaler (also known as weak, strong, and strict). Configural asks whether the structure is the same across groups, metric asks whether the loadings are the same (or do not differ very much), and scaler ask whether the means and intercepts of

the factors are the same. Testing for invariance is thus a set of comparisons of structures across groups. The `measurementInvariance` function in the *semTools* package (semTools Contributors, 2016) has been developed to do this in *lavaan*.

Structural Equation Modeling

Combining observations, latent variables and regression into one structural model (see Figure 1) seems to be an obvious step. How to combine the path tracing rules of Wright (1920,9) with factor analysis and reliability theory by using modern estimation algorithms was, however, an important insight. Developed independently by Keesling (1972), Jöreskog (1977) and Wiley (1973) the use of fitting regression models with latent variables was soon identified with a computer algorithm for Linear Structural Relations (LISREL) (Jöreskog, 1978; Joreskog and Sorbom, 1993) (see Tarka (2018) for a thorough history). Very influential texts on SEM include those of Bollen (1989, 2002), Loehlin and Beaujean (2017) and Mulaik (2009). How to report SEM is discussed by McDonald and Ho (2002) and others.

In addition to LISREL, the development of the proprietary programs EQS (Bentler, 1995) and MPLUS (Muthén and Muthén, 2007) made SEM available to many. Now, with the introduction into R of the *sem* (Fox, Nie, and Byrnes, 2013), *lavaan* (Rosseel, 2012) and OpenMx (Neale et al., 2016) packages, SEM and CFA are part of the open source armamentarium for all. Bayesian approaches are available in the *blavaan* (Merkle and Rosseel, 2016) which takes advantage of the *lavaan* package.

Combining formative causal variables with reflective indicator variables is done in MIMIC models (multiple indicators, multiple causes) where a latent variable is seen as formed from a set of causal variables but in turn causes another latent variable which is indicated by a number of reflective measured variables. An early example of a MIMIC model is the causative effect of education, occupation, and income on the latent variable of social status which in turn effects the latent variable of social participation which is measured by church attendance, memberships and the number of friends seen (Jöreskog and Goldberger, 1975).

Perhaps one of the most powerful uses of SEM techniques is in examining growth curves over time. Given a set of participants at time 1, what happens to them over weeks, months, or years? Growth curve models allow for the separation of trait and state effects (Cole, Martin, and Steiger, 2005) as well as an examination of change (McArdle and Bell, 2000; McArdle, 2009). Tutorials for using *lavaan* include growth curve analysis and are included in the help pages for *lavaan*.

In an important generalization of the problem of fungible regression weights (Waller, 2008; Waller and Jones, 2010) Lee, MacCallum, and Browne (2018) showed how with a very small decrease in fit (increase in RMSEA), path coefficients in equivalent models that fit equally well can actually differ in sign. This is just one of the many cautions in how to interpret SEM results. For SEM fit statistics are merely fits of a model to the data. What is needed is comparisons of the fit of alternative/equivalent models. It is important to realize

that reversing the direction of causal arrows in many SEM models does not change the fit, but drastically changes the interpretation (MacCallum, Wegener, Uchino, and Fabrigar, 1993).

Reliability: correlating a test with a test just like it

A powerful use of correlation is assessing reliability. All measures are contaminated with an unknown amount of error. Reliability is just the fraction of the measures that is not error and is the correlation of a measure with another measure that is just like the first measure (Revelle and Condon, 2018a,0). Using V_x to represent observed total test variance and σ_e^2 to represent unobserved error variance then the reliability (r_{xx}) of a measure is

$$r_{xx} = 1 - \frac{\sigma_e^2}{V_x}. \tag{10}$$

In terms of the observed and latent variables in Figure 5, $x = \chi + \epsilon_1$, and a test just like it is $x' = \chi + \epsilon_2$ with correlation r_{xx}. To infer the latent correlation between χ and η, $r_{\chi\eta}$, we can correct the observed correlation r_{xy} for the reliabilities r_{xx} and r_{yy}

$$r_{\chi\eta} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}, \tag{11}$$

Equation 11 was proposed by Spearman (1904b) and the problem since then has been how to estimate the reliability. This is important because if we underestimate the reliability, we will overestimate the disattenuated correlation (Equation 11). There are several different ways to estimate reliability. All agree on the basic principle that items represent some unknown amount of latent score and another unknown amount of error score; the problem is how to measure the relative contributions. Before the era of modern computers, several shortcuts were proposed that – with some very strong assumptions – would allow reliability to be found from the total test variance and the sum of the item variances. Equivalent forms of this procedure are known as (KR20, Kuder and Richardson, 1937) (λ_3, Guttman, 1945) and (α, Cronbach, 1951). Essentially these coefficients are a function of the average inter-item correlation and do not depend upon the structure of the test items. They are all available in the `alpha` function in *psych*.

Model based reliability measures

Procedures that take into account the internal structure of the test using factor analysis (so called model based procedures) include ω_t and ω_h of McDonald (1999), and various estimates of the *greatest lower bound* of the test reliability (Bentler, 2017). By applying factor analytic techniques, it is possible to estimate the amount of variance in a test that is attributable to a general factor (ω_h) (Revelle and Zinbarg, 2009; Zinbarg, Revelle, Yovel, and Li, 2005), as well as the general plus all group factors (total reliability or ω_t). With modern computing techniques, these model based estimates are easy to find

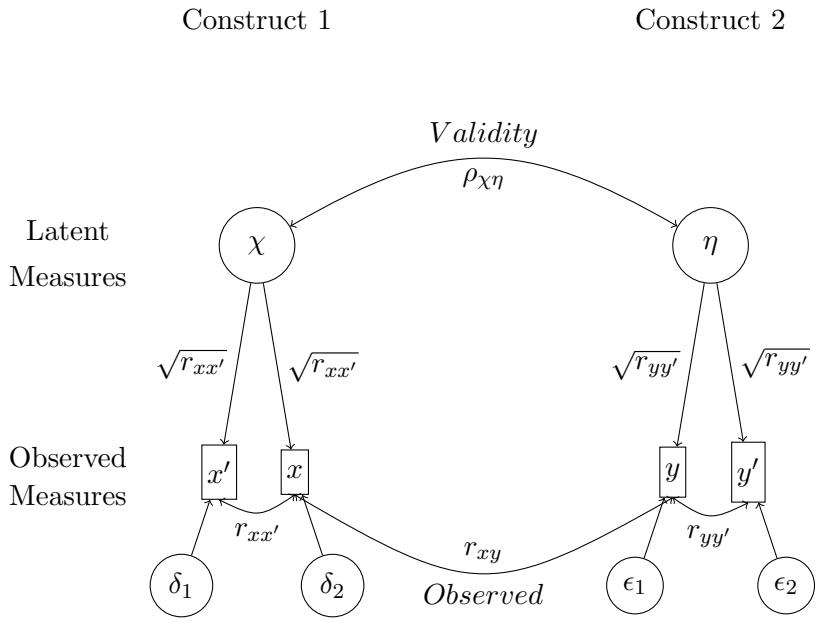

Figure 5. The basic concept of reliability and correcting for attenuation. All four observed variables (x, x', y, y') reflect the latent variables χ and η) but are contaminated by error ($\delta_{1..2}, \epsilon_{1..2}$). Adjusting observed correlations (r_{xy}) by reliabilities ($r_{xx'}$, $r_{yy'}$) estimates underlying latent correlations ($\rho_{\chi\eta}$). (See Equation 11). Observed variables and correlations are shown in conventional Roman fonts, latent variables and latent paths in Greek fonts.

(e.g., `omega` will find ω_h and ω_t as well as α). Estimates of ω_h and ω_t are preferred over α because they are sensitive to the actual structure of the test. α is an estimate based upon the average correlation and the very strong assumption that all the items are equally good measures of the latent trait. (More formally, the items are assumed to be τ equivalent: they all have equal loadings on the latent trait). This was a reasonable assumption to make as a short cut before we had computers. Now, there is no reason to settle for the shortcut. It is not uncommon to find tests with moderate levels of α that actually measure two unrelated or only partially related constructs. It is only by applying model based techniques that we can identify the problem (e.g., Rocklin and Revelle, 1981).

Reliabilty of raters

Rather than evaluating the reliability of a test, sometimes we want to know how much raters agree with each other when making judgements. If the ratings are numeric, this is done by finding one of several Intraclass Correlations (ICC). Depending upon whether we view raters as random or fixed, and whether the raters rate all subjects or just one each, and whether we pool the judgements of different raters, we end up with six different ICCs (Shrout and Fleiss, 1979; Revelle and Condon, 2018b), all of which are found by the `ICC` function. `ICC` uses the power of R to chain functions together: it performs a one or two way analysis of variance, extracts the mean squares or estimates of variance components, and find the appropriate ratio of variance components.

If the ratings are categorical rather than numeric, it is possible to compare the agreement of two raters in terms of Cohen's Kappa statistic (Cohen, 1960,9) which corrects the observed proportions of agreement for the expectations given the marginal base rates. This is done by the `cohen.kappa` function. An example of such ratings is given by Guo et al. (2016) who had raters evaluate the presence or absence of particular themes in a set of life narratives.

