Online supplement to Reliability from α to ω: A Tutorial

William Revelle and David Condon
Northwestern University and University of Oregon

Abstract

Elaborations on concepts in the original manuscript by Revelle and Condon. A flow chart and example code in R allow the interested reader to test out the various techniques for measuring reliability.

This supplement is in two parts: In the first we include several derivations and examples that are too long for the accompanying manuscript; In the second we provide a detailed set of examples in the open source statistical system R (R Core Team, 2019).These example data sets and the various functions are included in the *psych* (Revelle, 2019a) and *psychTools* packages (Revelle, 2019b) for R.

Traits and States over time

If we have only two time points, the correlation of the same test given twice is an unknown mixture of trait and state effects:

$$r_{t_1 t_2} = \frac{\sigma_T^2 + \tau^{t_2 - t_1}\sigma_S^2 + \sigma_s^2}{\sigma_T^2 + \sigma_S^2 + \sigma_s^2 + \sigma_e^2} \tag{1}$$

where τ (the auto-correlation due to short term state consistency) is less than 1 and thus the state effect ($\tau^{t_2 - t_1}\sigma_S^2$) will become smaller the greater the time lag. If the intervening time is long enough that the State effect is minimal, we will still have specific variance, and the correlation of a test with the same test later is

$$r_{xx} = \frac{\sigma_T^2 + \sigma_s^2}{V_x} = \frac{\sigma_T^2 + \sigma_s^2}{\sigma_T^2 + \sigma_S^2 + \sigma_s^2 + \sigma_e^2}. \tag{2}$$

As discussed in the article, the trait/state distinction can be estimated if we have at least three time points, although four or more is even better (Figure 1).

The MSQ and SAI data sets

Included in the *psychTools* package are several data sets including the `msqR` and the `sai` which we will use in the subsequent examples. The Motivational State Questionnaire (MSQ) (Revelle &

contact: William Revelle revelle@northwestern.edu
Supplementary material from a manuscript on Reliability (Revelle and Condon, 2019) Final submitted version of May 10, 2019

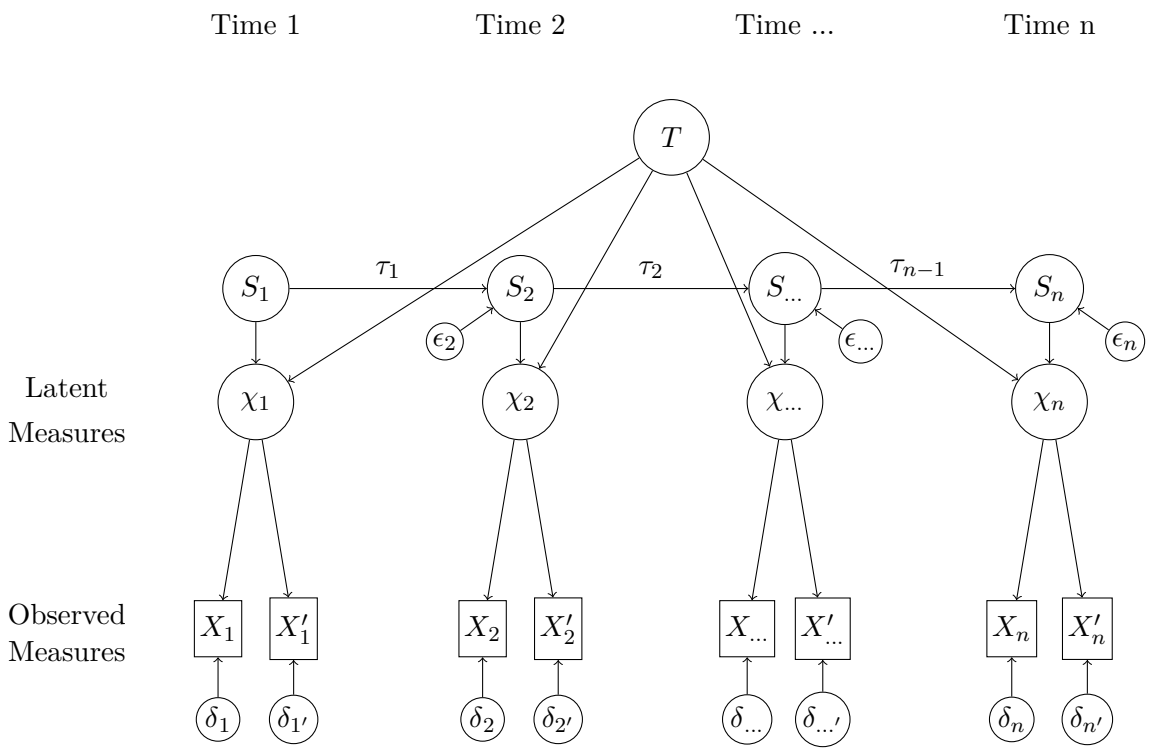

Figure 1. How traits and states affect short term and long term reliability. Observed scores (X_1 ... X_n) of the same trait measured over time reflect a latent Trait, time varying States (S_1 ... S_n) and measurement errors (δ_1 ... $\delta_{n'}$). The latent states are assumed to be auto-correlated (τ_i), such that over longer time periods the correlation will tend towards 0 ($\tau^n \rightarrow 0$) (see Equation 1). Adapted from (Cole et al., 2005).

Anderson, 1998) included 75 items taken from a number of mood questionnaires (e.g., Thayer, 1978; Larsen & Diener, 1992; Watson et al., 1988) and had 10 anxiety items that overlapped with the state version of the State Trait Anxiety Inventory (Spielberger et al., 1970). The MSQ and STAI were given before any motivational manipulations were given, and then sometimes given at the end of the experimental session. We will use these 10 items in the subsequent examples evaluating and comparing the immediate dependability and the 45 minute and multi-day stability coefficients of these measures. These 10 items were given as part of the STAI and then immediately given again (with a slightly different wording) as part of the MSQ. Five of the items were scored in a positive (anxious) direction, five in the negative (non-anxious) direction. The STAI items were given on a 1-4 scale, the MSQ on 0 to 3 scale. An additional 20 items of the STAI were given with trait instructions and are reported as the `tai` dataset. The movie manipulation used 7-9 minute film clips to induce fear (Halloween), depression (concentration camp), happiness (Parenthood), and a control movie (a nature film). The caffeine condition contrasted 4 mg/kg body weight of caffeine to a placebo.

Comparing reliability estimates

Table 1 integrates findings from a number of different procedures (detailed code in the R section) for finding reliability. The first set of these take advantage of the repeated nature of the `sai` and `msqR` data sets. The subsequent estimates of internal consistency are done with the

first administration of these state measures as well as trait measures of anxiety, impulsivity, and sociability.

As is seen in Table 1, the immediate correlations for the 10 item state scales (test dependability) was .85. As expected, the 45 minute correlations were smaller (.55-.76) and those with one day delay were smaller yet (ranging from .36 to .39) with a mean of .38. This is in contrast to the immediate correlations of state with trait measures (.48) and after two days (.43) suggesting that the state measure has a trait component (or that the trait measure has a state component). The state retest measures were also much lower than the retest correlations of the EPI Impulsivity (.70) and Sociability (.81) subscales.

Test-retest reliability and components of variance

To compare the effects of an immediate retest versus a short delay versus a somewhat longer delay, consider the `msqR`, `sai` and `tai` example data sets from the *psychTools* package (Revelle, 2019b) in the R open statistical system (R Core Team, 2019); the analyses discussed below are demonstrated in this supplementary online material. The data ($N = 3,032$) were collected as part of a long term series of studies of the interrelationships between the stable personality dimensions of extraversion, neuroticism, and impulsivity (e.g., Eysenck & Eysenck, 1964; Revelle et al., 1980), situational stressors (caffeine, time of day, movie induced affect, e.g., Anderson & Revelle, 1983, 1994; Rafaeli et al., 2007; Rafaeli & Revelle, 2006) momentary affective and motivational state (e.g. energetic and tense arousal (Thayer, 1978, 1989), state anxiety (Spielberger et al., 1970)), and cognitive performance (Humphreys & Revelle, 1984).

A powerful advantage of repeating items is that it allows for an assessment of subject consistency across time (the correlation for each subject of their pattern of responses across the two administrations) as well as the consistency of the items (the correlation across subjects of responses to each item) (DeSimone, 2015; Wood et al., 2017). This allows for identification of unstable items and inconsistent responders. In addition, by using multi-level analyses[1] it is possible to estimate the variance components due to people, items, the person x item interaction, time, the person x time interaction, and the residual (error) variance (DeSimone, 2015; Revelle & Wilt, 2019; Shrout & Lane, 2012). This is implemented for example as the `testRetest` function in the *psych* package. The responses to any particular item can be thought to represent multiple sources of variance, and the reliability of a test made up of items is thus a function of those individual sources of variance. If we let P_i represent the i_{th} person, I_j the j_{th} item, T_k the first or second administration of the item, then the response to any item (e.g., and anxiety item from the STAI) is a function of the mean anxiety item (μ), the particular subjects level of anxiety (P_i), the particular anxiety content of the item (I_j, e.g. feeling tense has a higher mean score than not feeling relaxed), the time or form administered (T_k, e.g. if the same item is administered twice, what is the effect of time, or if an equivalent item, of form), and the interactions of all of these terms.

$$X_{ijk} = \mu + P_i + I_j + T_k + P_iI_j + P_iT_k + I_jT_k + P_iI_jT_k + \epsilon. \tag{3}$$

With complete data, we can find these components using conventional repeated measures analysis of variance of the data (i.e.., `aov` in core R) or using multi-level functions such as `lmer` in the *lme4* package (Bates et al., 2015) for R. As an example of such a variance decomposition consider the 10 overlapping mood items in the STAI and MSQ discussed here (Table 2). 19% of the

[1] Analytic strategies for analyzing such multi-level data have been given different names in a variety of fields and are known by a number of different terms such as the random effects or random coefficient models of economics, multi-level models of sociology and psychology, hierarchical linear models of education or more generally, mixed effects models (Fox, 2016).

Table 1

Comparing multiple estimates of reliability. SAI and MSQ contained 10 items measuring anxious vs. calm mood. Test-retest values include very short term (dependability) and longer term (stability) measures. Short term dependabilities are of mood measures pre and post various mood manipulations. One- to two-day delay stabilities are mood measures pre any mood manipulation. Dependability measures are based upon these 10 items given in the same session although using two different questionnaires. Trait anxiety was given with "how you normally feel", state anxiety asked "do you feel ". The two-four week delay compares personality trait measures (Impulsivity and Sociability) given as part of a group testing session and then part of various experimental sessions. Internal consistency estimates for α and ω do not require retesting. When giving the test twice, it is possible to find the consistency of each item (r_{ii}).The particular functions for estimating these coefficients are all part of the psych package.

	State		Trait				estimation
	SAI	MSQ	SAI	TAI	EPI		functions
Types of reliability			MSQ	(SAI)	Imp	Soc	
Test-retest							
Short term (test dependability)							
45 minutes (control)	.76	.74					`testRetest`
45 minutes (caffeine)	.73	.74					`testRetest`
45 minutes (films)	.57	.55					`testRetest`
Longer delay (stability)							
1-2 days	.36	.39		.43*			`testRetest`
1-4 weeks					.70	.81	
Average r_{ii} over time (item dependability)							
45 minutes (control)	.60	.57					`testRetest`
45 minutes (caffeine)	.61	.58					`testRetest`
45 minutes (film)	.39	.40					`testRetest`
1-2 days	.29	.30					`testRetest`
1-4 weeks					.52	.56	`testRetest`
Parallel form approach							
Parallel tests	.74	.74		.48*			`scoreItems`
Duplicated tests (test dependability)			.85				`testRetest`
average r_{ii}(item dependability)			.67				`testRetest`
Internal consistency							
greatest split half (λ_4)	.92	.89		.94	.61	.83	`splitHalf`
ω_t	.90	.87		.92	.62	.80	`omega`
SMC adjusted (λ_6)	.89	.86		.92	.52	.78	`splitHalf`
α (λ_3)	.87	.83		.90	.51	.76	`alpha`
average split half	.86	.83		.90	.50	.76	`splitHalf`
ω_g	.57	.45		.67	.31	.62	`omega`
smallest split half	.66	.57		.79	.41	.66	`splitHalf`
worst split half (β)	.67	.57		.50	.05	.27	`iclust`
average r	.39	.33		.32	.11	.19	`alpha`
Other forms of reliability							
ICC	.87	.83		.90	.51	.76	`ICC`
kappa							

*Trait Anxiety x State Anxiety

variance of the anxiety scores was due to between person variability, 25% to the very short period of time, 19% to the interaction of person by time, etc. and 13% was residual (unexplained) variance. From these components of variance, we can find several different reliability estimates (Cranford et al., 2006; Shrout & Lane, 2012). The first is the reliability of the total score for each individual across the 10 overlapping items if the test is thought of as composed of those (fixed) 10 items.

$$R_{1F} = \frac{\sigma^2_{Person} + \frac{\sigma^2_{Person\ x\ Item}}{k}}{\sigma^2_{Person} + \frac{\sigma^2_{Person\ x\ Item}}{k} + \frac{\sigma^2_{Residual}}{k}} \tag{4}$$

The second is the reliability of the average of the two measurement occasions[2] this is

$$R_{kF} = \frac{\sigma^2_{Person} + \frac{\sigma^2_{Person\ x\ Item}}{k}}{\sigma^2_{Person} + \frac{\sigma^2_{Person\ x\ Item}}{k} + \frac{\sigma^2_{Residual}}{2k}}. \tag{5}$$

Additional estimates can be found for the reliability of a single item (R_{1R}) or the average of an item across both time points (R_{2R}) (Shrout & Lane, 2012).