Structure vs. Process

Factor analysis and estimates of reliability typically examine the structure of personality items and tests. In terms of Cattell's data box of people by measures by occassions (Cattell, 1946,9) this an example of what Cattell called "R" analysis (people over measures). They do not tell us about how people differ over time. Repeated measures allow us to examine the process of change. With the use of multi-level modeling techniques, it is possible to examine individual differences in within person structure over time (Cattell called this "S" analysis).

Statistical analysis of within subject variability

Although initially done using daily diaries (Bolger, Davis, and Rafaeli, 2003), with the use of personal digital assistants and now cell phone apps, it is possible to collect intensive within subject data once to many times per day for several weeks or even months (Fisher, 2015; Wilt, Funkhouser, and Revelle, 2011; Wilt, Bleidorn, and Revelle, 2017,0; Wilt and Revelle, 2017). Excellent reviews of how to analyze these intensive longitudinal data include Hamaker, Ceulemans, Grasman, and Tuerlinckx (2015) and Hamaker and Wichers (2017). Shrout and Lane (2012) and Revelle and Wilt (2019) provide useful tutorials for examining multilevel reliability, and the `multilevel.reliability` function will do all these calculations in R. A basic question is whether the data are *ergodic* (each individual subject can be seen as representing the entire group) or whether the patterning of each individual is a meaningful signal in its own right. Different approaches to ergodicity include those of Nesselroade and Molenaar (2016) and Revelle and Wilt (2016).

Functions to do these within subject analyses include multilevel regression using *lme4* and *nlme*, correlational structures within and between groups using `statsBy` and examining factor structures for invariance across subjects using `measurementinvariance` in the *semTools* package. The *multilevel* package (Bliese, 2016) comes with very useful documentation on doing some of the more complicated forms of multilevel analyses.

Computational modeling of process

Another very powerful approach to studying within person processes is computational modeling. This is not a statistical approach so much as a way to develop and test the plausibility of theories. It is however, an important use of computer analysis for it can compare and test the fit of alternative dynamic models. Read and Miller and their colleagues (Read, Vanman, and Miller, 1997; Read, Monroe, Brownstein, Yang, Chopra, and Miller, 2010; Yang, Read, Denson, Xu, Zhang, and Pedersen, 2014; Read, Brown, Wang, and Miller, 2018) have implemented a neural network model of the structure and dynamics of individual differences. (Read et al., 2010). Pickering (2008) has implemented several different representations of Gray's Reinforcement Sensitivity Theory (Gray, 1991; Gray and McNaughton, 2000). Although written in MatLab, it is straight forward to translate them into R. A model of the dynamics of action (DOA, Atkinson and Birch, 1970) was implemented as program for main frame computers (Atkinson, Bongort, and Price, 1977) and then reparameterized as the Cues-Tendency Action model and implemented as the `cta` function (Revelle, 1986; Revelle and Condon, 2015). The CTA model has been combined with the RST model into the `cta.rst` function by Brown (2017) and provided good fits to empirical data.

Other statistical techniques

Aggregating data by geographic location

Most personality analysis focuses on individuals, but many interesting questions may be asked concerning differences in the aggregated personality of groups. Whether aggregating the scores of subjects by college major, socioeconomic status, or geographic location, the typical first step is the same: find the mean score for each group. The `statsBy` function provides mean values for any number of variables by a selected grouping variable, as well as a suite of statistics and output for analysis of aggregated variables. For example, the `statsBy` function outputs a correlation matrix for both between groups (the `rbg` object) and within groups (the `rwg` object). The correlation between aggregated groups is known to sociologists as the ecological correlation (Robinson, 1950). The `rbg` object weights correlations by the number of subjects in each group due to the fact that estimates of a mean are more accurate with more subjects. A non-weighted between groups correlation matrix could be obtained by applying the `cor` function to the `mean` object of `statsBy` output.

When analyzing aggregated data, it is critical to keep in mind that a correlation between two variables may not be consistent between groups and within groups (Yule, 1903; Simpson, 1951). Kievit, Frankenhuis, Waldorp, and Borsboom (2013) provide some excellent illustrations of this phenomenon, known as the Yule-Simpson paradox, where the within group correlations are of the opposite sign of the between groups correlations; for example, although a higher dosage of a medication is positively related to the likelihood of patient recovery across genders, dosage is negatively correlated with patient recovery within each gender. Another striking example of the effect is the finding that although at the aggregate level, the University of California seemed to discriminate against women in their admission policy, the individual departments actually discriminate in their favor (Bickel, Hammel, and O'Connell, 1975). The *simpsons* package (Kievit and Epskamp, 2012) allows for detailed examination of data that can produce this effect. The important point, as made by Kievit et al. (2013) and Robinson (1950) is not to dismiss aggregate level relationships but to realize that that the level of generalization depends upon the level of analysis.

How large are the effects of aggregation? Two coefficients (intraclass correlations or ICCs) are reported when examining aggregated data. ICCs describe variance ratios for each aggregated variable in terms of within and between group variance components. ICC1 is an effect size that indicates the percentage of variance in subjects' scores that is explained by group membership (Shrout and Fleiss, 1979). Although sometimes expressed in terms of the between group and within group means squares from the traditional analysis of variance approach, a clearer definition may be expressed as the variance ratio of between group variance (σ^2_{bg}) to the total variance. The total is the sum of the between (σ^2_{bg}) and within group (σ^2_{wg}) variance. ICC2 (also known as ICC1k) takes into account the average number of observations within group ($\bar{n}$) and is a measure of the reliability of the group mean differences. It is the Spearman-Brown reliability formula applied to ICC1:

$$ICC1 = \frac{\sigma^2_{bg}}{\sigma^2_{bg} + \sigma^2_{wg}} \qquad ICC2 = \frac{\sigma^2_{bg}}{\sigma^2_{bg} + \frac{\sigma^2_{wg}}{\bar{n}}}. \tag{12}$$

In plain English, ICC2 indicates the extent to which the aggregated scores of a variable are reliably different from one another. Assuming one's current data are a random sample, if a new random sample is collected with the same average number of participants per group, ICC2 estimates the correlation between the group scores of the first sample and the second sample (James, 1982). ICC2 indicates that aggregated scores are still reliable (i.e., a high ICC2) even if there is a miniscule amount of variance explained in aggregation (i.e., a low ICC1), provided one has enough subjects per group (i.e., a large $\bar{n}$). The `statsBy` function outputs ICC1 (the `ICC1` object), ICC2 (the `ICC2` object) and the number of subjects in each group who responded to each variable (the `n` object).

In the last two decades, international collaborations and online personality assessments have collected enormous data sets with samples an order of magnitude larger than

what was possible a few decades ago (e.g., Gosling, Vazire, Srivastava, and John, 2004; Revelle, Condon, Wilt, French, Brown, and Elleman, 2016). *Geographical psychology* is a subfield that has taken advantage of these large data sets, investigating how and why psychological phenomena are aggregated by residence of geographic locations (Rentfrow and Jokela, 2016), both large (e.g., countries; McCrae and Terracciano, 2008) and small (e.g., postal codes; Jokela, Bleidorn, Lamb, Gosling, and Rentfrow, 2015) . Relatively straightforward correlational analyses at an aggregated geographic level (e.g., Rentfrow, Gosling, and Potter, 2008) are replicable (e.g., Elleman, Condon, Russin, and Revelle, 2018). Geographical psychology researchers have started to explore novel approaches, such as determining the extent to which a psychological phenomenon is spatially clustered between locations; for a psychological variable, is a given location more similar to neighbor locations than more distant locations (e.g., Jokela et al., 2015). Bleidorn, Schönbrodt, Gebauer, Rentfrow, Potter, and Gosling (2016) explored both individual and aggregated levels of their data to calculate "person-city personality fit" and found that this fit was related to the life satisfaction of individuals. The complexities of spatial analysis are beyond the scope of this chapter but see Rentfrow and Jokela (2016) for an overview and Rentfrow (2014) for details. The `spdep` package (Bivand and Piras, 2015) supplies functions pertaining to spatial autocorrelation, weights, statistics, and models.

Statistical Learning Theory

The study of individual differences has expanded beyond academic personality research to computer scientists who are taking advantage of the "big data" possible to collect though web based techniques. Algorithms popular among computer scientists to analyze individual differences data in a prediction context are known generically as "machine learning" or "statistical learning theory." Although some of these techniques are new, some of them repackage traditional methods with new labels (e.g., the ϕ coefficient of Pearson and Heron, 1913, has been 'rediscovered' as a measure of fit but with a new name: the Matthews Correlation Coefficient). We include a brief discussion of these techniques as many personality researchers will likely interact with computer scientists and it is worth learning the "new" vocabulary.

Machine learning is a term without a universally agreed-upon definition. In general, it concerns the prediction of outcome variables from models trained on other datasets and it can be used to refer to a broad range of techniques from logistic regression to neural networks (Hastie, Tibshirani, and Friedman, 2001). The core emphasis in machine learning is on generalization of algorithmic performance to new data, meaning that researchers must shift their focus away from explanation of underlying constructs and toward prediction when using these techniques (Yarkoni and Westfall, 2017). For the purposes of this chapter, it is useful to illustrate some of the methods in the domain of machine learning that are distinct from the traditional regression-based modeling that is more common in psychology.

One such method falls underneath the umbrella term of classification and regression tree (CART) methods. CART methods involve the recursive separation of observations

into distinct subgroups with a goal of enhancing subgroup homogeneity. The algorithm will first segment the observations based on the value of the variable that it has found to lead to the largest reduction in impurity, the exact measure of which depends on the type of problem at hand (e.g., Gini impurity index, entropy). Each of the subgroups created in this split will then be considered separately for further partitioning, until a desired fit to the training data has been reached. This process can be computed automatically through the *rpart* package (Therneau and Atkinson, 2018). CART methods are readily amenable to the creation of attractive output in the form of decision trees, which graphically display a series of sequential steps constituting the final partitions determined by the algorithm. These figures are interpretable by non-statisticians with little training, enhancing the applicability of these methods to a variety of applied contexts. Unfortunately, there is a downside to these methods: classification and regression trees tend to overfit the training sample data, meaning that they tend to extend beyond interpretable signal in a dataset to incorporate noise.Decision tree methods tend to produce models with a high amount of variability. As such, while predictive accuracy may be acceptable on the training dataset, these methods tend to not perform as well as simpler models when applied to out-of-sample data. Fortunately, their predictive capabilities can be dramatically enhanced by incorporating a class of techniques called ensemble methods, which aggregate many different trees to harness their power.

In general, ensemble methods take advantage of the power of averages to create estimates that have lower variance than any of the constituent observations. The logic behind the utility of averages can be understood through an extension of the Central Limit Theorem, in which the variance of the mean of a group of n independent observations, each with a variance of σ^2, is σ^2/n. As such, if we were able to create trees from many different training datasets and aggregate the results, the resultant model would have lower variance than any individual model alone (James, Witten, Hastie, and Tibshirani, 2013). Unfortunately, this is not typically a feasible process, as it would be rare to have many different training datasets at the ready. However, we are able to approximate this process through a method of bootstrap random sampling in which we repeatedly take random samples with replacement from our training dataset, creating many different bootstrapped datasets. This method of aggregation can be applied to a variety of statistical techniques, but applying this method to decision trees, we are able to fit a tree to each of these bootstrapped training datasets, aggregate the results, and get an ultimate prediction that does not suffer as much from the high variance concerns of the individual trees. Thus, ensemble methods form a relatively simple means by which the predictive power of classification and regression trees can be enhanced, making them compelling options for personality researchers with the goal of enhancing prediction.

Three of the most popular ensemble techniques used in the context of tree-based machine learning methods are bagging, random forests, and boosting. In bagging, observations are randomly sampled from the training dataset through bootstrapping methods to create many distinct datasets. A different tree is then grown based on each of these

bootstrapped datasets, resulting in many trees, each of which have a potentially different set of decision rules on which they based their predictions. The results of these trees are then aggregated to create a final prediction for each observation. Predictions are evaluated on the 'out-of-bag' observations that were left out of the bootstrap samples, providing a validation set of all observations not used in the creation of a subset of trees (James et al., 2013). Random forest methods can be viewed as special forms of bagging in which variables are randomly selected alongside the observations in the training dataset. More specifically, during the process of growing the tree from the training dataset, variables are randomly selected at each node to be used in the creation of the tree. Why add this additional step? While the general process of bagging reduces variance by bootstrapping multiple trees, each of the trees will be correlated due to the inclusion of identical variables at every split. In certain cases (i.e., when one variable is much more important in the prediction of the outcome than others) this will lead to many trees that are nearly identical. Averaging trees that are very similar will not lead to as large a reduction in variance as averaging uncorrelated trees. Random forest methods address this by changing the variables made available at each potential node split in the trees, effectively decorrelating the trees and theoretically decreasing the variance of a given predictive model. In general, if p is the number of predictors, $\sqrt{p}$ is the number of variables selected at each split for classification problems and p/3 variables are selected at each split for regression problems, although this number should be determined based on hyperparameter tuning given the dataset at hand (Hastie et al., 2001). Both bagging and random forest techniques can be computed in R using the *randomForest* package (Liaw and et al., 2002). By setting the 'mtry' argument to equal the number of variables in the model, a bagged model will be run, otherwise it will be random forest.

Boosting methods are conceptually similar to bagging and random forests, but alter the way that the trees update their information. Specifically, boosting involves sequentially updating each tree by fitting the residuals of the tree before it. In this manner, boosting attempts to target the areas of weakness of the previous trees and update them accordingly. The boosting algorithm completes this procedure slowly to avoid overfitting, each time fitting the residuals of the previous model before adding this fitted tree back to the original tree after applying a shrinkage parameter, thus updating the residuals. The power in the predictive ability of boosting comes from its slow progression and sequential growth based on residuals of previous trees. However, this practice of fitting residuals can lead to overfitting if the tree grows too fast and thus we want the algorithm to proceed slowly. Boosting tends to outperform bagging and random forest on prediction metrics, but may not be as conceptually clear as those methods due to the fitting of residuals. The preference is up to the researcher. Popular packages in R to fit boosted models include *xgboost* (Chen, He, Benesty, Khotilovich, Tang, Cho, Chen, Mitchell, Cano, Zhou, Li, Xie, Lin, Geng, and Li, 2018) and *gbm* (Ridgeway et al., 2017). *Python* is another language commonly used in machine learning as it allows somewhat faster processing of very large data sets.

Taken together, these ensemble methods are referred to as "black box" methods,

which means that their exact inner workings remain unknown to the practitioner. In other words, while a researcher may be able to tell that a random forest model provides excellent predictive accuracy, they would not be able to view each decision tree used in the crafting of the prediction. As such, these ensemble methods are best used when the major aim of a project is prediction and they may not be appropriate for situations in which a precise theoretical model is desired. This may be discomforting to personality psychologists who have largely been trained to prioritize theoretical modeling, but it is simply another possible way to analyze data with a predictive focus. Researchers are not left completely in the dark about the inner workings of the models, however: they are able to influence the way that the model constructs and aggregates the individual trees in these ensemble methods through the selection of various hyperparameters. These hyperparameters include the number of trees grown in all three methods, number of variables selected at each node in random forests, and how slowly the trees grow in boosting. Adjusting these hyperparameters to influence the performance of the ensemble methods is critical to predictive performance and can be done with through trial and error or cross validation.

The aforementioned methods are intriguing in that they provide the tools for creating a more prediction-focused study of individual differences. While promising in their potential impact on the field in the future, machine learning results should generally be viewed through a lens of caution, as many complicated methods may be outperformed by comparatively simpler methods of linear and logistic regression. A shift toward a more predictive science would be welcomed, but we must be sure to select methods to suit the problem at hand and not just apply these methods with abandon.

Conclusion

Personality research has come a long way from the simple correlation of Galton (1888), Pearson (1895), and Spearman (1904b). Advances in the past few years have brought powerful computation to the desk or lap of the individual researcher. Open source software has made complex research questions answerable by people anywhere in the world. Computational model techniques that used to take days on multi-million dollar computers can now be done in seconds on very affordable laptops. Data can be shared across the web, analyses can be duplicated using published and open source computer code. Statistical testing and modeling of psychological data and theory has never been easier for those willing to learn the modern methods.

References

Aiken, L. S., and West, S. G. (1991). *Multiple regression testing and interpretation*. Sage Publications, Inc.

Algina, J., Keselman, H. J., and Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods*, *10*, 317 – 328.

Atkinson, J. W., and Birch, D. (1970). *The dynamics of action*. New York, N.Y.: John Wiley.

Atkinson, J. W., Bongort, K., and Price, L. (1977). Explorations using computer simulation to comprehend thematic apperceptive measurement of motivation. *Motivation and Emotion*, *1*, 1–27.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*, 815–824.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. R package version 1.1-8.

Bechtoldt, H. (1961). An empirical study of the factor analysis stability hypothesis. *Psychometrika*, *26*, 405–432.

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988). The new S language. *Pacific Grove, Ca.: Wadsworth & Brooks, 1988*.

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA.: Multivariate Software, Inc.

Bentler, P. M. (2017). Specificity-enhanced reliability coefficients. *Psychological Methods*, *22*, 527 – 540.

Bernaards, C., and Jennrich, R. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement*, *65*, 676–696.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, *187*, 398–404.

Bivand, R., and Piras, G. (2015). Comparing implementations of estimation methods for spatial econometrics. *Journal of Statistical Software*, *63*(18), 1–36.

Bleidorn, W., Schönbrodt, F., Gebauer, J. E., Rentfrow, P. J., Potter, J., and Gosling, S. D. (2016). To live among like-minded others: Exploring the links between person-city personality fit and self-esteem. *Psychological Science*.

Bliese, P. (2016). *multilevel: Multilevel Functions*. R package version 2.6.