Table 2
A variance decomposition of the 10 overlapping items from the STAI and MSQ measures (N= 200). Data taken from the example for the `testRetest` *function in the psych package. Four different variance ratios and reliabilities may be found from these (Cranford et al., 2006; Shrout & Lane, 2012). Time is confounded by a subtle change in item wording from the STAI to the MSQ. For the specific code to do this example, see the R section below.*

Multilevel components of variance due to

	Variance	Percent
Subjects	0.34	0.19
Time	0.44	0.25
Items	0.17	0.10
Subjects x time	0.24	0.13
Subjects x items	0.01	0.01
time x items	0.34	0.19
Residual	0.23	0.13
Total	1.78	1.00

This leads to four reliability estimates:

R_{1F}	= 0.94	Reliability of average of all items for one time (Random time effects)
R_{kF}	= 0.97	Reliability of average of all ratings across all items and times (Fixed time effects)
R_{1R}	= 0.33	Generalizability of a single time point across all items (Random time effects)
R_{2R}	= 0.49	Generalizability of average time points across all items (Fixed time effects)

Split Half reliability

In order to understand these procedures, it is useful to think about what goes into the correlation between two tests or two times. Consider two tests, $\mathbf{X}$ and $\mathbf{X'}$, both made up of two subtests. The simple correlation $r_{xx'} = \frac{C_{xx'}}{\sqrt{V_x V_{x'}}}$ may be expressed in terms of elements (items) in

[2]Functionally, just the Spearman-Brown prophecy formula applied to R_{1F}. This will be discussed later in terms of split half reliability.

X and X'[3] The reliability of $\mathbf{X}$ is just its correlation with $\mathbf{X'}$ and can be thought of in terms of the elements of the variance-covariance matrix, $\Sigma_{XX'}$:

$$\Sigma_{XX'} = \begin{pmatrix} \mathbf{V_x} & \vdots & \mathbf{C_{xx'}} \\ \cdots & \cdots & \cdots \\ \mathbf{C_{xx'}} & \vdots & \mathbf{V_{x'}} \end{pmatrix} \tag{6}$$

and letting $V_{\mathbf{x}} = \mathbf{1'V_x1}$ and $C_{\mathbf{XX'}} = \mathbf{1'C}_{XX'}\mathbf{1}$ where $\mathbf{1}$ is a column vector of 1s and $\mathbf{1'}$ is its transpose, the correlation between the two tests will be

$$\rho_{xx'} = \frac{C_{xx'}}{\sqrt{V_x V_{x'}}}.$$

But the variance of a test is simply the sum of the true covariances and the error variances and we can break up each test ($\mathbf{X}$ and $\mathbf{X'}$) into their individual items ($\mathbf{x_1}$ $\mathbf{x_2}$ ... $\mathbf{x_i}$) and their respective variances and covariances. We can split each test into two parts and then the structure of the two tests seen in Equation 6 becomes

$$\Sigma_{XX'} = \left(\begin{array}{ccc|ccc} \mathbf{V_{x_1}} & \vdots & \mathbf{C_{x_1x_2}} & \mathbf{C_{x_1x'_1}} & \vdots & \mathbf{C_{x_1x'_2}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{C_{x_1x_2}} & \vdots & \mathbf{V_{x_2}} & \mathbf{C_{x_2x'_1}} & \vdots & \mathbf{C_{x_2x'_2}} \\ \hline \mathbf{C_{x_1x'_1}} & \vdots & \mathbf{C_{x_2x'_1}} & \mathbf{V_{x'_1}} & \vdots & \mathbf{C_{x'_1x'_2}} \\ \mathbf{C_{x_1x'_2}} & \vdots & \mathbf{C_{x_2x'_2}} & \mathbf{C_{x'_1x'_2}} & \vdots & \mathbf{V_{x'_2}} \end{array}\right) \tag{7}$$

But what if we don't have two tests? We need to make assumptions about the structure of what the covariance between a test ($\mathbf{X}$) and a test just like it ($\mathbf{X'}$) would be based upon what we know about test $\mathbf{X}$.

Because the splits are done at random and the second test is parallel with the first test, the expected covariances between splits are all equal to the true score variance of one split ($\mathbf{V_{t_1}}$), and the variance of a split is the sum of true score and error variances:

$$\Sigma_{XX'} = \left(\begin{array}{ccc|ccc} \mathbf{V_{t_1}} + \mathbf{V_{e_1}} & \vdots & \mathbf{V_{t_1}} & \mathbf{V_{t_1}} & \vdots & \mathbf{V_{t_1}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{V_{t_1}} & \vdots & \mathbf{V_{t_1}} + \mathbf{V_{e_1}} & \mathbf{V_{t_1}} & \vdots & \mathbf{V_{t_1}} \\ \hline \mathbf{V_{t_1}} & \vdots & \mathbf{V_{t_1}} & \mathbf{V_{t'_1}} + \mathbf{V_{e'_1}} & \vdots & \mathbf{V_{t'_1}} \\ \mathbf{V_{t_1}} & \vdots & \mathbf{V_{t_1}} & \mathbf{V_{t'_1}} & \vdots & \mathbf{V_{t'_1}} + \mathbf{V_{e'_1}} \end{array}\right)$$

The correlation between a test made up of two halves with intercorrelation ($r_1 = V_{t_1}/V_{x_1}$) with another such test is

$$r_{xx'} = \frac{4V_{t_1}}{\sqrt{(4V_{t_1} + 2V_{e_1})(4V_{t_1} + 2V_{e_1})}} = \frac{4V_{t_1}}{2V_{t_1} + 2V_{x_1}} = \frac{4r_1}{2r_1 + 2}$$

[3] We use **boldface** to represent matrices and vectors. The operation $\mathbf{1'C_{XX'}1}$ is the matrix equivalent of summing all the elements of the matrix $\mathbf{C_{XX}}$.

and thus

$$r_{xx'} = \frac{2r_1}{1 + r_1}. \tag{8}$$

There are a number of different approaches for estimating reliability when there is just one test and one time. The earliest was to split the test into two random split halves and then adjust the resulting correlation between these two splits using the Spearman-Brown prophecy formula (Brown, 1910; Spearman, 1910):

$$\frac{2 * r}{1 + r}. \tag{9}$$

Considering items as measuring ability (e.g., right with probability p and wrong with probability $q = 1 - p$), Kuder & Richardson (1937) proposed several estimates of the reliability of the average split half, with their most well known being their 20th equation (and thus known as KR20).

$$r_{tt} = \frac{n}{n-1}\frac{\sigma_t^2 - n\overline{pq}}{\sigma_t^2}. \tag{10}$$

Multilevel Reliability

Table 3
An abbreviated data set (adapted from Shrout & Lane (2012). Four subjects give responses to three items over four time points. See also Figure 2. Taken from the help file for the `multilevel.reliability` *function. One subject is deleted from the data set for simplicity.*

Variable	Person	Time	Item1	Item2	Item3
1	1	1	3	4	4
2	2	1	6	6	5
3	3	1	3	4	3
4	4	1	7	8	7
5	1	2	5	7	7
6	2	2	6	7	8
7	3	2	3	5	9
8	4	2	8	8	9
9	1	3	4	6	7
10	2	3	7	8	9
11	3	3	5	6	7
12	4	3	6	7	8
13	1	4	5	9	7
14	2	4	8	9	9
15	3	4	8	7	9
16	4	4	6	8	6

Shrout & Lane (2012) discuss six reliability (or generalizability) coefficients that may be found from these variance components. Although the next six equations are probably challenging, it might be a relief to realize that the resulting estimates are all reported in the `multilevel.reliablity` function.

If we are interested in how stable the between person differences are when averaged over all the (m) items and all (k) occasions, then we need to compare the variance due to people and to people by items to those variances plus error variance:

$$R_{kF} = \frac{\sigma_p^2 + \frac{\sigma_{p*i}^2}{m}}{\sigma_p^2 + \frac{\sigma_{p*i}^2}{m} + \frac{\sigma_e^2}{km}}. \tag{11}$$

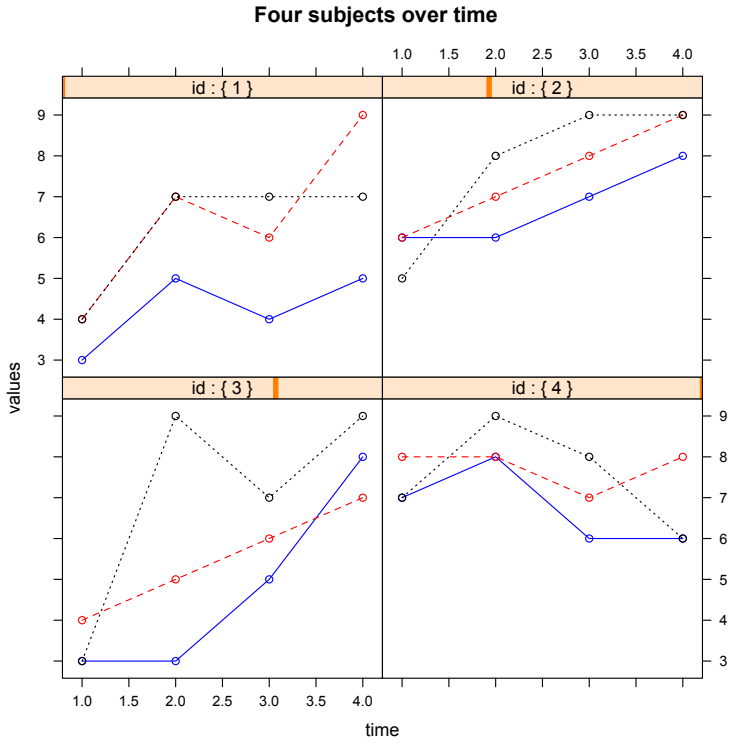

Figure 2. Four subjects are measured over four time points on three variables. The data are shown in Table 3 and adapted from Shrout & Lane (2012). Figure drawn using the `mlPlot` function in *psych*.

But, if the interest is individual differences from one randomly chosen time point (R), then we need to add time and its interaction with person in the denominator and we find

$$R_{1R} = \frac{\sigma_p^2 + \frac{\sigma_{p*i}^2}{m}}{\sigma_p^2 + \frac{\sigma_{p*i}^2}{m} + \sigma_t^2 + \sigma_{p*t}^2 + \frac{\sigma_e^2}{m}}. \tag{12}$$

An extension of the reliability coefficient from one randomly chosen time point (R_{1R}) to the average of k times (R_{kR}) is analogous to the benefit of Spearman-Brown formula and is

$$R_{kR} = \frac{\sigma_p^2 + \frac{\sigma_{p*i}^2}{m}}{\sigma_p^2 + \frac{\sigma_{p*i}^2}{m} + \frac{\sigma_t^2 + \sigma_{p*t}^2}{k} + \frac{\sigma_e^2}{km}}. \tag{13}$$

To measure the reliability of within individual change, we do not need to consider between person variability, just variability within people over time:

$$R_C = \frac{\sigma_{p*t}^2}{\sigma_{p*t}^2 + \frac{\sigma_e^2}{m}}. \tag{14}$$

The four equations above assume that time points are fixed and that all subjects are measured at the same time points (e.g., perhaps every evening, or at fixed times through out the day). But if

the timing differs across subjects we need to think of time as nested within subjects and we derive two more reliabilities, that between subjects and that within subjects:

$$R_{kRN} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{t(p)}^2}{k} + \frac{\sigma_e^2}{km}}. \tag{15}$$

$$R_{CN} = \frac{\sigma_{t(p)}^2}{\sigma_{t(p)}^2 + \frac{\sigma_e^2}{m}}. \tag{16}$$

The previous six equations might seem daunting, but are included to show the logic of generalizability theory as applied to the problems associated with measuring individual differences in mood over time. All six are included as the output for the `multilevel.reliabiity` function, see Table 4.

When doing multilevel reliability, it is straightforward to find the reliability of each individual subject over items and over time. People are not the same and the overall indices do not reflect how some subjects show a very different pattern of response. The `multilevel.reliability` function returns reliability estimates for each subject over time, as well as the six estimates shown in Table 4. In this supplement, we show multi-level reliabilities for the four subjects in Table 3 (Table 4) as well as 77 subjects on our ten anxiety items across four time points (see estimating multilevel reliability in the R code section).

Table 4

Alternative estimates of reliability based upon Generalizability theory for the example data set. Analysis done by the `multilevel.reliability` *function.*

RkF	= 0.92	Reliability of average of all ratings across all items and times (Fixed time effects)
R1R	= 0.25	Generalizability of a single time point across all items (Random time effects)
RkR	= 0.57	Generalizability of average time points across all items (Random time effects)
Rc	= 0.79	Generalizability of change (fixed time points, fixed items)
RkRn	= 0.44	Generalizability of between person differences averaged over time (time nested within people)
Rcn	= 0.73	Generalizability of within person variations averaged over items (time nested within people)

The data had 4 observations taken over 4 time intervals for 3 items.