Bock, R. D. (2007). Rethinking Thurstone. In R. Cudeck, and R. C. MacCallum (Eds.) *Factor analysis at 100: Historical developments and future directions*, (pp. 35–45). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Bolger, N., Davis, A., and Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, *54*, 579–616.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*, 605–634.

Borsboom, D., Mellenbergh, G. J., and van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*, 203–219.

Brogden, H. E. (1946). On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology*, *37*, 65 – 76.

Bromley, A. G. (1982). Charles Babbage's Analytical Engine, 1838. *IEEE annals of the history of computing*, *4*, 196–217.

Brown, A. D. (2017). *The Dynamics of Affect: Using Newtonian Mechanics, Reinforcement Sensitivity Theory, and the Cues-Tendencies-Actions Model to Simulate Individual Differences in Emotional Experience*. Ph.D. thesis, Northwestern University.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*, 296–322.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150.

Butcher, J. N., Dahlstrom, W., Graham, J., Tellegen, A., and Kaemmer, B. (1989). *MMPI-2: Manual for administration and scoring*. Minneapolis: University of Minnesota Press.

Cattell, R. B. (1946). Personality structure and measurement. I. The operational determination of trait unities. *British Journal of Psychology*, *36*, 88–102.

Cattell, R. B. (1966a). The data box: Its ordering of total resources in terms of possible relational systems. In R. B. Cattell (Ed.) *Handbook of multivariate experimental psychology*, (pp. 67–128). Chicago: Rand-McNally.

Cattell, R. B. (1966b). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276.

Champely, S. (2018). *pwr: Basic Functions for Power Analysis*. R package version 1.2-2.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., and Li, Y. (2018). *xgboost: Extreme Gradient Boosting*. R package version 0.71.1.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(37-46).

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *The Journal of Abnormal and Social Psychology*, *65*, 145–153.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J.: L. Erlbaum Associates, 2nd ed ed.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*, 155–159.

Cohen, J. (1994). The earth is round ($p < .05$). *American psychologist*, *49*, 997–1003.

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, N.J.: L. Erlbaum Associates, 3rd ed ed.

Cole, D. A., Martin, N. C., and Steiger, J. H. (2005). Empirical and conceptual problems with longitudinal trait-state models: Introducing a trait-state-occasion model. *Psychological Methods*, *10*, 3–20.

Crawford, C. B., and Ferguson, G. A. (1970). A general rotation criterion and its use in orthogonal rotation. *Psychometrika*, *35*, 321–332.

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

Danielson, J. R., and Clark, J. H. (1954). A personality inventory for induction screening. *Journal of Clinical Psychology*, *10*, 137 – 143.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*, 571–582.

Dixon, W. J., and Brown, M. B. (1979). *BMDP-79: Biomedical computer programs P-series*. Univ of California Press.

Eckart, C., and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*, 211–218.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1-26).

Efron, B., and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, *37*, 36–48.

Elleman, L. G., Condon, D. M., Russin, S. E., and Revelle, W. (2018). The personality of U.S. states: Stability from 1999 to 2015. *Journal of Research in Personality*, *72*, 64 – 72. Special issue of Replication of Critical Findings in Personality Psychology.

Erceg-Hurn, D. M., and Mirosevich, V. M. (2008). Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*, 591.

Eysenck, H. J. (1944). Types of personality: a factorial study of seven hundred neurotics. *The British Journal of Psychiatry*, *90*(381), 851–861.

Fisher, A. J. (2015). Toward a dynamic model of psychological assessment: Implications for personalized care. *Journal of Consulting and Clinical Psychology*, *83*, 825 – 836.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 3–32.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fox, J., Nie, Z., and Byrnes, J. (2013). *sem: Structural Equation Models*. R package version 3.1-3.

Fox, J., Nie, Z., and Byrnes, J. (2016). *sem: Structural Equation Models*. R package version 3.1-7.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263.

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

Garcia, D. M., Schmitt, M. T., Branscombe, N. R., and Ellemers, N. (2010). Women's reactions to ingroup members who protest discriminatory treatment: The importance of beliefs about inequality and response appropriateness. *European Journal of Social Psychology*, *40*, 733–745.

Gosling, S. D., Vazire, S., Srivastava, S., and John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*, 93–104.

Gray, J. A. (1991). The neuropsychology of temperament. In J. Strelau, and A. Angleitner (Eds.) *Explorations in temperament: International perspectives on theory and measurement*, (pp. 105–128). New York, NY: Plenum Press.

Gray, J. A., and McNaughton, N. (2000). *The Neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system*. Oxford: Oxford University Press, 2nd ed.

Green, D. M., and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. Oxford: John Wiley.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*, 430–450.

Guo, J., Klevan, M., and McAdams, D. P. (2016). Personality traits, ego development, and the redemptive self. *Personality and Social Psychology Bulletin*, *42*, 1551–1563.

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

Hamaker, E. L., Ceulemans, E., Grasman, R., and Tuerlinckx, F. (2015). Modeling affect dynamics: State of the art and future challenges. *Emotion Review*, *7*, 316–322.

Hamaker, E. L., and Wichers, M. (2017). No time like the present. *Current Directions in Psychological Science*, *26*, 10–15.

Harman, H. H. (1976). *Modern factor analysis*. Chicago: University of Chicago Press, 3d ed., rev ed.

Harman, H. H., and Jones, W. (1966). Factor analysis by minimizing residuals (minres). *Psychometrika*, *31*, 351–368.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. Springer-Verlag New York, Inc., New York, 2 ed.

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

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York: Guilford Press.

Hendrickson, A. E., and White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, *17*, 65–70.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–83.

Hofmann, R. J. (1978). Complexity and simplicity as objective indices descriptive of factor solutions. *Multivariate Behavioral Research*, *13*, 247–250.

Holzinger, K., and Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*, 41–54.

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

Horn, J. L., and Engstrom, R. (1979). Cattell's scree test in relation to Bartlett's chi-square test and other observations on the number of factors problem. *Multivariate Behavioral Research*, *14*, 283–300.

Isaacson, W. (2014). *The Innovators: How a Group of Inventors, Hackers, Geniuses and Geeks Created the Digital Revolution*. Simon and Schuster.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*. vol. 112. Springer.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology*, *67*, 219 – 229.

Jennrich, R. I. (1979). Admissible values of γ in direct oblimin rotation. *Psychometrika*, *44*, 173–177.

Jokela, M., Bleidorn, W., Lamb, M. E., Gosling, S. D., and Rentfrow, P. J. (2015). Geographically varying associations between personality and life satisfaction in the london metropolitan area. *Proceedings of the National Academy of Sciences*, *112*, 725–730.

Jöreskog, K. G. (1977). *Applications of statistics: proceedings of the symposium held at Wright State University*, chap. Structural Equation Models in the social sciences: Spcification, estimation, and testing. North Holland.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, *43*, 443–477.

Jöreskog, K. G., and Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable,. *Journal of the American Statistical Association*, *70*(351a), 631–639.

Joreskog, K. G., and Sorbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.*. Lisrel 8: Lawrence Erlbaum Associates, Inc.

Judd, C. M., and McClelland, G. H. (1989). *Data analysis : a model-comparison approach*. San Diego: Harcourt Brace Jovanovich.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187–200.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika*, *35*, 401–415.

Kaiser, H. F., and Caffrey, J. (1965). Alpha factor analysis. *Psychometrika*, *30*, 1–14.

Keesling, W. (1972). *Maximum likelihood approaches to causal flow analysis.*. Ph.D. thesis, University of Chicago.

Kelley, K. (2017). *MBESS: The MBESS R Package*. R package version 4.4.1.

Kievit, R. A., and Epskamp, S. (2012). *Simpsons: Detecting Simpson's Paradox. R package version 0.1.0.*.

Kievit, R. A., Frankenhuis, W. E., Waldorp, L. J., and Borsboom, D. (2013). Simpson's paradox in psychological science: a practical guide. *Frontiers in Psychology*, *4*(513), 1–14.

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

Lawley, D. N., and Maxwell, A. E. (1962). Factor analysis as a statistical method. *The Statistician*, *12*, 209–229.

Lawley, D. N., and Maxwell, A. E. (1963). *Factor analysis as a statistical method*. London: Butterworths.

Lee, T., MacCallum, R. C., and Browne, M. W. (2018). Fungible parameter estimates in structural equation modeling. *Psychological Methods*, *23*, 58 – 75.

Liaw, A., and et al., M. W. (2002). Classification and regression by randomforest. *R news*, *2*, 18–22.

Loehlin, J. C. (2004). *Latent variable models: an introduction to factor, path, and structural equation analysis*. Mahwah, N.J.: L. Erlbaum Associates, 4th ed.

Loehlin, J. C., and Beaujean, A. (2017). *Latent variable models: an introduction to factor, path, and structural equation analysis*. Mahwah, N.J.: Routledge, 5th ed.

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

Lovelace, A. A. (1842). Sketch of the analytical engine invented by Charles Babbage, by LF Menabrea, officer of the military engineers, with notes upon the memoir by the translator. *Taylor's Scientific Memoirs*, *3*, 666–731.

MacCallum, R. C., Browne, M. W., and Cai, L. (2007). Factor analysis models as approximations. In R. Cudeck, and R. C. MacCallum (Eds.) *Factor analysis at 100: Historical developments and future directions*, (pp. 153–175). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., and Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological bulletin.*, *114*, 185–199.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York, NY US: Lawrence Erlbaum Associates Taylor & Francis Group.

Mair, P., Schoenbrodt, F., and Wilcox, R. (2017). *WRS2: Wilcox robust estimation and testing*. R package 0.9-2.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*, pp. 519–530.

Marsh, H. W., Hau, K.-T., and Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*, 320–341.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*, 577–605.

McArdle, J. J., and Bell, R. Q. (2000). Recent trends in modeling longitudinal data by latent growth curve methods. In T. D. Little, K. U. Schnabel, and J. Baumert (Eds.) *Modeling longitudinal and multiple-group data: practical issues, applied approaches, and scientific examples*, (pp. 69–107). Mahwah, NJ: Lawrence Erlbaum Associates.

McCrae, R. R., and Terracciano, A. (2008). *Multilevel analysis of individuals and cultures.*, chap. The Five-Factor Model and its correlates in individuals and cultures, (pp. 249–283.). New York, NY: Taylor & Francis Group/Lawrence Erlbaum Associates.

McDonald, R. P. (1985). *Factor Analysis and Related Methods*. Hillsdale, NJ:: Erlbaum.

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

McDonald, R. P., and Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological methods*, *7*, 64–82.

Merkle, E. C., and Rosseel, Y. (2016). blavaan: Bayesian structural equation matrix models via parameter expansion. *arXiv*, *1511.05604*.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Chapman & Hall/CRC statistics in the social and behavioral sciences series. Boca Raton: CRC Press.

Muthén, L., and Muthén, B. (2007). *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén, fifth edition ed.

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kickpatrick, R. M., Estabrook, R., Bates, T. C., Maes, H. H., and Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*.

Nesselroade, J. R., and Molenaar, P. C. M. (2016). Some behavioral science measurement concerns and proposals. *Multivariate Behavioral Research*, *51*, 396–412.

Neuhaus, J., and Wrigley, C. (1954). The quartimax method: an analytical approach to orthogonal simple structure. *British Journal of Statistical Psychology*, *7*, 81–91.

Ozer, D. J. (2007). Evaluating effect size in personality research. In R. W. Robins, R. C. Fraley, and R. F. Krueger (Eds.) *Handbook of research methods in personality psychology*, (pp. 495–501). New York, NY: Guilford Press.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society. London Series*, *LVIII*, 240–242.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philisopical Transactions of the Royal Society of London. Series A*, *187*, 254–318.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, *13*, 25–45.

Pearson, K., and Heron, D. (1913). On theories of association. *Biometrika*, *9*(1/2), 159–315.

Pek, J., and Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, *25*, 208–225.

Pickering, A. D. (2008). Formal and computational models of reinforcement sensitivity theory. In P. J. Corr (Ed.) *The Reinforcement Sensivity Theory*, (pp. 453–481). Cambridge: Cambridge University Press.

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

Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology*, *66*, 825–852.

Preacher, K. J., Rucker, D. D., and Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate behavioral research*, *42*, 185–227.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Read, S. J., Brown, A. D., Wang, P., and Miller, L. C. (2018). The virtual personalities neural network model: Neurobiological underpinnings. *Personality Neuroscience.*.

Read, S. J., Monroe, B. M., Brownstein, A. L., Yang, Y., Chopra, G., and Miller, L. C. (2010). A neural network model of the structure and dynamics of human personality. *Psychological Review*, *117*, 61 – 92.

Read, S. J., Vanman, E. J., and Miller, L. C. (1997). Connectionism, parallel constraint satisfaction processes, and gestalt principles: (re)introducing cognitive dynamics to social psychology. *Personality and Social Psychology Review*, *1*, 26–53.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667–696.

Rentfrow, P. J. (Ed.) (2014). *Geographical Psychology: Exploring the Interaction of Environment and Behavior*. American Psychological Association.

Rentfrow, P. J., Gosling, S. D., and Potter, J. (2008). A theory of the emergence, persistence, and expression of geographic variation in psychological characteristics. *Perspectives on Psychological Science*, *3*, 339–369.

Rentfrow, P. J., and Jokela, M. (2016). Geographical psychology: The spatial organization of psychological phenomena. *Current Directions in Psychological Science*, *25*, 393–398.

Revelle, W. (1986). Motivation and efficiency of cognitive performance. In D. R. Brown, and J. Veroff (Eds.) *Frontiers of Motivational Psychology: Essays in honor of J. W. Atkinson*, chap. 7, (pp. 105–131). New York: Springer.

Revelle, W. (2007). Experimental approaches to the study of personality. In R. Robins, R. C. Fraley, and R. F. Krueger (Eds.) *Handbook of research methods in personality psychology.*, (pp. 37–61). New York: Guilford.

Revelle, W. (2018). *psych: Procedures for Personality and Psychological Research*. Northwestern University, Evanston, https://CRAN.r-project.org/package=psych. R package version 1.8.12.

Revelle, W., Condon, D., Wilt, J., French, J. A., Brown, A. D., and Elleman, L. G. (2016). Web and phone based data collection using planned missing designs. In G. B. Nigel G. Fielding, Raymond M. Lee (Ed.) *The Sage Handbook of Online Research Methods*, chap. 33, (pp. 578–595). SAGE Publications, 2nd ed.

Revelle, W., and Condon, D. M. (2015). A model for personality at three levels. *Journal of Research in Personality*, *56*, 70–81.

Revelle, W., and Condon, D. M. (2018a). Reliability. In P. Irwing, T. Booth, and D. J. Hughes (Eds.) *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*. London: John Wily & Sons.

Revelle, W., and Condon, D. M. (2018b). Reliability from α to ω: A tutorial. *(under review: https://osf.io/e685p/)*.

Revelle, W., and Rocklin, T. (1979). Very Simple Structure - alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, *14*, 403–414.

Revelle, W., and Wilt, J. (2016). The data box and within subject analyses: A comment on Nesselroade and Molenaar. *Multivariate Behavioral Research*, *51*(2-3), 419–421.

Revelle, W., and Wilt, J. A. (2019). Analyzing dynamic data: a tutorial. *Personality and Individual Differences*, *136*, 38–51.

Revelle, W., and Zinbarg, R. E. (2009). Coefficients alpha, beta, omega and the glb: comments on Sijtsma. *Psychometrika*, *74*, 145–154.

Ridgeway, G., et al. (2017). *gbm: Generalized Boosted Regression Models*. R package version 2.1.3.

Rindskopf, D., and Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, *23*, 51–67.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*, 351–357.

Rocklin, T., and Revelle, W. (1981). The measurement of extraversion: A comparison of the Eysenck Personality Inventory and the Eysenck Personality Questionnaire. *British Journal of Social Psychology*, *20*, 279–284.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*, 1 – 12.

Rosenthal, R. (1994). Parametric measures of effect size. *The handbook of research synthesis*, (pp. 231–244).

Rosenthal, R., and Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, *74*, 166 – 169.

Rosnow, R. L., and Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *57*, 221–237.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

semTools Contributors (2016). *semTools: Useful tools for structural equation modeling*. R package version 0.4-13.

Shapiro, A., and ten Berge, J. M. (2002). Statistical inference of minimum rank factor analysis. *Psychometrika*, *67*, 79–94.

Shrout, P. E., and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Shrout, P. E., and Lane, S. P. (2012). Psychometrics. In *Handbook of research methods for studying daily life*. Guilford Press.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *13*, 238–241.

Spearman, C. (1904a). "General Intelligence," objectively determined and measured. *American Journal of Psychology*, *15*, 201–292.

Spearman, C. (1904b). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72–101.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271–295.

Spearman, C. (1927). *The abilities of man*. Oxford England: Macmillan.

SPSS (2008). Version 17.0. *Chicago: SPSS Inc*.

Streiner, D. L. (2003). Unicorns *Do* exist: A tutorial on "proving" the null hypothesis. *The Canadian Journal of Psychiatry*, *48*, 756–761.

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

Student (1908). The probable error of a mean. *Biometrika*, *6*, 1–25.

Tal-Or, N., Cohen, J., Tsfati, Y., and Gunther, A. C. (2010). Testing causal direction in the influence of presumed media influence. *Communication Research*, *37*, 801–824.

Tarka, P. (2018). An overview of structural equation modeling: its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & Quantity*, *52*, 313–354.

Taylor, H. C., and Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology*, *23*, 565 – 578.

Therneau, T., and Atkinson, B. (2018). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-13.

Thurstone, L. L. (1933). *The theory of multiple factors*. Ann Arbor, Michigan: Edwards Brothers.

Thurstone, L. L. (1934). The vectors of mind. *Psychological Review*, *41*, 1.

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

Thurstone, L. L. (1947). *Multiple-factor analysis: a development and expansion of The vectors of the mind*. Chicago, Ill.: The University of Chicago Press.