Source		Variance	Percentage
Person	σ_p^2	0.57	0.15
Time	σ_t^2	0.82	0.21
Items	σ_i^2	0.48	0.12
Person x Time	σ_{pt}^2	0.84	0.22
Person x Items	σ_{pi}^2	0.12	0.03
Time x items	σ_{ti}^2	0.31	0.08
Residual	σ_e^2	0.68	0.18
Total	σ_T^2	3.82	1.00
	Nested model		
Person	σ_p^2	0.38	0.11
Time(person)	$\sigma_{p(t)}^2$	1.46	0.43
Residual	σ_e^2	1.58	0.46
Total	σ_T^2	3.43	1.00

Reliability of raters

Consider the case where we are rating numerous subjects with only a few judges. We might do a small study first to determine how much our judges agree with each other, and depending upon this result, decide upon how many judges to use going forward. As an example, examine the data from 5 judges (raters) who are rating the anxiety of 10 subjects (Table 5). If raters are expensive, we might want to use the ratings of just one judge rather than all five. In this case, we

will want to know how ratings of any single judge will agree with those from the other judges. In this case, differences in leniency (the judges' means) between judges will make a difference in their judgements. In addition, different judges might use the scale differently, with some having more variance than others. We also need to think about how we will use the judges. Will we use their ratings as given, will we use their ratings as deviations from their mean, or will we pool the judges? All of these choices lead to different estimates of generalizability. Shrout & Fleiss (1979) provide a very clear exposition of three different cases and the resulting equations for reliability. Although they express their treatment in terms of Mean Squares derived from an analysis of variance (e.g., the `aov` function in R), it is equally easy to do this with variance components estimated using a mixed effects linear model (e.g., `lmer` from the *lme4* package (Bates et al., 2015) in R). Both of these procedures are implemented in the `ICC` function in the *psych* package.

In Case 1 each subject is rated by k randomly chosen judges. The variance of the ratings is thus a mixture of between person and between judges. We can estimate these variance components from a one way analysis of variance treating subjects as random effects. Within person variance is an unknown mixture of rater and and residual (which includes error and the interaction) effects. Reliability for a single judge (ICC_1) is the ratio of person variance to total variance, while reliability for multiple judges (ICC_{1k}) adjusts the residual variance (σ_w^2) by the number of judges:

$$ICC_1 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_w^2} \qquad ICC_{1k} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_w^2}{k}}. \tag{17}$$

Case 2 is more typical, in that we are still using the ratings of k randomly chosen judges, but each judge rates all subjects. We are trying to generalize to another set of randomly chosen judges. This is a two way random effects model where both subjects and raters are chosen at random. By partitioning out the raters effects (σ_r^2) from the residual, we improve our estimate for the person variance (σ_p^2). Once again, by having multiple raters, the residual term ($\sigma_r^2 + \sigma_e^2$) is reduced by the number of raters ($\frac{\sigma_r^2+\sigma_e^2}{k}$):

$$ICC_2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_r^2 + \sigma_e^2} \qquad ICC_{2k} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_r^2+\sigma_e^2}{k}}. \tag{18}$$

Case 3 is unusual when considering judges, but is typical when considering items. It assumes judges are fixed rather than random effects. Thus, this is a two-way mixed model (subjects are random, judges are fixed). The estimate of the person variance is the same as in Case 2, but by assuming judges are fixed, the variance associated with judges is removed from the divisor of our reliability coefficient:

$$ICC_3 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_e^2} \qquad ICC_{3k} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_e^2}{k}}. \tag{19}$$

Each of these cases can be examined for the reliability of a single judge, or for k judges. The effect of pooling judges is identical to the effect of pooling items and is just the Spearman-Brown correction. When applied to items, the ICC_{3k} is the same as α because we typically associate item differences as fixed effects. The `ICC` function reports 6 different relliability estimates: three for the case of single judges, three for the case of multiple judges. It also reports the results in terms of a traditional analysis of variance aa well as a mixed effects linear model as well as confidence intervals for each coefficient (Table 6).

The intraclass correlation is appropriate when ratings are numerical, but sometimes ratings are categorical (particularly in clinical diagnosis or in evaluating themes in stories). This then

Table 5

An example of rater reliability. Five judges (raters) rate 10 subjects on a trait. The subjects differ in their overall mean score across judges, the judges differ in their mean ratings. These data produce six different estimates of reliability (Table 6).

	Judge or Rater					
Subject	J1	J2	J3	J4	J5	Mean
S1	1	1	3	2	3	2.0
S2	1	1	3	2	5	2.4
S3	1	2	3	2	4	2.4
S4	4	1	2	6	2	3.0
S5	2	4	3	5	3	3.4
S6	3	4	3	4	6	4.0
S7	1	4	6	4	6	4.2
S8	3	4	4	4	6	4.2
S9	3	5	4	6	5	4.6
S10	5	5	5	6	5	5.2
Mean	2	2.4	2.4	3	3.4	4.0

Table 6

Intra class correlations summarize the amount of variance due to subjects, raters, and their interactions. Depending upon the type of generalization to be made, one of six different reliability coefficients is most appropriate. Scores may be analyzed as a one way (Case 1) or two way (Cases 2 and 3) ANOVAs with random (Cases 1 and 2) or mixed effects (Case 3). Variance components may be derived from the MS from ANOVA or directly from the `ICC` *output: For Case 1,* $\sigma_p^2 = \frac{MS_p - MS_w}{k}$. *Similarly, for Cases 2 and 3 ,* $\sigma_p^2 = \frac{MS_p - MS_{residual}}{k}$. *ICCs may be based upon 1 rater or k raters. Data from Table 5 are analyzed using the* `ICC` *function from the psych package. See later for data and code.*

Analysis of Variance and the resulting decomposition into variance components

	df	SS	MS	Variance	Value by case			% of total		
					C1	C2	C3	C1	C2	C3
Person	9	51.22	5.69	σ_p^2	0.76	0.87	0.87	29	32	40
Within	40	75.20	1.88	σ_w^2	1.88			71		
Rater	4	27.32	6.83	σ_r^2		0.55			20	
Residual (P x R)	36	47.88	1.33	σ_e^2		1.33	1.33		48	60
Total				σ_t^2	2.64	2.75	2.20	100	100	100

Intraclass correlations and their confidence intervals (From the `ICC` function).

Variable	type	ICC	F	df1	df2	p	lower bound	upper bound
Single_raters_absolute	ICC1	0.29	3.03	9	40	0.01	0.04	0.66
Single_random_raters	ICC2	0.32	4.28	9	36	0.00	0.09	0.67
Single_fixed_raters	ICC3	0.40	4.28	9	36	0.00	0.13	0.74
Average_raters_absolute	ICC1k	0.67	3.03	9	40	0.01	0.19	0.91
Average_random_raters	ICC2k	0.70	4.28	9	36	0.00	0.32	0.91
Average_fixed_raters	ICC3k	0.77	4.28	9	36	0.00	0.42	0.93

Number of subjects = 10 Number of raters = 5

leads to measures of agreement of nominal ratings. Rediscovered multiple times and given different names (Conger, 1980; Scott, 1955; Hubert, 1977; Zapf et al., 2016) perhaps the most standard coefficient is known as Cohen's Kappa (Cohen, 1960, 1968) which adjusts observed proportions of agreement by the expected proportion:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{f_o - f_e}{N - f_e} \tag{20}$$

where $p_o = \frac{f_o}{N}$ is the observed proportion (p_o) or frequency of agreement (f_o) between two observers, and $p_e = \frac{f_e}{N}$ is the expected proportion or frequency of agreement (f_e) (Cohen, 1960). Because raw agreements will reflect the base rates of judgements, κ corrects for the expected agreement on the assumption of independence of the raters. Thus, if two raters each use one category 60% of the time, we would expect them to agree by chance 36% of the time in their positive judgements and 16% in their negative judgements. Various estimates of correlations of nominal data have been proposed and differ primarily in the treatment of the correction for chance agreement (Feng, 2015). Thus, κ adjusts for differences in the marginal likelihood of judges, while Krippendorf's α_k does not (Krippendorff, 1970, 2004). To Krippendorff (2004) this is a strength of α_k, but to Fleiss it is not (Krippendorff & Fleiss, 1978).

If some disagreements are more important than others, we have weighted κ which with appropriate weights is equal to the intraclass correlation between the raters (Cohen, 1968; Fleiss & Cohen, 1973). For multiple raters, the average κ is known as Light's κ (Conger, 1980; Light, 1971).

Consider a study where four coders are asked to rate 10 narratives for three mutually exclusive categories: Achievement, Intimacy, and Power. A hypothetical example of such data is shown in Table 7. Raters 1 and 2 and 3 and 4 show high agreement, but there is no agreement between raters 2 and 4.

Real life examples of a range of κ values are given by Freedman et al. (2013) in a discussion of the revised DSM where the κ values for clinical diagnoses range from "very good agreement" ($>$ .60) for major neurocognitive disorders or post-traumatic stress disorder, to "good" (.40-.60) for bipolar II, or schizophrena, to "questionable agreement" (.2-4) for generalized anxiety or obsessive compulsive disorder, to values which did not exceed the confidence values of 0. When comparing the presence or absence of each of five narrative themes in a life story interview, Guo et al. (2016) report how two independent raters of each of 12 different interview segments showed high reliability of judgements with κ values ranging from .61 (did the story report early advantage) to .83 (did the story discuss prosocial goals).

Reliability of a difference score

It is frequently tempting to compare individual patterns such as being more spatially versus verbally skilled. Unfortunately, the reliability of such a difference is much lower than either that of either score.

$$\rho_{x-y} = \frac{\rho_{xx} + \rho_{yy} - 2r_{xy}}{2 * (1 - r_{xy})} \tag{21}$$

Consider the 16 item ability test included as the `ability` data set in the *psychTools*. This test reflects 16 items for 1525 subjects collected from the SAPA project and had four subscales, verbal, quantitative, spatial and matrix reasoning. Some researchers would like to form two composites from these scales (verbal and spatial) and then compare the difference between these two composites. Unfortunately, although the composites are reasonably reliable (verbal α = .77, spatial α= .72, verbal ω_t = .78 and spatial ω_t = .76), the reliability of the difference scores is much less .39 based

Table 7

Cohen's κ can be used to assess the change corrected agreement between raters for categorical data. κ adjusts observed agreement by expected agreement. It is found using the `cohen.kappa` *function.*

Hypothethetical ratings from four raters for 10 subjects on three strivings.

Subject	R1	R2	R3	R4
1	Achieve	Achieve	Achieve	Power
2	Achieve	Achieve	Intimacy	Power
3	Achieve	Achieve	Intimacy	Power
4	Achieve	Achieve	Power	Power
5	Achieve	Intimacy	Achieve	Achieve
6	Intimacy	Achieve	Achieve	Achieve
7	Intimacy	Intimacy	Intimacy	Intimacy
8	Intimacy	Power	Intimacy	Intimacy
9	Power	Power	Intimacy	Intimacy
10	Power	Power	Power	Power

Produces this measure of % agreement

Rater	R1	R2	R3	R4
R1	100	70	50	40
R2	70	100	40	30
R3	50	40	100	70
R4	40	30	70	100

Which, when adjusted for chance becomes
Kappa by each pair of raters
(Unweighted below the diagonal, weighted above)

Rater	R1	R2	R3	R4
R1	1.00	0.78	0.30	-0.14
R2	0.52	1.00	0.29	-0.17
R3	0.24	0.13	1.00	0.52
R4	0.15	-0.01	0.57	1.00

Average Cohen kappa for all raters	0.27
Average weighted kappa for all raters	0.26

upon α reliability, and .46 based upon ω_t reliability. Show the code for these calculations in the section on R code. Although both the verbal (.77) and spatial (.72) have reasonable α values for 8 item scales, the α of the composite is much lower (.19). The reliability of the difference score using equation 21 is slightly higher ($\frac{(.77+.72-2*.58)}{2*(1-.58)} = .39$), suggesting that the α values are underestimates. Using the ω_t values for the two composites, and then finding the ω_t value for the difference scores, shows more agreement with equation 21 ($\frac{.79+.76-2*.58)}{(2*(1-.58)} = .46$ versus the .45 from `omega`.

Item response theory

Reliability is a joint property of the test and the people being measured by the test. For fixed amount of error, reliability is a function of the variance of the people being assessed. A test of ability will be reliable if given to a random sample of 18-20 year olds, but much less reliable if given to students at a particularly selective college because there will be less between person variance. The reliability of a test of emotional stability will be higher if given to a mixture of psychiatric patients and their spouses than it will be if given just to the patients. That is, reliability is not a property of test independent of the people taking it. This is the basic concept of Item Response Theory (IRT), called by some the "new psychometrics" (Embretson, 1996, 1999; Embretson & Reise, 2000) and which models the individual's patterns of response as a function of parameters (discrimination, difficulty) of the item. Classical test theory has been likened to a "flogging wall" where we count the number of whips hitting subjects as they move down a conveyer belt as a measure of height rather than calibrating items to the targets (Lumsden, 1976).

By focusing on item difficulty (endorsement frequency) it is possible to consider the range of application of our scores. Items are most informative if they are equally likely to be passed or failed (endorsed or not endorsed). But this can only be the case for a particular person taking the

test and can not be the case for a person with a higher or lower latent score. Although tests are maximally reliable if all of the items are equally difficult, such a test will not be very discriminating at any other than at that level (Loevinger, 1954). Thus, we need to focus on spreading out the items across the range to be measured.

The essential assumptions of IRT is that items can differ in how hard they are, as well as how well they measure the latent trait. Although seemingly quite different from classical approaches, there is a one-to-one mapping between the difficulty and discrimination parameters of IRT and the factor loadings and item response thresholds found by factor analysis of the polychoric correlations of the items (Kamata & Bauer, 2008; McDonald, 1999).

Two of the IRT parameters are item difficulty (location = δ_j) and item discrimination (α_j). These may be found from factor analysis of the tetrachoric or polychoric correlations with the τ value representing the frequency of response in terms of the normal distribution and λ representing the factor loading of the item. D may be 1 or 1.702 and is conversion factor from the normal curve to a logistic function.

$$\delta_j = \frac{D\tau}{\sqrt{1-\lambda_j^2}}, \qquad \alpha_j = \frac{\lambda_j}{\sqrt{1-\lambda_j^2}} \tag{22}$$

FA parameters from IRT

$$\lambda_j = \frac{\alpha_j}{\sqrt{1+\alpha_j^2}}, \qquad \tau_j = \frac{\delta_j}{\sqrt{1+\alpha_j^2}}. \tag{23}$$

Item information is the reciprocal of the standard error of the item, test information is just the sum of the item information, and test reliability at any lolocation is 1- 1/(test information) at that location. The relationship of the IRT approach to classical reliability theory is given a very clear explication by Markon (2013) who examines how test information (and thus the reliability) varies by subject variance as well as trait level. A test can be developed to be reliable for certain discriminations (e.g. between psychiatric patients) and less reliable for discriminating between members of a control group. The particular strength of IRT approaches is the use in tailored or adaptive testing where the focus is on the reliability for a particular person at a particular level of the latent trait. We can see this most clearly when we plot the item information for the 10 `sai` items used in our examples. The test is most informative (reliable) for mid to high levels of anxiety (Figure 3).

R Code examples

Here we include R code (R Core Team, 2019) to find the various reliability estimates discussed above. All of these examples require installing the *psych* (Revelle, 2019a) and *psychTools* (Revelle, 2019b) packages and then making them active. To install R on your computer, go to `https://cran.r-project.org` and install the most recent version that is appropriate for your computer (PC, MacOS, Linux). For details on installing and using R, go to the `https://personality-project.org/r`. Many people find that RStudio (RStudio Team, 2016) is a very convenient interface to R. It may be downloaded from `https://www.rstudio.com`.

In the following detail, we leave out the ">" prompt with which R prefaces every line where it is waiting for input. We also use the # symbol to add comments to the R code. Copying these lines as written into R will allow you to run the code.

R is a set of functions that act upon input. Thus each command is asking to execute the function on a particular input and then return the result to some object. If the object is not named, then the result is printed to the screen. Because each R command is a function, it must be followed

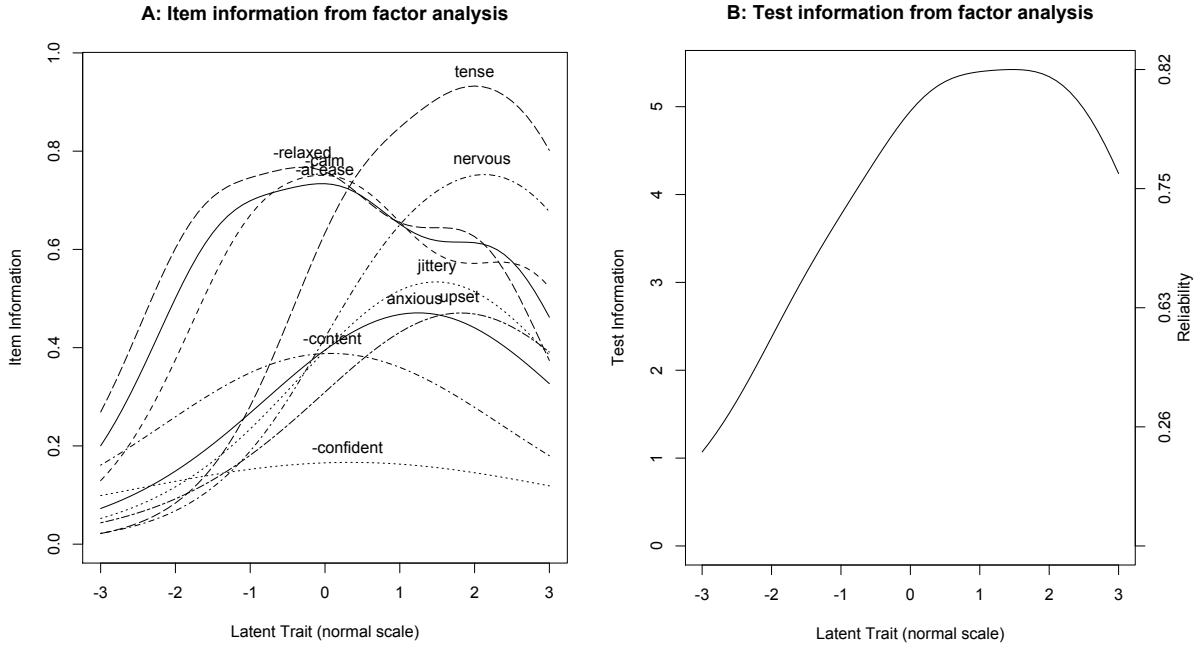

Figure 3. Panel A: Item information for 10 `sai` items varies as a function of the latent trait of anxiety. Panel B: The total test information is just the sum of the item information. Test reliability varies as 1-1/(test information) and thus reliability varies across the range of individual differences on the latent trait. The 10 anxiety items are most reliable for moderate to high levels of anxiety and much less reliable at low levels of anxiety.

by a parenthetical expression containing the inputs. Always use these parentheses. If you make a mistake, R is relatively forgiving and you can repeat the command by just using the up arrow on your computer keyboard. R is case sensitive, so make sure that you follow the capitalization of the functions. Some functions are "camelCase" which means that parts of the function are emphasized by starting with a capital letter.

First steps: installing *psych* and making it active

To use any R package, it is necessary to first install it from the Comprehensive R Archive Network (CRAN) which may be found at https://cran.r-project.org. This needs to be done once, and then from then on the package merely needs to be made active using the `library` command.

R code

```
install.packages("psych",dependencies = TRUE)  #Just need to do this once
install.packages("psychTools")  #Just do this once as well
library(psych) #make the psych package active-- need to do this everytime you start R
library(psychTools) #if you want to use the example data sets and some convenient tools
```

Detailed instructions for using *psych* may be found by reading the accompanying *vignettes*[4]

[4]On a Mac, vignettes may be found in directly in the help menu. For RStudio, using the command `vignette(topic="intro",package="psych")` and `vignette(topic="overview",package="psychTools")` will open a pdf window.

or reading the series of "HowTo"s from the https://personality-project.org/r. As is true of all R packages, help for individual functions may be obtained by entering ? followed by the command you do not understand.

R code

```
#Getting help
help("psych") #opens a help window overview of the package
help{"psychTools") #opens a help window listing the various data sets in psychTools
vignette(topic="intro",package="psych") #opens an extensive  pdf document
vignette(topic="overview",package=psychTools") #opens the second part of this vignette
?omega #opens the specific help page for e.g., the omega function
```

Perhaps the most powerful feature of R is that all functions in R operate upon objects and then return the result as another object. This is the real power of R for it allows us to do a particular analysis, look at the results (if we want) and then do a subsequent analysis on these results. Most functions in R have default values for certain options. These can be changed by specifying the option by name and giving the desired value. To find the complete list of options for any functions, you can ask for `help` for that function. RStudio will prompt with the available options when you type the name of the function.

Entering your data

For the examples below, we will use datasets already available in R and the *psychTools* package. However, it is important to know how to enter your own data into R as well. The easiest way of doing this is to read from an external file where the first row of the file gives the names of the variables and the subsequent rows are one row for each subject. If you have a data file that is a text file with the suffix .text, .txt, or .csv, or that has been saved from e.g., SPSS as a .sav file, then you can read the data using the `read.file` command. This will open a search window on your computer and you can locate the file. Alternatively, you can copy the data to your clipboard and use the `read.clipboard` command.

By default, `read.file` and `read.clipboard` assume that the first line of the file includes "header" information. That is, the names of the variables. If this is not true, then specify that header=FALSE. Examples of the code one might use to enter your data are given below. All of the 'R code' chunks in this file can be copied and pasted directly into the R console.

R code

```
my.data <- read.file()#opens an OS dependent search window and reads data according to the suffix
#or first copy your data to the clipboard and then
my.data <- read.clipboard() #assumes that header information is on the first line
my.data <- read.clipboard(header=FALSE) #no header information: the data start on the first line
```

These read operations will result in your data being stored as a `data.frame` which is just a two dimensional table of data. Typically subjects will be rows and variables will be the columns. In order to see the size of the resulting data frame, use `dim` to show the number of rows and columns. To show just a few lines from the top and bottom of a matrix or data.frame, use the `headTail` function. To see all the items in an object, you may use the `view` function, but if the object is very big this is not recommended.

R code

```
dim(sai)   #how many rows (subjects) and columns (variables) in the built in data set sai
dim(msqR) #how many rows (subjects) and columns  variables) in the msqR data set
```

```
headTail(sai) # show the first and last 3 lines of the sai data set
```

```
dim(sai)   #how many rows (subjects) and columns (variables) in the built in data set sai
[1] 5378   23
> dim(msqR) #how many rows (subjects) and columns  variables) in the msqR data set
[1] 6411   88
> headTail(sai) # show the first and last 3 lines of the sai data set
      study time  id calm secure tense regretful ... rattled joyful pleasant
1     AGES    1   1    3      3     2         1  ...      1      3        3
2     AGES    1   2    3      3     2         2   ...     1      1        2
3     AGES    1   3    3      3     2         1  ...      1      3        3
4     AGES    1   4    3      3     1         1  ...      1      2        3
...   <NA>  ... ...  ...    ...   ...       ...   ... ...    ...      ...
5375  XRAY    2 197    2      2     3      <NA> ...    <NA>   <NA>     <NA>
5376  XRAY    2 198    4      4     1         1 ...      1      3        4
5377  XRAY    2 199    4      4     1         1 ...      1      3        4
5378  XRAY    2 200    3      3     2         2 ...      1      3        3
```

Specifying the items we want

For these examples we use small subsets of the larger `msqR` and `sai` data sets (in *psychTools* and then specify which items to score for which analysis. The `msqR` data set is stored as a `data.frame` which may be thought of a spreadsheet with subjects as rows and variables as columns. (Using the $ command specfies a particular column by name). Both of these data sets represent data collected in multiple different studies with different designs. Thus, to show the different studies and the number of subjects per occasion we use the `table` command. `table(msqR$study,msqR$time)` does a cross tabulation of two variables within the `msqR` data.frame, the `study` and the `time` variables.

Because the entire data set includes 6,411 row for 3,032 unique subjects (some studies included multiple administrations), we will select just subjects from studies that meet particular criteria. That is, for short term test-dependability, those studies where the SAI and MSQ was given twice in the same session (time = 1 and 2). For longer term stability (over 1-2 days), those studies where the SAI and MSQ were given on different days (time = 1 and 3). We use the `subset` function to choose just those subjects who meet certain conditions (e.g., the first occasion data). We use "==" to represent equality.

R code

```
?msqR   #ask for information about the sai data set
table(msqR$study,msqR$time)   #show the study names and  sample sizes
#Now, select some subsets for analysis using the subset function.
#the short term consistency sets
sai.control <- subset(sai,is.element(sai$study,c("Cart", "Fast", "SHED",  "SHOP")) )
#pre and post drug studies
sai.drug <- subset(sai,is.element(sai$study, c("AGES","SALT","VALE","XRAY")))
#pre and post film studies
sai.film <- subset(sai,is.element(sai$study, c("FIAT","FLAT", "XRAY") ))
msq.control <- subset(msqR,is.element(msqR$study,c("Cart", "Fast", "SHED",  "SHOP")) )
#pre and post drug studies
msq.drug <- subset(msqR,is.element(msqR$study, c("AGES","CITY","EMIT","SALT","VALE","XRAY")))
#pre and post film studies
msq.film <- subset(msqR,is.element(msqR$study, c("FIAT","FLAT", "MAPS", "MIXX","XRAY") ))
msq.films4 <- subset(msqR, is.element(msqR$study, c("FLAT", "MAPS", "XRAY") ))
msq1 <- subset(msqR,msqR$time == 1)  #just the first day measures
sai1 <- subset(sai,sai$time==1)  #just the first set of observations for the SAI
sam.rim <- subset(sai,(sai$study %in% c("SAM" ,"RIM")))#choose SAM and RIM for 2 day test retest
vale <- subset(sai,sai$study=="VALE") #choose the VALE study for multilevel analysis
```

Two basic R commands (`dim` and (`table`) allow us to the see the size (dimensions) of these data sets, and then count the cases meeting certain conditions.

R code

```
dim(msq.control)  #how many subjects and how many items?
dim(sai.control)  #show the number of subjects and items for the second subset
table(sam.rim$time) #how many were in each time point
table(vale$time) #how many were repeated twice on day 1 and then on day 2
```

Produces this output:

```
> dim(msq1)  #how many subjects and how many items?
[1] 3032   88
> dim(sai1)  #show the number of subjects and items for the second subset
[1] 3032   23
> table(rim$time) #how many were in each time point
  1   3
666 666
> table(vale$time) #how many were repeated twice on day 1 and then on day 2
 1  2  3  4
77 77 70 70
```

We want to do analyses on those items in the `sai` and `msqR` data sets that overlap. The items, although all thought to measure anxiety reflect two subdimensions, positive and negative affect/tension. We can score them for positive affect, negative affect, and total anxiety on one subset of `sai` items. Of the 20 items, 10 overlap with the `msqR` items, and we can use the other 10 as a logical alternate form. We indicate reversed keyed items by a negative sign.

We specify the scoring keys for all seven of these scales. Note that these keys include overlapping items. This will artificially inflate correlations of these scales. We form a `list` of the seven different keys, where each key is given a name and is formed by concatenating (using the `c` command) the separate elements of the key.