Thurstone, L. L., and Thurstone, T. G. (1941). *Factorial studies of intelligence*. Chicago, Ill.: The University of Chicago press.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., and Imai, K. (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software*, *59*, 1–38.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples (preliminary report) (abstract). *Ann. Math. Statist.*, *29*, 614.

Velicer, W. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*, 321–327.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83*, 213–217.

Waller, N. G. (2008). Fungible weights in multiple regression. *Psychometrika*, *73*, 691–703.

Waller, N. G., and Jones, J. A. (2010). Correlation weights in multiple regression. *Psychometrika*, *75*, 58–69.

Wiggins, J. S. (1973). *Personality and prediction: principles of personality assessment*. Reading, Mass.: Addison-Wesley Pub. Co.

Wilcox, R. R. (2001). Modern insights about Pearson's correlation and least squares regression. *International Journal of Selection and Assessment*, *9*, 195–205.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing*. Statistical modeling and decision science. Amsterdam: Boston: Elsevier/Academic Press, 2nd ed.

Wilcox, R. R., and Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, *8*, 254–274.

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

Wilt, J., Bleidorn, W., and Revelle, W. (2016). Finding a life worth living: Meaning in life and graduation from college. *European Journal of Personality*, *30*, 158–167.

Wilt, J., Bleidorn, W., and Revelle, W. (2017). Velocity explains the links between personality states and affect. *Journal of Research in Personality*, *69*, 86-95.

Wilt, J., Funkhouser, K., and Revelle, W. (2011). The dynamic relationships of affective synchrony to perceptions of situations. *Journal of Research in Personality*, *45*, 309–321.

Wilt, J., and Revelle, W. (2017). The big five, situational context, and affective experience. *Personality and Individual Differences*.

Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K. A. Bollen, and J. S. Long (Eds.) *Testing structural equation models*, chap. 11, (pp. 256–293). Newbury Park: Sage Publications.

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceedings of the National Academy of Sciences*, *6*, 320–332.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*, 557–585.

Yang, Y., Read, S. J., Denson, T. F., Xu, Y., Zhang, J., and Pedersen, W. C. (2014). The key ingredients of personality traits: Situations, behaviors, and explanations. *Personality and Social Psychology Bulletin*, *40*, 79–91.

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

Yates, A. (1988). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis*. Suny Press.

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, *2*, 121–134.

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, *LXXV*, 579–652.

Zinbarg, R. E., Revelle, W., Yovel, I., and Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*, 123–133.

Appendix

The R code for the various examples is shown here.

Table 1 was a subset of the `msqR` data set which is included in the *psych* package. Here we show the size of the entire data set (6411 rows by 79 columns) , the number of subjects with repeated measures (2086), and how to form a subset of the first eight cases for both time 1 and time 2.

R code

```
msq.items <- c("anxious" ,  "at.ease" ,  "calm" , "confident", "content",
 "jittery",  "nervous", "relaxed" , "tense" , "upset" ) #these overlap with the sai
dim(msqR)  #show the dimensions of the data set
colnames(msqR)  #what are the variables

table(msqR$time)  #show the number of observations with various repeated values
example <- msqR[c(1:8,69:76),c(cs(id,time),msq.items)]
df2latex(example)  #make a \LateX table of the example data
```

Descriptive statistics

Table 2 shows descriptive statistics.

R code

```
describe(msqR[c(1:8,69:76),c(cs(id,time),msq.items)],IQR=TRUE) #for the data in table 1
describe(msqR[c(cs(id,time),msq.items) #describe the entire data set
```

Correlation and regression

Table 5 shows the correlation matrix from the Tal-Or et al. (2010) data set. Here we show several different ways to show those correlations and to test for their significance. In this and the subsequent examples, we use the standard notation for $y \sim x$. Unfortunately, the $\sim$ symbol renders poorly and for those who want to directly copy from the pdf, the $\sim$ symbol should be written in by hand.

R code

```
describe(Tal_Or)  #the descriptive statistics for the data.


t.test(reaction ~ cond, data=Tal_Or)  # The t.test of interest
t.test(pmi ~ cond, data=Tal_Or)  # Also test the effects on pmi
t.test(import ~ cond, data=Tal_Or)  #and import

cor(Tal_Or)  #the core-R command displays to 9 decimals

#or  just show the lower diagonal of the correlations,
lowerCor(Tal_Or) #round the results to two decimals and abbreviate the names

corr.test(Tal_Or) # find the correlations, the raw p values and the adjusted p values

 cor.ci(Tal_Or[1:4], n.iter=1000)
cor2latex(Tal_Or[1:4],stars=TRUE,adjust="none")  #create the Table
```

Produces this output:

```
> describe(Tal_Or)  #the descriptive statistics for the data.
          vars   n  mean   sd median trimmed  mad min max range  skew kurtosis   se
cond         1 123  0.47 0.50   0.00    0.46 0.00   0   1     1  0.11    -2.00 0.05
pmi          2 123  5.60 1.32   6.00    5.78 1.48   1   7     6 -1.17     1.30 0.12
import       3 123  4.20 1.74   4.00    4.26 1.48   1   7     6 -0.26    -0.89 0.16
reaction     4 123  3.48 1.55   3.25    3.44 1.85   1   7     6  0.21    -0.90 0.14
gender       5 123  1.65 0.48   2.00    1.69 0.00   1   2     1 -0.62    -1.62 0.04
age          6 123 24.63 5.80  24.00   23.76 1.48  18  61    43  4.71    24.76 0.52

> t.test(reaction ~ cond, data=Tal_Or)  # The t.test of interest

        Welch Two Sample t-test

data:  reaction by cond
t = -1.7964, df = 120.98, p-value = 0.07492
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.04196792  0.05058861
sample estimates:
mean in group 0 mean in group 1
        3.25000         3.74569
    ... output omitted

 cor(Tal_Or)  #the core-r command
                cond          pmi     import    reaction      gender          age
cond       1.00000000  0.180773560 0.18091083  0.16026292 -0.12717905  0.025245417
pmi        0.18077356  1.000000000 0.28207107  0.44649392 -0.02112095 -0.004947199
import     0.18091083  0.282071074 1.00000000  0.46477681  0.02700985  0.073431563
reaction   0.16026292  0.446493916 0.46477681  1.00000000  0.01436459 -0.083728952
gender    -0.12717905 -0.021120953 0.02700985  0.01436459  1.00000000 -0.318450715
age        0.02524542 -0.004947199 0.07343156 -0.08372895 -0.31845072  1.000000000



lowerCor(Tal_Or) #round the results to two decimals and abbreviate the names
         cond  pmi   imprt rectn gendr age
cond      1.00
pmi       0.18  1.00
import    0.18  0.28  1.00
reaction  0.16  0.45  0.46  1.00
gender   -0.13 -0.02  0.03  0.01  1.00
age       0.03  0.00  0.07 -0.08 -0.32  1.00
>
> corr.test(Tal_Or) # find the correlations, the raw p values and the adjusted p values
Call:corr.test(x = Tal_Or)
Correlation matrix
          cond   pmi import reaction gender   age
cond      1.00  0.18   0.18     0.16  -0.13  0.03
pmi       0.18  1.00   0.28     0.45  -0.02  0.00
import    0.18  0.28   1.00     0.46   0.03  0.07
reaction  0.16  0.45   0.46     1.00   0.01 -0.08
gender   -0.13 -0.02   0.03     0.01   1.00 -0.32
age       0.03  0.00   0.07    -0.08  -0.32  1.00
Sample Size
[1] 123
Probability values (Entries above the diagonal are adjusted for multiple tests.)
```

```
          cond  pmi import reaction gender age
cond      0.00 0.50   0.50     0.69      1   1
pmi       0.05 0.00   0.02     0.00      1   1
import    0.05 0.00   0.00     0.00      1   1
reaction  0.08 0.00   0.00     0.00      1   1
gender    0.16 0.82   0.77     0.87      0   0
age       0.78 0.96   0.42     0.36      0   0

 To see confidence intervals of the correlations, print with the short=FALSE option
>
>  cor.ci(Tal_Or[1:4], n.iter=1000)
Call:corCi(x = x, keys = keys, n.iter = n.iter, p = p, overlap = overlap,
    poly = poly, method = method, plot = plot, minlength = minlength)

 Coefficients and bootstrapped confidence intervals
          cond pmi  imprt rectn
cond      1.00
pmi       0.18 1.00
import    0.18 0.28 1.00
reaction  0.16 0.45 0.46  1.00

 scale correlations and bootstrapped confidence intervals
            lower.emp lower.norm estimate upper.norm upper.emp    p
cond-pmi         0.02       0.02     0.18       0.35      0.34 0.03
cond-imprt       0.00       0.00     0.18       0.35      0.34 0.05
cond-rectn      -0.02      -0.02     0.16       0.33      0.32 0.09
pmi-imprt        0.10       0.11     0.28       0.44      0.43 0.00
pmi-rectn        0.30       0.30     0.45       0.57      0.58 0.00
imprt-rectn      0.32       0.31     0.46       0.60      0.59 0.00
 To see confidence intervals of the correlations, print with the short=FALSE option
> cor2latex(Tal_Or[1:4],stars=TRUE,adjust="none")  #create the Table
.... omitted
```

Mediation and Moderation

Mediation is just a different way of thinking of regression. It can be done using the `mediate` function. The first example just shows the regression analysis and draws the figure, The second example adds pmi and import as mediators. Compare the two outputs. See Figure 3.

R code

```
reg <-  mediate(reaction ~   pmi +cond +  import,data=Tal_Or)
moderate.diagram(reg,main="Regression")
reg

med <-  mediate(reaction ~ cond + (pmi)+ (import),data=Tal_Or)
print(med,short=FALSE)
```

```
> reg <-  mediate(reaction ~   pmi +cond +  import,data=Tal_Or)
> moderate.diagram(reg,main="Regression")
> reg

Mediation/Moderation Analysis
```

```
Call: mediate(y = reaction ~ pmi + cond + import, data = Tal_Or)

The DV (Y) was  reaction . The IV (X) was  pmi cond import . The mediating variable(s) =  .
 DV =  reaction
       slope   se    t       p
pmi     0.40 0.09 4.26 4.0e-05
cond    0.10 0.24 0.43 6.7e-01
import  0.32 0.07 4.59 1.1e-05

With R2 =  0.33
R = 0.57 R2 = 0.33   F = 19.11 on 3 and 119 DF   p-value:  3.5e-10
>
> med <-  mediate(reaction ~ cond + (pmi)+ (import),data=Tal_Or)
> print(med,short=FALSE)

Mediation/Moderation Analysis
Call: mediate(y = reaction ~ cond + (pmi) + (import), data = Tal_Or)

The DV (Y) was  reaction . The IV (X) was  cond . The mediating variable(s) =  pmi import .

Total effect(c) of  cond  on  reaction  =  0.5   S.E. =  0.28  t  =  1.79  df=  119
                    with p =  0.077
Direct effect (c') of  cond  on  reaction  removing  pmi import  =  0.1
                    S.E. =  0.24  t  =  0.43  df=  119   with p =  0.67
Indirect effect (ab) of  cond  on  reaction  through  pmi import   =  0.39
Mean bootstrapped indirect effect =  0.4  with standard error =  0.17
                                              Lower CI =  0.09    Upper CI =  0.73
R = 0.57 R2 = 0.33   F = 19.11 on 3 and 119 DF   p-value:  3.5e-10


 Full output

 Total effect estimates (c)
     reaction   se    t  df   Prob
cond      0.5 0.28 1.79 119 0.0766

Direct effect estimates     (c')
       reaction   se    t  df     Prob
cond       0.10 0.24 0.43 119 6.66e-01
pmi        0.40 0.09 4.26 119 4.04e-05
import     0.32 0.07 4.59 119 1.13e-05

 'a'  effect estimates
       cond   se    t  df   Prob
pmi    0.48 0.24 2.02 121 0.0454
import 0.63 0.31 2.02 121 0.0452

 'b'  effect estimates
       reaction   se    t  df     Prob
pmi        0.40 0.09 4.26 119 4.04e-05
import     0.32 0.07 4.59 119 1.13e-05

 'ab'  effect estimates
     reaction boot   sd lower upper
cond     0.39  0.4 0.17  0.09  0.73

 'ab' effects estimates for each mediator
      pmi boot   sd lower upper
```

```
cond 0.19 0.19 0.11  0.01  0.42
     import boot   sd lower upper
cond    0.2  0.2 0.11  0.01  0.45
```

To show moderation, we use the Garcia et al. (2010) data set. We use the `scale` and `lm` functions from Core-R to do the regressions. We compare the mean centered versus non-mean centered results. Then we use `setCor` to combine these two steps. We include a demonstration of how to create the interaction plot of Figure 3

R code

```
#First do the regular linear model
mod1 <- lm(respappr ~ prot2 * sexism ,data=Garcia)  #do not mean center
centered <- scale(Garcia,scale=FALSE) #mean center, do not standardize
centered.df <- data.frame(centered)   #convert to a data frame
mod.centered <- lm(respappr ~ prot2 * sexism ,data=centered.df)
summary(mod1)   #the uncentered model
summary(mod.centered) #the centered model

par(mfrow=c(1,2))
#compare two models  (bootstrapping n.iter set to 5000 by default
# 1) mean center the variables prior to taking product terms
mod <- setCor(respappr ~ prot2 * sexism ,data=Garcia,
 ,main="A:  Moderated regression (std. and mean centered)")
mod
#demonstrate interaction plots
plot(respappr ~ sexism, pch = 23- protest, bg = c("black","red", "blue")[protest],
data=Garcia, main = "B:  Response to sexism varies as type of protest")
by(Garcia,Garcia$protest, function(x) abline(lm(respappr ~ sexism,
   data =x),lty=c("solid","dashed","dotted")[x$protest+1]))
text(6.5,3.5,"No protest")
text(3.1,3.9,"Individual")
text(3.1,5.2,"Collective")
```

```
> summary(mod1)   #the uncentered model

Call:
lm(formula = respappr ~ prot2 * sexism, data = Garcia)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4984 -0.7540  0.0801  0.8301  3.1853

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.5667     1.2095   5.429 2.83e-07 ***
prot2         -2.6866     1.4515  -1.851  0.06654 .
sexism        -0.5290     0.2359  -2.243  0.02668 *
prot2:sexism   0.8100     0.2819   2.873  0.00478 **
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 1.144 on 125 degrees of freedom
Multiple R-squared:  0.2962,        Adjusted R-squared:  0.2793
```

```
F-statistic: 17.53 on 3 and 125 DF,  p-value: 1.456e-09

> summary(mod.centered) #the centered model

Call:
lm(formula = respappr ~ prot2 * sexism, data = centered.df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4984 -0.7540  0.0801  0.8301  3.1853

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.01184    0.10085  -0.117  0.90671
prot2         1.45803    0.21670   6.728 5.52e-10 ***
sexism        0.02354    0.12927   0.182  0.85579
prot2:sexism  0.80998    0.28191   2.873  0.00478 **
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 1.144 on 125 degrees of freedom
Multiple R-squared:  0.2962,        Adjusted R-squared:  0.2793
F-statistic: 17.53 on 3 and 125 DF,  p-value: 1.456e-09


> #compare two models  (bootstrapping n.iter set to 5000 by defalt
> # 1) mean center the variables prior to taking product terms
> mod <- setCor(respappr ~ prot2 * sexism ,data=Garcia,
+  ,main="A:  Moderated regression (std. and mean centered)")
> mod
Call: setCor(y = respappr ~ prot2 * sexism, data = Garcia,
            main = "A:  Moderated regression (std. and mean centered)")

Multiple Regression from raw data

 DV =  respappr
             slope   se    t       p VIF
prot2         0.51 0.08 6.73 5.5e-10   1
sexism        0.01 0.08 0.18 8.6e-01   1
prot2*sexism  0.22 0.08 2.87 4.8e-03   1

 Multiple Regression
            R  R2  Ruw R2uw Shrunken R2 SE of R2 overall F df1 df2        p
respappr 0.54 0.3 0.42 0.18        0.28     0.06     17.53   3 125 1.46e-09
> #demonstrate interaction plots
> plot(respappr ~ sexism, pch = 23- protest, bg = c("black","red", "blue")[protest],
+ data=Garcia, main = "B:  Response to sexism varies as type of protest")
> by(Garcia,Garcia$protest, function(x) abline(lm(respappr ~ sexism,
+    data =x),lty=c("solid","dashed","dotted")[x$protest+1]))
Garcia$protest: 0
NULL
------------------------------------------------------------------------------------------------------------------------
Garcia$protest: 1
NULL
------------------------------------------------------------------------------------------------------------------------
Garcia$protest: 2
NULL
> text(6.5,3.5,"No protest")
```

```
> text(3.1,3.9,"Individual")
> text(3.1,5.2,"Collective")
>
```

Decision theory and Area under the curve

Table 6 and Figure 4 are example of signal detection theory. This is done by giving the four cells to the `AUC` function.

R code

```
AUC(c(49,40,79,336))
```

```
Decision Theory and Area under the Curve

The original data implied the following 2 x 2 table
         Predicted.Pos Predicted.Neg
True.Pos         0.097         0.079
True.Neg         0.157         0.667

Conditional probabilities of
         Predicted.Pos Predicted.Neg
True.Pos          0.55          0.45
True.Neg          0.19          0.81

Accuracy =  0.76  Sensitivity =  0.55   Specificity =  0.81
with Area Under the Curve =  0.76
d.prime =  1  Criterion =  0.88  Beta =  0.15
Observed Phi correlation =  0.32
Inferred latent (tetrachoric) correlation  =  0.53
>
```

EFA

The factor analysis of the `Thurstone` data set was done using the `fa` function. We specify that the number of subjects was 213. By default, we find a *minres* solution and use the `oblminin` rotation. We also show how to specify other factor extraction techniques, and other rotations. We just show the first solution.