R code

```
#create keying information for several analyses
sai.alternate.forms <- list(
sai=c( "anxious", "jittery", "nervous" ,"tense", "upset",   "-at.ease" ,  "-calm" ,
"-confident", "-content","-relaxed", "regretful", "worrying", "high.strung",
"worried", "rattled", "-secure",   "-rested", "-comfortable", "-joyful",  "-pleasant"),
anx1 = c("anxious", "jittery", "nervous" ,"tense", "upset","-at.ease" ,  "-calm" ,
       "-confident", "-content","-relaxed"),
 pos1 =c( "at.ease","calm","confident","content","relaxed"),
 neg1 = c("anxious", "jittery", "nervous" ,"tense"  ,   "upset"),
 anx2 = c("regretful","worrying", "high.strung","worried", "rattled", "-secure",
    "-rested", "-comfortable", "-joyful",  "-pleasant" ),
  pos2=c(  "secure","rested","comfortable" ,"joyful" , "pleasant" ),
  neg2=c("regretful","worrying", "high.strung","worried", "rattled" )
)

 anx.keys <- sai.alternate.forms$anx1   #the  overlapping keys for scoring sai and msq

 select <- selectFromKeys (anx.keys)  #to be used later in alpha and multilevel.reliability
```

Three measures of test-retest reliability using `testRetest` function

In some studies, the STAI and MSQ were given before and after a control, drug, or film manipulation. This allows us to test for the short term dependability of these measures (see

Table 1). Here we show the commands do this but just the full output for one set. To run the `testRetest` function, the data need to be in one of two forms: one object where the subjects are identified with an identification number and the time (1 or 2) of testing is specified or two data objects with an equal number of rows. Here we show the first way of finding test-retest measures.

As is true of most R functions, the `testRetest` returns many different objects (results). The `print` function is called automatically by just specifying the name of the resulting object. `print` will give the output that the author of the function thinks is most useful in a somewhat neat manner. Some functions also have an associated `summary` function that gives even less output. To see all of the output, you need to inspect the various objects returned by the function.

R code

```
sai.test.retest.control     <-     testRetest(sai.control, keys=anx.keys)
sai.test.retest.drug        <-     testRetest(sai.drug, keys=anx.keys)
sai.test.retest.film        <-     testRetest(sai.film, keys=anx.keys)

msq.test.retest.control     <-     testRetest(msq.control, keys=anx.keys)
msq.test.retest.drug        <-     testRetest(msq.drug, keys=anx.keys)
msq.test.retest.film        <-     testRetest(msq.film, keys=anx.keys)

sai.test.retest.control        #show the complete ouput
summary( sai.test.retest.drug) #give just a  brief summary of the outpu
summary( sai.test.retest.film)  #many functions can be summarized
summary(msq.test.retest.control)
summary( msq.test.retest.drug)
summary(msq.test.retest.film )
```

Show the complete output for the control condition, and then just the summary statistics for the others.

Short Term – dependability. Short term dependability is just the test-retest correlation after a short delay. Here we show the test-retest correlation for the 313 subjects in the control condition

```
 sai.test.retest.control        #show the complete ouput

Test Retest reliability
Call: testRetest(t1 = sai.control, keys = anx.keys)

Number of subjects =  313  Number of items =  10
 Correlation of scale scores over time 0.76
 Alpha reliability statistics for time 1 and time 2
       raw G3 std G3   G6 av.r  S/N   se lower upper var.r
Time 1   0.86   0.86 0.88 0.38 6.24 0.04  0.76  0.93  0.03
Time 2   0.87   0.87 0.89 0.40 6.72 0.04  0.79  0.94  0.02

 Mean between person, across item reliability =  0.6
 Mean within person, across item reliability =  0.67
with standard deviation of  0.27

 Mean within person, across item d2 =  0.54
R1F  =  0.91 Reliability of average of all items for one  time (Random time effects)
RkF  =  0.95 Reliability of average of all items and both times (Fixed time effects)
R1R  =  0.48 Generalizability of a single time point across all items (Random time effects)
Rc   =  0.91 Generalizability of change (fixed time points, fixed items)
Multilevel components of variance
             variance Percent
ID               0.21    0.18
Time             0.01    0.01
Items            0.28    0.24
ID x time        0.19    0.17
ID x items       0.05    0.04
```

```
time x items      0.21    0.18
Residual          0.20    0.18
Total             1.14    1.00

 To see the item.stats, print with short=FALSE.
To see the subject reliabilities and differences, examine the 'scores' object.
```

Several studies had a drug or film manipulation between the two adminstrations of the `sai`. Summarize the results using the `summary` function.

```
> summary( sai.test.retest.drug) #a brief summary
Call: testRetest(t1 = sai.drug, keys = anx.keys)

Test-retest correlations and reliabilities

Test retest correlation =  0.73
 Alpha reliabilities for both time points
  raw G3 std G3   G6 av.r  S/N   se lower upper var.r
x   0.88   0.88 0.90 0.43 7.49 0.03  0.81  0.93  0.03
y   0.89   0.89 0.91 0.45 8.05 0.03  0.81  0.91  0.03


> summary( sai.test.retest.film)
Call: testRetest(t1 = sai.film, keys = anx.keys)

Test-retest correlations and reliabilities

Test retest correlation =  0.57
 Alpha reliabilities for both time points
  raw G3 std G3   G6 av.r  S/N   se lower upper var.r
x   0.88   0.88 0.91 0.43 7.66 0.03  0.81  0.93  0.02
y   0.88   0.88 0.91 0.42 7.25 0.03  0.78  0.91  0.03


> summary(msq.test.retest.control)
Call: testRetest(t1 = msq.control, keys = anx.keys)

Test-retest correlations and reliabilities

Test retest correlation =  0.74
 Alpha reliabilities for both time points
  raw G3 std G3   G6 av.r  S/N   se lower upper var.r
x    0.8    0.8 0.84 0.29 4.10 0.09  0.62  0.95  0.03
y    0.8    0.8 0.84 0.28 3.92 0.09  0.59  0.93  0.04


> summary( msq.test.retest.drug)
Call: testRetest(t1 = msq.drug, keys = anx.keys)

Test-retest correlations and reliabilities

Test retest correlation =  0.74
 Alpha reliabilities for both time points
  raw G3 std G3   G6 av.r  S/N   se lower upper var.r
x   0.83   0.83 0.87 0.33 5.03 0.06  0.69  0.94  0.03
y   0.84   0.84 0.88 0.35 5.44 0.05  0.69  0.92  0.04


> summary(msq.test.retest.film )
Call: testRetest(t1 = msq.film, keys = anx.keys)

Test-retest correlations and reliabilities

Test retest correlation =  0.55
 Alpha reliabilities for both time points
  raw G3 std G3   G6 av.r  S/N   se lower upper var.r
```

```
x   0.85   0.85 0.88 0.37 5.87 0.05  0.75  0.94  0.03
y   0.83   0.83 0.87 0.33 4.96 0.06  0.66  0.92  0.04
```

Test dependability by examining duplicated items. Ten of the `sai` items were given immediately afterwards as part of the `msqR`. This gives us a measure of immediate dependability (see Table 2). An alternative way to do `testRetest` to include two data sets with the same number of rows in each set. The code for this is taken from the example for `testRetest` in the *psych*. We find the overlapping items by using the `is.element` function which is identical to the %in% function. We use the example from the help pages for the `testRetest` function.

R code

```
 sai.xray1 <- subset(sai,((sai$time==1) & sai$study=="XRAY"))
msq.xray <- subset(psychTools::msqR,
 (psychTools::msqR$study=="XRAY") & (psychTools::msqR$time==1))
select <- colnames(sai.xray1)[is.element(colnames(sai.xray1 ),colnames(psychTools::msqR))]

select <-select[-c(1:3)]  #get rid of the id information
#The case where the two times are in the form x, y
#show the items we are including
select
dependability <-  testRetest(sai.xray1,msq.xray,keys=select)
names(dependability)  #what are the objects included that could be examined
dependability  #print the main results by specifying the name of the object to be printed.
```

```
   select
 [1] "calm"      "tense"     "at.ease"   "upset"     "anxious"   "confident" "nervous"
   "jittery"   "relaxed"   "content"
> dependability <-  testRetest(sai.xray1,msq.xray,keys=select)
> names(dependability)  #what are the objects included that could be examined
 [1] "r12"        "alpha"      "rqq"        "dxy"        "item.stats" "scores"     "xy.df"
   "key"       "ml"         "Call"

 dependability  #print the main results by specifying the name of the object to be printed.


  Test Retest reliability
Call: testRetest(t1 = sai.xray1, t2 = msq.xray, keys = select)

Number of subjects =  200  Number of items =  10
 Correlation of scale scores over time 0.92
 Alpha reliability statistics for time 1 and time 2
       raw G3 std G3   G6 av.r  S/N   se lower upper var.r
Time 1   0.90   0.90 0.91 0.46 8.54 0.02  0.83  0.92  0.03
Time 2   0.87   0.87 0.89 0.40 6.54 0.04  0.78  0.94  0.02

 Mean between person, across item reliability =  0.71
 Mean within person, across item reliability =  0.6
with standard deviation of  0.3

 Mean within person, across item d2 =  1.38
R1F  =  0.94 Reliability of average of all items for one  time (Random time effects)
RkF  =  0.97 Reliability of average of all items and both times (Fixed time effects)
R1R  =  0.33 Generalizability of a single time point across all items (Random time effects)
Rc   =  0.91 Generalizability of change (fixed time points, fixed items)
Multilevel components of variance
             variance Percent
ID               0.34    0.19
Time             0.44    0.25
Items            0.17    0.10
ID x time        0.24    0.13
ID x items       0.01    0.01
```

```
time x items     0.34    0.19
Residual         0.23    0.13
Total            1.78    1.00

 To see the item.stats, print with short=FALSE.
To see the subject reliabilities and differences, examine the 'scores' object.
```

Longer term – stability

There are several data sets that allow us to examine temporal stability over several days or weeks. The `msqR` and `sai` were given with a one to 2 delay in two studies, 666 subjects were available in two studies with within study repeated measures over two days.

R code

```
#select the two day subjects
sai.2 <- subset(sai, sai$study %in% cs(RIM,SAM))
msqR.2 <- subset(msqR, msqR$study %in% cs(RIM,SAM))
sai.stability <- testRetest(sai.2 ,keys = select)
msqR.stability <- testRetest(msqR.2, keys = select)
summary(sai.stability)
summary(msqR.stability)
```

Gives this output.

```
>  sai.2 <- subset(sai, sai$study %in% cs(RIM,SAM))
>  msqR.2 <- subset(msqR, msqR$study %in% cs(RIM,SAM))
>  sai.stability <- testRetest(sai.2 ,keys = select)
>  msqR.stability <- testRetest(msqR.2, keys = select)


summary(sai.stability)
Call: testRetest(t1 = sai.2, keys = select)
Test-retest correlations and reliabilities
Test retest correlation =  0.36
 Alpha reliabilities for both time points
  raw G3 std G3   G6 av.r  S/N   se lower upper var.r
x   0.86   0.86 0.89 0.38 6.07 0.04  0.73  0.92  0.03
y   0.87   0.87 0.90 0.41 6.85 0.04  0.78  0.93  0.03


>  summary(msqR.stability)
Call: testRetest(t1 = msqR.2, keys = select)
Test-retest correlations and reliabilities
Test retest correlation =  0.39
 Alpha reliabilities for both time points
  raw G3 std G3   G6 av.r  S/N   se lower upper var.r
x   0.81   0.82 0.86 0.31 4.44 0.08  0.63  0.93  0.04
y   0.82   0.83 0.86 0.32 4.72 0.07  0.65  0.93  0.04
```

Trait and State measures

Just as we can find the 1-2 day stability of the state scores of the `msqR` and `sai` data sets, so we can find the correlation between the trait measures on day 1 with the state measures at the same time as well as after a delay.

To do this, we first need to find the scores for the state measures and the trait measures and then correlate these measures. The `psych` package includes multiple ways of finding scale scores from items. Details on these may be found in the online "Vignette" included in the package. Here

we show one way of obtaining scores based upon the items using the `scoreItems` function. When we discuss the `alpha` function, we will show how to use that function to obtain scores as well.

There are two traditions in converting items into scales. One is to find the average item response, the other is to total the item responses. Although both options are available in `scoreItems`, the default is to find item averages. This has the advantage that the scores are in the same metric as the item. When scoring items, it is important to recognize that some items need to be reverse keyed. Although it is possible to do this by manually recoding items, it is easier to let the function do it for you by specifying the direction of the items in a *keying* list. We will use the keys we created before.

R code

```
sai.1 <- sai[sai$time == 1,]   #get just first measures for the sai
dim(sai.1)
dim(tai)  #note that these are the same
sai.1.scores <- scoreItems(keys=sai.alternate.forms , items = sai.1)
#now, show the correlations correcting for overlap.
sai.1.overlap <- scoreOverlap(keys = sai.alternate.forms, r = sai.1)
sai.1.overlap
```

```
#First the dimensional information

dim(sai.1)
[1] 3032   23
> dim(tai)  #note that these are the same
[1] 3032   23


#now the scoring results for the sai
Call: scoreOverlap(keys = sai.alternate.forms, r = sai.1)

(Standardized) Alpha:
 sai anx1 pos1 neg1 anx2 pos2 neg2
0.91 0.87 0.86 0.82 0.80 0.83 0.73

(Standardized) G6*:
 sai anx1 pos1 neg1 anx2 pos2 neg2
0.86 0.78 0.87 0.84 0.72 0.84 0.79

Average item correlation:
 sai anx1 pos1 neg1 anx2 pos2 neg2
0.34 0.40 0.54 0.48 0.28 0.50 0.36

Median item correlation:
 sai anx1 pos1 neg1 anx2 pos2 neg2
0.32 0.39 0.56 0.55 0.24 0.47 0.28

Number of items:
 sai anx1 pos1 neg1 anx2 pos2 neg2
  20   10    5    5   10    5    5

Signal to Noise ratio based upon average r and n
 sai anx1 pos1 neg1 anx2 pos2 neg2
10.2  6.6  5.9  4.7  4.0  4.9  2.8

Scale intercorrelations corrected for item overlap and attenuation
 adjusted for overlap correlations below the diagonal, alpha on the diagonal
 corrected correlations above the diagonal:
       sai  anx1  pos1  neg1  anx2  pos2  neg2
sai   0.91  1.00 -0.92  0.85  1.00 -0.82  0.85
anx1  0.89  0.87 -0.90  0.89  1.00 -0.77  0.87
pos1 -0.81 -0.77  0.86 -0.60 -0.95  0.95 -0.57
neg1  0.74  0.75 -0.50  0.82  0.81 -0.40  0.98
```

```
anx2  0.86  0.83 -0.78  0.66  0.80 -0.86  0.82
pos2 -0.71 -0.66  0.80 -0.33 -0.70  0.83 -0.41
neg2  0.70  0.70 -0.45  0.76  0.63 -0.32  0.73

 In order to see the item by scale loadings and frequency counts of the data
 print with the short option = FALSE
```

We then score the trait items (20 items are keyed) for the `tai`. Once again, we form positively and negatively keyed subscales, as well as an overall scale.

R code

```
tai.keys <- list(tai =c("-pleasant" ,  "nervous" , "not.satisfied" , "wish.happy" ,
   "failure"  ,  "-rested" ,  "-calm" , "difficulties", "worry",   "-happy",
   "disturbing.thoughts",    "lack.self.confidence", "-secure", "decisive"  ,
    "inadequate", "-content",  "thoughts.bother" , "disappointments" ,
    "-steady",  "tension"  ) ,
tai.pos = c( "pleasant" , "-wish.happy" ,"rested" ,"calm"  ,  "happy", "secure" ,
        "content"  , "steady"  ) ,
tai.neg =  c( "nervous" , "not.satisfied" , "failure"  , "difficulties","worry",
    "disturbing.thoughts" , "lack.self.confidence", "decisive","inadequate",
    "thoughts.bother","disappointments","tension"  )   )

tai.scores <- scoreItems(keys=tai.keys, items =tai)   #find the scores
tai.scores   #show the output
```

As mentioned before, R return a great deal of output, most of which is not shown to user unless requested. In the case of `scoreItems` one of the objects returned is a matrix of scores for each subject on each scale. We can then correlate these scores from the state scores at time one with the trait scores at time one. If we do this for the state scores taken several days after the trait measures, this gives us the amount of the state measure that is actually trait. Although there are many functions that can find the correlations between two data sets, we use `cor2` which automatically uses the "pairwise.complete" option in `cor` as well as rounds to 2 decimal places. `corr.test` would also give the correlations, as well as their confidence intervals, but given the sample size of 3,032, this not necessary.

R code

```
cor2(sai.1.scores$scores,tai.scores$scores)
#this is the same as
 R <- cor(sai.1.scores$scores,tai.scores$scores, use="pairwise")
 #but with useful defaults.
 #to find the confidence intervals of the  the correlations, we can use cor.test
ci <- corr.test(sai.1.scores$scores,tai.scores$scores)
```

We just show the results of (cor2)

```
 cor2(sai.1.scores$scores,tai.scores$scores)
       tai tai.pos tai.neg
sai   0.53   -0.52    0.46
anx1  0.47   -0.47    0.40
pos1 -0.50    0.55   -0.38
neg1  0.30   -0.23    0.31
anx2  0.55   -0.53    0.47
pos2 -0.48    0.54   -0.36
neg2  0.41   -0.31    0.42
```

Now, do this again, but find the correlation of trait on day 1 with the state measured on days 2 or 3. We choose the subjects from time 3 in the SAM and RIM experiments.

R code

```
sai.sam.rim.time3 <- sam.rim[sam.rim$time==3,]
tai.sam.rim <- tai[tai$study %in% cs(SAM, RIM),]
sai.scores <- scoreItems(keys=  sai.alternate.forms, items = sai.sam.rim.time3)
tai.scores.sam.rim <- scoreItems(keys = tai.keys, items = tai.sam.rim)
cor2(sai.scores$scores,tai.scores.sam.rim$scores)
```

Compare these delayed values to the immediate values.

```
 cor2(sai.scores$scores,tai.scores$scores)
       tai tai.pos tai.neg
sai   0.48   -0.47    0.42
anx1  0.43   -0.43    0.36
pos1 -0.45    0.48   -0.35
neg1  0.28   -0.24    0.28
anx2  0.50   -0.48    0.44
pos2 -0.44    0.46   -0.35
neg2  0.37   -0.30    0.38
```

Item measures of dependability and stability

Consistency using the testRetest function

To run the `testRetest` function, the data need to be in one of two forms: two data objects with an equal number of rows or one object where the subjects are identified with an identification number and the time (1 or 2) of testing is specified. Here we show the second way of doing this. We use the `sai` example data file included in *psychTools* package and then extract just a small subset (the RIM data set measured State Anxiety on two different days).

R code

```
sam.rim.test.retest <-  testRetest(sam.rim,keys=anx.keys)  #do the analysis
sam.rim.test.retest    #show the results
```

This results in the following output

```
Test Retest reliability
Call: testRetest(t1 = sam.rim, keys = anx.keys)

Number of subjects =  666  Number of items =  10
 Correlation of scale scores over time 0.36    <- This is the test-retest correlation
 Alpha reliability statistics for time 1 and time 2
       raw G3 std G3   G6 av.r  S/N   se lower upper var.r
Time 1   0.86   0.86 0.89 0.38 6.07 0.04  0.73  0.92  0.03
Time 2   0.87   0.87 0.90 0.41 6.85 0.04  0.78  0.93  0.03

 Mean between person, across item reliability =  0.29
 Mean within person, across item reliability =  0.48
with standard deviation of  0.37

 Mean within person, across item d2 =  0.99
R1F  =  0.79 Reliability of average of all items for one  time (Random time effects)
RkF  =  0.88 Reliability of average of all items and both times (Fixed time effects)
R1R  =  0.46 Generalizability of a single time point across all items (Random time effects)
Rc   =  0.77 Generalizability of change (fixed time points, fixed items)
Multilevel components of variance
            variance Percent
ID              0.10    0.10
```

```
Time              0.00    0.00
Items             0.21    0.21
ID x time         0.11    0.11
ID x items        0.17    0.17
time x items      0.10    0.10
Residual          0.32    0.32
Total             1.02    1.00
 To see the item.stats, print with short=FALSE.
To see the subject reliabilities and differences, examine the 'scores' object.
```

Split reliability using the splitHalf function

To find split half reliabilities and to graph the distributions of split halves (e.g., Figure 4) requires three lines. Here we use the built in `ability` data set of 16 items for 1,525 participants from *psychTools*. The data were taken from the Synthetic Aperture Personality Assessment (SAPA) project (https://sapa-project.org) (Revelle et al., 2016) and reported in (Condon & Revelle, 2014).

R code

```
sp <- splitHalf(ability,raw=TRUE, brute=TRUE)
sp #show the results
hist(sp$raw,breaks=101, xlab="Split half reliability",
      main="Split half reliabilities of 16 ICAR ability items")
```

```
Split half reliabilities
Call: splitHalf(r = ability, raw = TRUE, brute = TRUE)

Maximum split half reliability (lambda 4) =  0.87
Guttman lambda 6                          =  0.84
Average split half reliability            =  0.83
Guttman lambda 3 (alpha)                  =  0.83
Guttman lambda 2                          =  0.83
Minimum split half reliability  (beta)    =  0.73
Average interitem r =  0.23  with median =  0.21
                                             2.5% 50% 97.5%
 Quantiles of split half reliability      =  0.77 0.83 0.86
```

Internal consistency using the alpha and omega functions

Although we do not recommend α as a measure of consistency, many researchers want to report it. The `alpha` function will do that. Confidence intervals from normal theory (Duhachek & Iacobucci, 2004) as well as from the bootstrap are reported. We use 10 items from the anxiety inventory as an example. We use all the cases from the `msqR` data set. By default, items that are negatively correlated with the total score are *not* reversed. However, if we specify that `check.keys=TRUE`, then items with negative correlations with the total score are automatically reversed keys. A warning is produced.

R code

```
select  #show the items we want to score
alpha(msq1[select],check.keys=TRUE)  #find  alpha -- reverse code some items automatically
```

```
select  #show the items we want to score
 [1] "anxious"   "jittery"   "nervous"   "tense"     "upset"     "at.ease"   "calm"
```

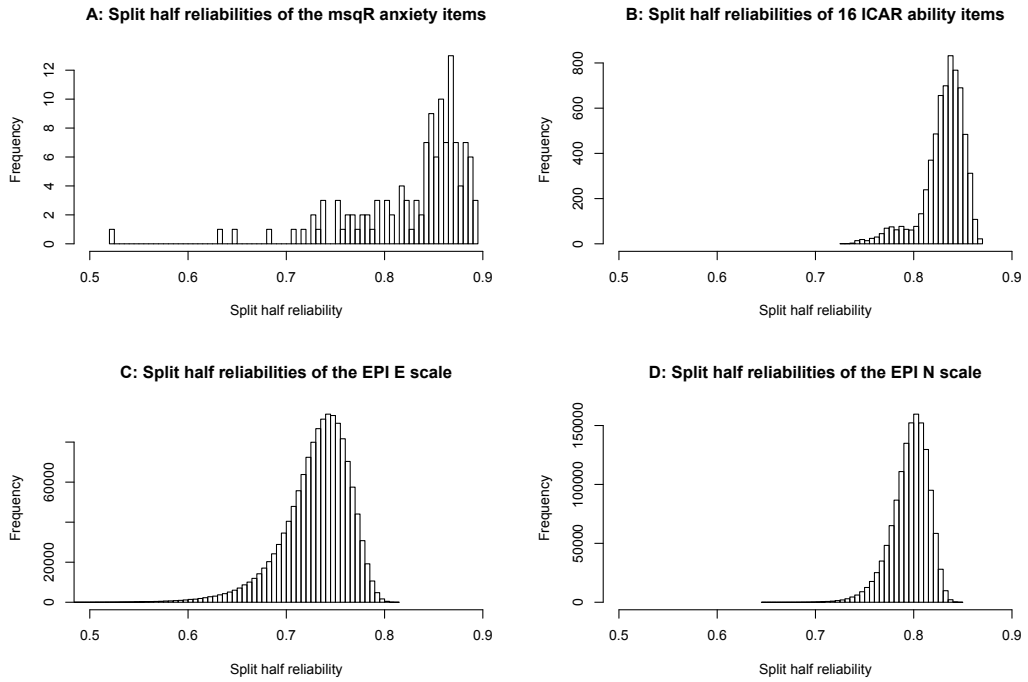

Figure 4. The distribution of 126 split half reliabilities for the 10 state anxiety items (panel A) and the 1,352,078 splits of the 24 EPI Extraversion items (panel C) suggests that the tests are not univocal while that of the 6,435 splits of the ICAR ability items (panel B) and the 1,352,078 splits of the EPI N scale (panel D) suggests greater homogeneity .

```
          "confident" "content"   "relaxed"

Reliability analysis
Call: alpha(x = msq1[select], 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 4.9 0.0046    2 0.53     0.31

 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-       0.82      0.82    0.85      0.34 4.5   0.0048 0.028  0.33
jittery-       0.82      0.82    0.85      0.34 4.7   0.0048 0.029  0.31
nervous-       0.82      0.81    0.84      0.33 4.4   0.0050 0.031  0.32
tense-         0.81      0.81    0.84      0.32 4.2   0.0052 0.030  0.30
upset-         0.82      0.82    0.85      0.34 4.7   0.0049 0.035  0.34
at.ease        0.80      0.80    0.83      0.31 4.1   0.0056 0.031  0.31
calm           0.80      0.81    0.84      0.32 4.2   0.0054 0.033  0.30
confident      0.83      0.83    0.85      0.36 5.0   0.0046 0.023  0.33
content        0.82      0.82    0.84      0.33 4.5   0.0050 0.027  0.32
relaxed        0.80      0.80    0.84      0.31 4.1   0.0055 0.033  0.30

 Item statistics
             n raw.r std.r r.cor r.drop mean   sd
anxious-  1871  0.58  0.60  0.55   0.45  2.3 0.86
jittery-  3026  0.53  0.56  0.49   0.42  2.3 0.83
nervous-  3017  0.59  0.65  0.60   0.52  2.6 0.67
tense-    3017  0.66  0.71  0.68   0.59  2.5 0.77
```

```
upset-    3020  0.52  0.56  0.48   0.43  2.7 0.66
at.ease   3018  0.77  0.73  0.71   0.66  1.6 0.93
calm      3020  0.73  0.70  0.67   0.62  1.6 0.90
confident 3021  0.52  0.47  0.41   0.36  1.5 0.91
content   3010  0.65  0.60  0.57   0.51  1.5 0.92
relaxed   3024  0.75  0.72  0.69   0.64  1.7 0.89

Non missing response frequency for each item
             0    1    2    3 miss
anxious   0.53 0.29 0.13 0.04 0.38
jittery   0.54 0.31 0.12 0.04 0.00
nervous   0.71 0.22 0.06 0.02 0.00
tense     0.61 0.28 0.09 0.03 0.00
upset     0.76 0.18 0.04 0.02 0.00
at.ease   0.13 0.32 0.36 0.18 0.00
calm      0.12 0.34 0.37 0.17 0.00
confident 0.14 0.33 0.38 0.15 0.00
content   0.17 0.35 0.35 0.13 0.01
relaxed   0.10 0.30 0.41 0.19 0.00
```

Now do it again, using the `omegaSem` function which calls the *lavaan* package (Rosseel, 2012) to do a SEM analysis and report both the EFA and CFA solutions. `omega` just reports the EFA solution. `omegaSem` reports both the EFA and the CFA solution. Note that they differ somewhat with lower ω_h estimates from the EFA solution (.42) as contrasted to the CFA solution (.55). By forcing cross loadings to 0 in the CFA, more variance is accounted for by the general factor.