R code

```
fa(Thurstone,nfactors=3,n.obs=213)
fa(Thurstone,nfactors=3,n.obs=213,fm="mle")  #use the maximum likelihood algorithm
fa(Thurstone,nfactors=3,n.obs=213, rotate="Varimax") #use an orthogonal rotation.
```

```
> fa(Thurstone,nfactors=3,n.obs=213)
Factor Analysis using method =  minres
Call: fa(r = Thurstone, nfactors = 3, n.obs = 213)
Standardized loadings (pattern matrix) based upon correlation matrix
                    MR1   MR2   MR3   h2   u2 com
Sentences          0.90 -0.03  0.04 0.82 0.18 1.0
Vocabulary         0.89  0.06 -0.03 0.84 0.16 1.0
```

```
Sent.Completion    0.84  0.03  0.00 0.74 0.26 1.0
First.Letters      0.00  0.85  0.00 0.73 0.27 1.0
Four.Letter.Words -0.02  0.75  0.10 0.63 0.37 1.0
Suffixes           0.18  0.63 -0.08 0.50 0.50 1.2
Letter.Series      0.03 -0.01  0.84 0.73 0.27 1.0
Pedigrees          0.38 -0.05  0.46 0.51 0.49 2.0
Letter.Group      -0.06  0.21  0.63 0.52 0.48 1.2

                       MR1  MR2  MR3
SS loadings           2.65 1.87 1.49
Proportion Var        0.29 0.21 0.17
Cumulative Var        0.29 0.50 0.67
Proportion Explained  0.44 0.31 0.25
Cumulative Proportion 0.44 0.75 1.00

 With factor correlations of
     MR1  MR2  MR3
MR1 1.00 0.59 0.53
MR2 0.59 1.00 0.52
MR3 0.53 0.52 1.00

Mean item complexity =  1.2
Test of the hypothesis that 3 factors are sufficient.

The degrees of freedom for the null model are  36  and the objective function was  5.2 with
                  Chi Square of  1081.97
The degrees of freedom for the model are 12  and the objective function was  0.01

The root mean square of the residuals (RMSR) is  0.01
The df corrected root mean square of the residuals is  0.01

The harmonic number of observations is  213 with the empirical chi square  0.52  with prob <  1
The total number of observations was  213  with Likelihood Chi Square =  2.98  with prob <  1

Tucker Lewis Index of factoring reliability =  1.026
RMSEA index =  0  and the 90 % confidence intervals are  0 0
BIC =  -61.36
Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                 MR1  MR2  MR3
Correlation of (regression) scores with factors   0.96 0.92 0.90
Multiple R square of scores with factors          0.93 0.85 0.82
Minimum correlation of possible factor scores     0.86 0.71 0.63
```

Reliability

Here we find the reliability of the `msqR` items found in the first example. We select just the time 1 data. We show several different approaches. Because we have just 8 items and they represent two subfactors, we find ω_h using a two factor solution.

R code

```
msq.items <- c("anxious" ,  "at.ease" ,  "calm" , "confident", "content",
 "jittery",  "nervous", "relaxed" , "tense" , "upset" ) #these overlap with the sai
msq1 <- subset(msqR,msqR$time==1)
alpha(msq1[msq.items], check.keys=TRUE)
```

```
omega(msq1[msq.items], nfactors=2)
```

```
 alpha(msq1[msq.items], check.keys=TRUE)

Reliability analysis
Call: alpha(x = msq1[msq.items], check.keys = TRUE)

  raw_alpha std.alpha G6(smc) average_r S/N    ase mean   sd median_r
      0.83      0.83    0.86      0.33   5 0.0046    2 0.54     0.32

 lower alpha upper     95% confidence boundaries
0.82 0.83 0.84

 Reliability if an item is dropped:
          raw_alpha std.alpha G6(smc) average_r S/N alpha se var.r med.r
anxious-       0.83      0.83    0.85      0.34 4.7   0.0047 0.026  0.34
at.ease        0.80      0.80    0.83      0.31 4.1   0.0055 0.028  0.32
calm           0.80      0.81    0.84      0.32 4.2   0.0054 0.030  0.32
confident      0.83      0.83    0.85      0.36 5.0   0.0046 0.022  0.32
content        0.82      0.82    0.84      0.34 4.6   0.0049 0.025  0.32
jittery-       0.83      0.83    0.85      0.35 4.8   0.0047 0.027  0.33
nervous-       0.82      0.82    0.84      0.33 4.4   0.0049 0.030  0.32
relaxed        0.80      0.81    0.84      0.31 4.1   0.0055 0.030  0.31
tense-         0.81      0.81    0.83      0.32 4.2   0.0051 0.029  0.32
upset-         0.82      0.82    0.85      0.34 4.7   0.0049 0.033  0.35

 Item statistics
             n raw.r std.r r.cor r.drop mean   sd
anxious-  1871  0.54  0.56  0.51   0.42  2.3 0.86
at.ease   3018  0.77  0.74  0.72   0.67  1.6 0.94
calm      3020  0.74  0.71  0.68   0.63  1.6 0.92
confident 3021  0.54  0.50  0.43   0.38  1.5 0.93
content   3010  0.64  0.59  0.55   0.50  1.4 0.92
jittery-  3026  0.52  0.55  0.48   0.41  2.3 0.83
nervous-  3017  0.59  0.64  0.60   0.52  2.6 0.68
relaxed   3023  0.76  0.73  0.70   0.66  1.6 0.91
tense-    3017  0.67  0.71  0.69   0.60  2.4 0.78
upset-    3019  0.54  0.58  0.50   0.45  2.6 0.68

Non missing response frequency for each item
             0    1    2    3 miss
anxious   0.53 0.29 0.13 0.04 0.38
at.ease   0.14 0.33 0.35 0.18 0.00
calm      0.14 0.34 0.36 0.17 0.00
confident 0.16 0.33 0.37 0.14 0.00
content   0.17 0.35 0.35 0.13 0.01
jittery   0.54 0.31 0.12 0.04 0.00
nervous   0.70 0.22 0.06 0.02 0.00
relaxed   0.12 0.30 0.40 0.18 0.00
tense     0.59 0.28 0.10 0.03 0.00
upset     0.74 0.18 0.05 0.02 0.00
Warning message:
In alpha(msq1[msq.items], check.keys = TRUE) :
  Some items were negatively correlated with total scale and were automatically reversed.
 This is indicated by a negative sign for the variable name.
> omega(msq1[msq.items], nfactors=2)
```

```
Three factors are required for identification -- general factor loadings set to be equal.
Proceed with caution.
Think about redoing the analysis with alternative values of the 'option' setting.

Omega
Call: omega(m = msq1[msq.items], nfactors = 2)
Alpha:                 0.83
G.6:                   0.86
Omega Hierarchical:    0.45
Omega H asymptotic:    0.51
Omega Total            0.87

Schmid Leiman Factor loadings greater than  0.2
             g   F1*   F2*   h2   u2   p2
anxious-  0.36       -0.57 0.46 0.54 0.28
at.ease   0.52  0.59       0.64 0.36 0.43
calm      0.49  0.47 -0.21 0.51 0.49 0.48
confident 0.31  0.58       0.46 0.54 0.21
content   0.40  0.65       0.59 0.41 0.26
jittery-  0.35       -0.52 0.40 0.60 0.31
nervous-  0.43       -0.57 0.51 0.49 0.36
relaxed   0.51  0.48 -0.22 0.53 0.47 0.48
tense-    0.50       -0.62 0.63 0.37 0.39
upset-    0.35       -0.29 0.25 0.75 0.50

With eigenvalues of:
  g F1* F2*
1.8 1.6 1.5

general/max  1.13   max/min =   1.05
mean percent general =  0.37    with sd =  0.1 and cv of  0.28
Explained Common Variance of the general factor =  0.37

The degrees of freedom are 26  and the fit is  0.24
The number of observations was  3032  with Chi Square =  721.36  with prob <  2.4e-135
The root mean square of the residuals is  0.04
The df corrected root mean square of the residuals is  0.05
RMSEA index =  0.094  and the 10 % confidence intervals are  0.088 0.1
BIC =  512.92

Compare this with the adequacy of just a general factor and no group factors
The degrees of freedom for just the general factor are 35  and the fit is  1.67
The number of observations was  3032  with Chi Square =  5055.64  with prob <  0
The root mean square of the residuals is  0.21
The df corrected root mean square of the residuals is  0.24

RMSEA index =  0.218  and the 10 % confidence intervals are  0.213 0.223
BIC =  4775.04

Measures of factor score adequacy
                                                 g  F1*  F2*
Correlation of scores with factors             0.67 0.77 0.76
Multiple R square of scores with factors       0.45 0.60 0.59
Minimum correlation of factor score estimates -0.09 0.19 0.17

 Total, General and Subset omega for each subset
                                                g  F1*  F2*
```

```
Omega total for total scores and subscales    0.87 0.84 0.79
Omega general for total scores and subscales  0.45 0.33 0.30
Omega group for total scores and subscales    0.36 0.51 0.49
```