R code

```
omegaSem(msq1[select],nfactors = 2)   #specify a two factor solution
```

```
Call: omegaSem(m = msq1[select], nfactors = 2)
Omega
Call: omega(m = m, nfactors = nfactors, fm = fm, key = key, flip = flip,
    digits = digits, title = title, sl = sl, labels = labels,
    plot = plot, n.obs = n.obs, rotate = rotate, Phi = Phi, option = option)
Alpha:                 0.83
G.6:                   0.86
Omega Hierarchical:    0.42
Omega H asymptotic:    0.48
Omega Total            0.87

Schmid Leiman Factor loadings greater than  0.2
              g   F1*   F2*   h2   u2   p2
anxious    0.38  0.62       0.53 0.47 0.27
jittery    0.35  0.53       0.41 0.59 0.30
nervous    0.42  0.59       0.53 0.47 0.33
tense      0.48  0.63       0.63 0.37 0.36
upset      0.32  0.29       0.22 0.78 0.47
at.ease-   0.51       -0.60 0.64 0.36 0.40
calm-      0.47  0.22 -0.47 0.49 0.51 0.45
confident- 0.29       -0.58 0.45 0.55 0.18
content-   0.40       -0.68 0.64 0.36 0.25
relaxed-   0.48  0.23 -0.47 0.51 0.49 0.46

With eigenvalues of:
  g F1* F2*
1.7 1.7 1.6

general/max  1.04   max/min =   1.01
mean percent general =  0.35    with sd =  0.1 and cv of  0.28
Explained Common Variance of the general factor =  0.34
```

```
The degrees of freedom are 26  and the fit is  0.25
The number of observations was  3032  with Chi Square =  756.57  with prob <  9.5e-143
The root mean square of the residuals is  0.04
The df corrected root mean square of the residuals is  0.05
RMSEA index =  0.096  and the 10 % confidence intervals are  0.09 0.102
BIC =  548.13

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  1.82
The number of observations was  3032  with Chi Square =  5493.51  with prob <  0
The root mean square of the residuals is  0.22
The df corrected root mean square of the residuals is  0.25

RMSEA index =  0.227  and the 10 % confidence intervals are  0.222 0.232
BIC =  5212.91

Measures of factor score adequacy
                                                 g  F1*  F2*
Correlation of scores with factors             0.66 0.78 0.79
Multiple R square of scores with factors       0.43 0.61 0.62
Minimum correlation of factor score estimates -0.14 0.22 0.23

 Total, General and Subset omega for each subset
                                                  g  F1*  F2*
Omega total for total scores and subscales    0.87 0.79 0.84
Omega general for total scores and subscales  0.42 0.28 0.31
Omega group for total scores and subscales    0.38 0.52 0.53

 The following analyses were done using the  lavaan  package

 Omega Hierarchical from a confirmatory model using sem =  0.55
 Omega Total  from a confirmatory model using sem =  0.88
With loadings of
              g  F1*  F2*   h2   u2   p2
anxious    0.41 0.60      0.53 0.47 0.32
jittery    0.47 0.41      0.39 0.61 0.57
nervous    0.44 0.60      0.55 0.45 0.35
tense      0.54 0.58      0.63 0.37 0.46
upset      0.34 0.32      0.22 0.78 0.53
at.ease-   0.66      0.47 0.66 0.34 0.66
calm-      0.69      0.31 0.58 0.42 0.82
confident-           0.77 0.61 0.39 0.02
content-   0.31      0.74 0.65 0.35 0.15
relaxed-   0.68      0.33 0.57 0.43 0.81

With eigenvalues of:
  g F1* F2*
2.5 1.3 1.6

The degrees of freedom of the confimatory model are  25  and the fit is  506.0981  with p =  0
general/max  1.57   max/min =   1.18
mean percent general =  0.47    with sd =  0.26 and cv of  0.57
Explained Common Variance of the general factor =  0.46

Measures of factor score adequacy
                                                 g  F1*  F2*
Correlation of scores with factors             0.86 0.80 0.88
Multiple R square of scores with factors       0.74 0.65 0.77
Minimum correlation of factor score estimates 0.48 0.29 0.55

 Total, General and Subset omega for each subset
                                                  g  F1*  F2*
Omega total for total scores and subscales    0.88 0.81 0.87
Omega general for total scores and subscales  0.55 0.35 0.41
Omega group for total scores and subscales    0.33 0.46 0.46

To get the standard sem fit statistics, ask for summary on the fitted object
```

Parallel Forms

The `sai` data set includes 20 items. 10 overlap with the `msqR` data set and are used for most examples. But we may also score anxiety from the second set of items. We can use either the `scoreItems` or the `scoreOverlap` functions. The latter function corrects for the fact that the positive and negative subsets of the anxiety scales overlap with the total scale.

R code

```
sai.parallel <- scoreOverlap(sai.alternate.forms,sai1)
sai.parallel
```

```
Call: scoreOverlap(keys = sai.alternate.forms, r = sai1)

(Standardized) Alpha:
pos1 neg1 anx1 pos2 neg2 anx2
0.86 0.82 0.87 0.83 0.73 0.80

(Standardized) G6*:
pos1 neg1 anx1 pos2 neg2 anx2
0.87 0.84 0.78 0.84 0.79 0.72

Average item correlation:
pos1 neg1 anx1 pos2 neg2 anx2
0.54 0.48 0.40 0.50 0.36 0.28

Number of items:
pos1 neg1 anx1 pos2 neg2 anx2
   5    5   10    5    5   10

Signal to Noise ratio based upon average r and n
pos1 neg1 anx1 pos2 neg2 anx2
 5.9  4.7  6.6  4.9  2.8  4.0

Scale intercorrelations corrected for item overlap and attenuation
 adjusted for overlap correlations below the diagonal, alpha on the diagonal
 corrected correlations above the diagonal:
      pos1  neg1  anx1  pos2  neg2  anx2
pos1  0.86 -0.60 -0.90  0.95 -0.57 -0.95
neg1 -0.50  0.82  0.89 -0.40  0.98  0.81
anx1 -0.77  0.75  0.87 -0.77  0.87  1.00
pos2  0.80 -0.33 -0.66  0.83 -0.41 -0.86
neg2 -0.45  0.76  0.70 -0.32  0.73  0.82
anx2 -0.78  0.66  0.83 -0.70  0.63  0.80

 In order to see the item by scale loadings and frequency counts of the data
 print with the short option = FALSE
```

Inter rater reliability using the `ICC` function

We use the same data as in Table 6. They are displayed here in a compact form and then analyzed.

R code

```
example <- structure(list(c(1, 1, 1, 4, 2, 3, 1, 3, 3, 5, 2), c(1, 1, 2,
1, 4, 4, 4, 4, 5, 5, 2.4), c(3, 3, 3, 2, 3, 3, 6, 4, 4, 5, 2.4
), c(2, 2, 2, 6, 5, 4, 4, 4, 6, 6, 3), c(3, 5, 4, 2, 3, 6, 6,
6, 5, 5, 3.4), c(2, 2.4, 2.4, 3, 3.4, 4, 4.2, 4.2, 4.6, 5.2,
4)), .Names = c("V1", "V2", "V3", "V4", "V5", "Mean"), row.names = c("S1",
"S2", "S3", "S4", "S5", "S6", "S7", "S8", "S9", "S10", "Mean"), class = "data.frame")
```

```
example #show it

 ICC(example[1:10,1:5])  #find the ICCs for the 10 subjects and 5 judges:
```

```
Call: ICC(x = example[1:10, 1:5])

Intraclass correlation coefficients
                         type  ICC   F df1 df2       p lower bound upper bound
Single_raters_absolute   ICC1 0.29 3.0   9  40 0.00748       0.045        0.66
Single_random_raters     ICC2 0.32 4.3   9  36 0.00078       0.085        0.67
Single_fixed_raters      ICC3 0.40 4.3   9  36 0.00078       0.125        0.74
Average_raters_absolute ICC1k 0.67 3.0   9  40 0.00748       0.190        0.91
Average_random_raters   ICC2k 0.70 4.3   9  36 0.00078       0.317        0.91
Average_fixed_raters    ICC3k 0.77 4.3   9  36 0.00078       0.418        0.93

 Number of subjects = 10     Number of Judges =  5
```

Reliability over time: the `multilevelReliability` function

In Table 3 we showed four subjects from Shrout & Lane (2012) and the associated analysis (Table 4). Here we show the R code for that analysis, as well as a more realistic analysis of 77 subjects with repeated measures on anxiety over four time points.

The data from Shrout & Lane (2012) are included in the help page for `multilevel.reliability`. We take those data, deleting the first subject, and then do the analysis. The data are stored in the help page in a compact form which we show here.

R code

```
shrout <- structure(list(Person = c(1L, 2L, 3L, 4L, 5L, 1L, 2L, 3L, 4L,
5L, 1L, 2L, 3L, 4L, 5L, 1L, 2L, 3L, 4L, 5L), Time = c(1L, 1L,
1L, 1L, 1L, 2L, 2L, 2L, 2L, 2L, 3L, 3L, 3L, 3L, 3L, 4L, 4L, 4L,
4L, 4L), Item1 = c(2L, 3L, 6L, 3L, 7L, 3L, 5L, 6L, 3L, 8L, 4L,
4L, 7L, 5L, 6L, 1L, 5L, 8L, 8L, 6L), Item2 = c(3L, 4L, 6L, 4L,
8L, 3L, 7L, 7L, 5L, 8L, 2L, 6L, 8L, 6L, 7L, 3L, 9L, 9L, 7L, 8L
), Item3 = c(6L, 4L, 5L, 3L, 7L, 4L, 7L, 8L, 9L, 9L, 5L, 7L,
9L, 7L, 8L, 4L, 7L, 9L, 9L, 6L)), .Names = c("Person", "Time",
"Item1", "Item2", "Item3"), class = "data.frame", row.names = c(NA,
-20L))

small.shrout <- shrout[-c(1,6,11,16),] #these are the data in the table
multilevel.reliability(small.shrout,grp="Person",Time="Time") #do the analysis
mlPlot(small.shrout,grp="Person",Time="Time",col=gray(c(0,.1,.2)),
       lty=cs(solid,dashed,dotted)) #draw the figure
```

See Table 4 for the output.

The VALE data set has four replications of the anxiety items. We use the `multilevel.reliability` function. The first and third were separated by several days, the second and fourth were 30 minutes following the first and third obserations.

R code

```
select #show the items to score
vale.mlr <- multilevel.reliability(vale,grp="id",Time="time",items=select)
vale.mlr    #show the output
```

```
select
 [1] "anxious"   "jittery"   "nervous"   "tense"     "upset"     "at.ease"   "calm"
```

```
             "confident" "content"   "relaxed"

Multilevel Generalizability analysis
Call: multilevel.reliability(x = vale, grp = "id", Time = "time", items = select)

The data had  77  observations taken over  4  time intervals for  10 items.

 Alternative estimates of reliabilty based upon Generalizability theory

RkF  =  0.9 Reliability of average of all ratings across all items and  times (Fixed time effects)
R1R  =  0.58 Generalizability of a single time point across all items (Random time effects)
RkR  =  0.84 Generalizability of average time points across all items (Random time effects)
Rc   =  0.37 Generalizability of change (fixed time points, fixed items)
RkRn =  0.7 Generalizability of between person differences averaged over time (time nested within people)
Rcn  =  0 Generalizability of within person variations averaged over items  (time nested within people)

 These reliabilities are derived from the components of variance estimated by ANOVA
             variance Percent
ID               0.04    0.04
Time             0.00    0.00
Items            0.35    0.36
ID x time        0.02    0.02
ID x items       0.28    0.29
time x items     0.01    0.01
Residual         0.28    0.29
Total            0.98    1.00

 The nested components of variance estimated from lme are:
         variance Percent
id        5.5e-02 5.6e-02
id(time)  1.6e-09 1.6e-09
residual  9.2e-01 9.4e-01
total     9.8e-01 1.0e+00

To see the ANOVA and alpha by subject, use the short = FALSE option.
 To see the summaries of the ICCs by subject and time, use all=TRUE
 To see specific ks select from the following list:
 ANOVA s.lmer s.lme alpha summary.by.person summary.by.time ICC.by.person ICC.by.time lmer long Call
```

Reliability of difference scores

The `ability` data set includes 1525 subjects and their scores on 16 items. We can find 4 subtests of these 16 items, or two higher level subtests (verbal and spatial). Although these two subtests have α reliabiities of .77 and .72, the reliability of their difference scroe is substantiailly less.

R code

```
#score the four subscales, as well as the two higher level composites
keys <- keys <- list(
 ICAR16=colnames(ability),reasoning =  cs(reason.4,reason.16,reason.17,reason.19),
  letters=cs(letter.7, letter.33,letter.34,letter.58, letter.7),
  matrix=cs(matrix.45,matrix.46,matrix.47,matrix.55),
  rotate=cs(rotate.3,rotate.4,rotate.6,rotate.8),
  verbal =cs(reason.4,reason.16,reason.17,reason.19,letter.7, letter.33,letter.34,
        letter.58, letter.7),
  spatial =cs(matrix.45,matrix.46,matrix.47,matrix.55, rotate.3,rotate.4,rotate.6,rotate.8),
  verbal_spatial = cs(reason.4,reason.16,reason.17,reason.19,letter.7, letter.33,letter.34,
    letter.58, letter.7, -matrix.45, -matrix.46, -matrix.47, -matrix.55,-rotate.3,
         -rotate.4,- rotate.6,-rotate.8)  )
 ability.scores <- scoreOverlap(keys,ability)
 summary(ability.scores)
 keys <- c(rep(1,8),rep(-1,8))
 ability.diff <- reverse.code(keys,ability)
 omega.diff <- omega(ability.diff, flip=FALSE)
 summary(omega.diff)
```

```
  summary(ability.scores)
Call: scoreOverlap(keys = keys, r = ability)

Scale intercorrelations adjusted for item overlap
Scale intercorrelations corrected for attenuation
 raw correlations (corrected for overlap) below the diagonal, (standardized) alpha on the diagonal
 corrected (for overlap and reliability) correlations above the diagonal:
              ICAR16 reasoning letters matrix rotate verbal spatial verbal_spatial
ICAR16         0.831      0.90    0.89  0.893   0.77   0.94   0.939           0.17
reasoning      0.658      0.65    0.80  0.774   0.54   0.95   0.737           0.39
letters        0.664      0.53    0.67  0.777   0.51   0.95   0.721           0.42
matrix         0.598      0.46    0.47  0.539   0.54   0.82   0.855           0.28
rotate         0.611      0.38    0.37  0.350   0.77   0.56   0.896          -0.40
verbal         0.753      0.67    0.68  0.529   0.43   0.77   0.774           0.43
spatial        0.727      0.50    0.50  0.534   0.67   0.58   0.722          -0.13
verbal_spatial 0.067      0.14    0.15  0.089  -0.15   0.16  -0.048           0.19

#the omega summary
Omega
Alpha:                 0.19
G.6:                   0.39
Omega Hierarchical:    0.04
Omega H asymptotic:    0.09
Omega Total            0.45
```

Kappa

κ is a measure of agreement between raters (Cohen, 1960). Consider the data in Table 7. We can apply the `cohen.kappa` function:

R code

```
#Get the data
ratings <- structure(list(R1 = structure(c(1L, 1L, 1L, 1L, 1L, 2L, 2L, 2L,
3L, 3L), .Label = c("Achieve", "Intimacy", "Power"), class = "factor"),
    R2 = structure(c(1L, 1L, 1L, 1L, 2L, 1L, 2L, 3L, 3L, 3L), .Label = c("Achieve",
    "Intimacy", "Power"), class = "factor"), R3 = structure(c(1L,
    2L, 2L, 3L, 1L, 1L, 2L, 2L, 2L, 3L), .Label = c("Achieve",
    "Intimacy", "Power"), class = "factor"), R4 = structure(c(3L,
    3L, 3L, 3L, 1L, 1L, 2L, 2L, 2L, 3L), .Label = c("Achieve",
    "Intimacy", "Power"), class = "factor")), row.names = c(NA,
-10L), class = "data.frame")

cohen.kappa(ratings)
```

Produces this output:

```
Cohen Kappa (below the diagonal) and Weighted Kappa (above the diagonal)
For confidence intervals and detail print with all=TRUE
     R1     R2   R3    R4
R1 1.00  0.783 0.30 -0.14
R2 0.52  1.000 0.29 -0.17
R3 0.24  0.130 1.00  0.52
R4 0.15 -0.014 0.57  1.00

Average Cohen kappa for all raters  0.27
Average weighted kappa for all raters  0.26
```

Item Response Theory estimates of reliability

The essential assumptions of Item Response Theory (IRT) is that items can differ in how hard they are, as well as how well they measure the latent trait. Although seemingly quite different

from classical approaches, there is a one-to-one mapping between the difficulty and discrimination parameters of IRT and the factor loadings and item response thresholds found by factor analysis of the polychoric correlations of the items (Kamata & Bauer, 2008; McDonald, 1999). The relationship of the IRT approach to classical reliability theory is given a very clear explication by Markon (2013) who examines how test information (and thus the reliability) varies by subject variance as well as trait level. A test can be developed to be reliable for certain discriminations (e.g. between psychiatric patients) and less reliable for discriminating between members of a control group. The particular strength of IRT approaches is the use in tailored or adaptive testing where the focus is on the reliability for a particular person at a particular level of the latent trait. By using IRT approaches, we can see that reliability varies as a function of the person's location on the underlying attribute dimension. We show this by examining the responses to the `sai` items using the `irt.fa` and `plot.Irt` functions (Figure 3).

R code

```
sai.irt <- irt.fa(sai1[4:23])
plot(sai.irt,type="test",main="Test information for the sai items")
sai.irt  #show the results
```

```
 Summary information by factor and item
 Factor =  1
                -3   -2   -1    0    1    2    3
calm          0.42 0.53 0.61 0.66 0.57 0.33 0.13
secure        0.35 0.48 0.54 0.55 0.46 0.29 0.13
tense         0.61 0.78 0.75 0.56 0.28 0.11 0.03
regretful     0.31 0.35 0.33 0.25 0.16 0.09 0.05
at.ease       0.47 0.62 0.65 0.74 0.70 0.51 0.20
upset         0.60 0.70 0.58 0.35 0.16 0.06 0.02
worrying      0.24 0.31 0.34 0.30 0.22 0.14 0.08
rested        0.10 0.14 0.18 0.20 0.21 0.18 0.15
anxious       0.27 0.35 0.38 0.33 0.24 0.15 0.08
comfortable   0.25 0.40 0.49 0.53 0.49 0.37 0.21
confident     0.17 0.23 0.27 0.27 0.24 0.18 0.12
nervous       0.52 0.61 0.55 0.38 0.20 0.09 0.03
jittery       0.26 0.33 0.34 0.28 0.20 0.13 0.07
high.strung   0.28 0.35 0.36 0.30 0.20 0.12 0.07
relaxed       0.35 0.59 0.65 0.74 0.72 0.56 0.25
content       0.23 0.39 0.52 0.57 0.51 0.36 0.19
worried       0.39 0.52 0.52 0.40 0.24 0.12 0.05
rattled       0.24 0.28 0.27 0.22 0.16 0.10 0.06
joyful        0.09 0.12 0.15 0.17 0.17 0.16 0.14
pleasant      0.19 0.33 0.46 0.51 0.46 0.34 0.20
Test Info     6.31 8.42 8.94 8.32 6.62 4.39 2.27
SEM           0.40 0.34 0.33 0.35 0.39 0.48 0.66
Reliability   0.84 0.88 0.89 0.88 0.85 0.77 0.56
```

Updates to R and the psych package

The R Core Team maintains and improves R. Updates are usually added every six-12 months, (numbered with a trailing zero. e.g., 3.6.0). Minor corrections are then added in the subsequent months (indicated by a change in the last digit). These new releases may be downloaded from CRAN at https://cran.r-project.org. The *psych* package is also improved with new releases added about every 4-6 months. The numbering system of *psych* reflects the year and month of release (e.g., 1.8.12 was released to CRAN in December, 2018.) Development versions of *psych* may be downloaded from the Personality-Project web site at https://personality-project.org/r.

References

Anderson, K. J., & Revelle, W. (1983). The interactive effects of caffeine, impulsivity and task demands on a visual search task. *Personality and Individual Differences*, *4*(2), 127-134. doi: 10.1016/0191-8869(83)90011-9

Anderson, K. J., & Revelle, W. (1994). Impulsivity and time of day: Is rate of change in arousal a function of impulsivity? *Journal of Personality and Social Psychology*, *67*(2), 334-344. doi: 10.1037/0022-3514.67.2.334

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67(1), 1-48*, *67*(1), 1-48. (R package version 1.1-8) doi: 10.18637/jss.v067.i01.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(37-46). doi: 10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213-220. doi: 10.1037/h0026256

Cole, D. A., Martin, N. C., & Steiger, J. H. (2005). Empirical and conceptual problems with longitudinal trait-state models: Introducing a trait-state-occasion model. *Psychological Methods*, *10*(1), 3–20. doi: 10.1037/1082-989X.10.1.3

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

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*(2), 322 - 328. doi: 10.1037/0033-2909.88.2.322

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin*, *32*(7), 917-929. doi: 10.1177/ 0146167206287721

DeSimone, J. A. (2015). New techniques for evaluating temporal consistency. *Organizational Research Methods*, *18*(1), 133-152. doi: 10.1177/1094428114553061

Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ase): An accurate and precise confidence interval estimate. *Journal of Applied Psychology*, *89*(5), 792-808.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*(4), 341-349. doi: 10.1037/1040-3590.8.4.341

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*(4), 407–433. doi: 10.1007/BF02294564

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, N.J.: L. Erlbaum Associates.

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

Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology*. doi: 10.1027/1614-2241/a000086

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*(3), 613-619. doi: 10.1177/001316447303300309

Fox, J. (2016). *Applied regression analysis and generalized linear models* (3rd ed.). Sage.

Freedman, R., Lewis, D. A., Michels, R., Pine, D. S., Schultz, S. K., Tamminga, C. A., ... Yager, J. (2013). The initial field trials of DSM-5: New blooms and old thorns. *American Journal of Psychiatry*, *170*(1), 1-5. doi: 10.1176/appi.ajp.2012.12091189

Guo, J., Klevan, M., & McAdams, D. P. (2016). Personality traits, ego development, and the redemptive self. *Personality and Social Psychology Bulletin*, *42*(11), 1551-1563. doi: 10.1177/0146167216665093

Hubert, L. (1977). Kappa revisited. *Psychological Bulletin*, *84*(2), 289 - 297. doi: 10.1037/0033-2909.84.2.289

Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance: A theory of the relationship between individual differences and information processing. *Psychological Review*, *91*(2), 153-184. doi: 10.1037/0033-295X.91.2.153

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(1), 136-153. doi: 10.1080/10705510701758406

Krippendorff, K. (1970). Bivariate agreement coefficients for reliability of data. *Sociological Methodology*, *2*, 139–150. doi: 10.2307/270787

Krippendorff, K. (2004, 7). Reliability in content analysis. *Human Communication Research*, *30*(3), 411–433. doi: 10.1111/j.1468-2958.2004.tb00738.x

Krippendorff, K., & Fleiss, J. L. (1978). Reliability of binary attribute data. *Biometrics*, *34*(1), 142–144.

Kuder, G., & Richardson, M. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151-160. doi: 10.1007/BF02288391

Larsen, R. J., & Diener, E. (1992). Promises and problems with the circumplex model of emotion. In M. S. Clark (Ed.), *Emotion* (p. 25-59). Thousand Oaks, CA: Sage Publications, Inc.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, *76*(5), 365 - 377. doi: 10.1037/h0031643

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, *51*(5), 493 - 504. doi: 10.1037/h0058543

Lumsden, J. (1976). Test theory. *Annual Review of Psychology*, *27*, 251-280. doi: 10.1146/annurev.ps.27.020176.001343

Markon, K. E. (2013). Information utility: Quantifying the total psychometric information provided by a measure. *Psychological Methods*, *18*(1), 15-35. doi: 10.1037/a0030638

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

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rafaeli, E., & Revelle, W. (2006). A premature consensus: Are happiness and sadness truly opposite affects? *Motivation and Emotion*, *30*(1), 1-12. doi: 10.1007/s11031-006-9004-2

Rafaeli, E., Rogers, G. M., & Revelle, W. (2007). Affective synchrony: Individual differences in mixed emotions. *Personality and Social Psychology Bulletin*, *33*(7), 915-932. doi: 10.1177/0146167207301009

Revelle, W. (2019a, June). psych: Procedures for personality and psychological research [Computer software manual]. https://CRAN.r-project.org/package=psych. Retrieved from https://CRAN.R-project.org/package=psych (R package version 1.9.6)

Revelle, W. (2019b, June). psychtools: Tools to accompany the psych package for psychological research [Computer software manual]. https://CRAN.r-project.org/package=psychTools. Retrieved from https://CRAN.R-project.org/package=psychTools (R package version 1.9.6)

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

Revelle, W., Condon, D. M., Wilt, J., French, J. A., Brown, A., & Elleman, L. G. (2016). Web and phone based data collection using planned missing designs. In N. G. Fielding, R. M. Lee, & G. Blank (Eds.), *Sage handbook of online research methods* (2nd ed., p. 578-595). Sage Publications, Inc.

Revelle, W., Humphreys, M. S., Simon, L., & Gilliland, K. (1980). Interactive effect of personality, time of day, and caffeine: A test of the arousal model. *Journal of Experimental Psychology General*, *109*(1), 1-31. doi: 10.1037/0096-3445.109.1.1

Revelle, W., & Wilt, J. A. (2019). Analyzing dynamic data: a tutorial. *Personality and Individual Differences*, *136*(1), 38-51. doi: /10.1016/j.paid.2017.08.020

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi: 10.18637/jss.v048.i02

RStudio Team. (2016). Rstudio: Integrated development environment for r [Computer software manual]. Boston, MA. Retrieved from http://www.rstudio.com/

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *The Public Opinion Quarterly*, *19*(3), 321-325. doi: 10.1086/266577

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420-428. doi: 10.1037/0033-2909.86.2.420

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

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

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Thayer, R. E. (1978). Toward a psychological theory of multidimensional activation (arousal). *Motivation and Emotion*, *2*(1), 1-34. doi: 10.1007/BF00992729

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

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063-1070. doi: 10.1037/0022-3514.54.6.1063

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, *8*(4), 454-464. doi: 10.1177/1948550617703168

Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, *16:93*. doi: 10.1186/s12874-016-0200-9