

Classical Test Theory

Reliability

Christophe Lalanne
ch.lalanne@gmail.com

November, 2009

Summary

“ Here we are interested in assessing the reliability of test scores, that is the reproducibility of the observed results on a given questionnaire, in the context of the CTT framework. ”

© 2009, www.aliquote.org

Outline

- 1.
- 2.
- 3.
- 4.

*Various functions used throughout this chapter were collated in the package **Psychomisc**.*

© 2009, www.aliquote.org

1

What are the sources of scores reliability?

A better formulation is: What are the sources of scores variability? Then, other questions arise such as: How can we measure it? How does it impact inference made on subjects population?

Replicated measurement may occur as a consequence of [11, Chap. 2]:

- repeated assessment of several subjects by the same rater or clinician;
- alternative assessment of a given subject by different raters;
- alternative administration of the same questionnaire or a parallel form;
- use of different subscales of a single questionnaire to infer one's performance.

© 2009, www.aliquote.org

2

Ways to measure reliability

Reliability can be assessed using:

- linear decomposition of variance components in CTT,
- structural equation model,
- item response theory model.

Hereafter, we shall focus on the first option but we will show how all three measures converge albeit focusing on a different conception of what is a score and how it is measured.

© 2009, www.aliquote.org

3

Underlying concepts

One must distinguish between *reliability* and *significance* [33, p. 7]:

- “statistical” significance evaluates the probability or likelihood of the sample results, with reference to a reference population where the null is exactly true; “practical” and “clinical” significance are closely related concepts underlying the extent to which sample results diverge from the null (as measured by effect size statistics) or the way treated patients may not be distinguished from control or normal ones.
- However, “statistical”, “practical”, and “clinical” significance all stand on the assumption that scores are meaningful and reliable indicators of individuals’ performance. . .

Caveats

Classical measures of association such as chi-square, phi coefficients, etc. are not measures of reliability *per se*. Indeed, they only show that scores are correlated in some sense but they do not necessarily underly the same construct (which is the purpose of the measurement).

One must also bear in mind that structural stability is different from temporal stability, and that we are always assuming that the construct does not evolve between occasion. This is particularly important when studying responsiveness.

Reliability is not a fixed property



Reliability is not a property of a measurement instrument itself but of scores delivered throughout it to some individuals sampled from a larger population, hence it is not a fixed parameter.

It is important to remember that a test is not reliable or unreliable. Reliability is a property of the scores on a test for a particular population of examinees (Feldt & Brennan, 1989). Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric.

Wilkinson & APA Task Force, [22]

Gold standards for reliability index?

In fact, there is no standard of reliability although some authors did propose practical rule of thumbs, e.g. 0.70–0.90 for Cronbach’s alpha [20, 35, p. 264].

Measurement model

From CTT, we already know that a given individual score, x_i , may be expressed as the sum of its *true* score plus an additional measurement error which comes from the fact that we assessed his performance once.

$$x_i = \tau_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0; \sigma_e^2). \quad (1)$$

As can be seen, $\mathbb{E}(x) = \tau$, but how about the error terms? In fact, we could also start from:

$$\mathbb{V}(X) = \mathbb{V}(T) + \mathbb{V}(E), \quad (2)$$

with the additional assumption that T and E are independent.

Measurement model (Con't)

The *coefficient of reliability* of X is defined as

$$\begin{aligned} R_X &= \frac{\mathbb{V}(T)}{\mathbb{V}(X)} \\ &= \frac{\mathbb{V}(T)}{\mathbb{V}(T) + \mathbb{V}(E)}. \end{aligned} \quad (3)$$

Although it amounts to a standard coefficient of determination, or the η^2 in the ANOVA framework, here we will treat this quantity as a reliability ratio whose square root is called the standard error of measurement (SEM). Note that it is a random variable, i.e. it is not a fixed characteristic of the instrument.

Extension of our simple measurement model

Now, suppose that ratings depend not only on subject's true score and random measurement error, but also on the rater (I). Then,

$$\mathbb{V}(X) = \mathbb{V}(T) + \mathbb{V}(I) + \mathbb{V}(E). \quad (4)$$

Now, what is the reliability of the instrument? If all subjects are assessed by only one rater, R_X is the same as in Equation 3. But, if subjects are each going to be assessed by a rater randomly drawn from a large pool of raters, then

$$R'_X = \frac{\mathbb{V}(T)}{\mathbb{V}(T) + \mathbb{V}(I) + \mathbb{V}(E)}, \quad (5)$$

and obviously $R_X > R'_X$.

Another extension: Two instruments

Suppose we now have a collection of paired observations on two measurement instruments, say \mathcal{I}_X and \mathcal{I}_Y , where sum scores are expressed under the same linear decomposition:

$$X = T_X + E_X$$

$$Y = T_Y + E_Y$$

What is the correlation between X and Y ?



Be aware that both measures are random realizations of X and Y but above all fallible indicators. . . Reliabilities of X and Y should be taken into account.

Another extension: Two instruments (Con't)

The true correlation between X and Y is given by $\rho_{XY} = \frac{\text{cov}(T_X, T_Y)}{\sqrt{\text{V}(T_X)\text{V}(T_Y)}}$. But, our estimate will be

$$\begin{aligned}\hat{\rho}_{XY} &= \frac{\text{cov}(X, Y)}{\sqrt{\text{V}(X)\text{V}(Y)}} \\ &= \frac{\text{cov}(T_X + E_X, T_Y + E_Y)}{\sqrt{\text{V}(T_X + E_X)\text{V}(T_Y + E_Y)}} = \rho_{XY} \sqrt{R_X R_Y}\end{aligned}\quad (6)$$

Hence, $\hat{\rho}_{XY}$ is a shrunk estimator of the *true* correlation. In fact, the correlation of empirical measures is subjected to *attenuation*.

Another extension: Two instruments (Con't)

If we were using a linear regression to predict Y from X , with a model like $\mathbb{E}(T_Y|T_X) = \beta_0 + \beta_1 T_X$, the true slope would be $\beta_1 = \text{cov}(T_X, T_Y) / \text{V}(T_X)$. However, we actually estimate

$$\begin{aligned}\hat{\beta}_1 &= \frac{\text{cov}(X, Y)}{\text{V}(X)} \\ &= \frac{R_X \text{cov}(T_X, T_Y)}{\text{V}(T_X)} = R_X \beta_1\end{aligned}\quad (7)$$

The regression slope is also shrunk toward zero.

Internal consistency: Cronbach's alpha

The Cronbach's alpha is a well-known index used thoroughly in scientific publications to ascertain the reliability of an instrument. However, it is worth noting that it is no more than an indicator of variance shared by all items considered in its calculation. Therefore, it should not be confused with an absolute measure of reliability.

It is readily available in any decent statistical software [32], and also in several R packages.

If the variables being tested are all dichotomous, Cronbach's alpha is the same as Kuder-Richardson coefficient. Note that when alpha is .70, the standard error of measurement will be over half (0.55) a standard deviation. Alpha increases when increasing the number of items.

In addition, the following assumptions are made: (i) no residual correlations, (ii) items have identical loadings, and (iii) the scale is unidimensional.

Inside the Cronbach's alpha

For dichotomous items, [18] proposes as a measure of internal consistency the KR-20 coefficient which is

$$KR-20 = \frac{K}{K-1} \left[1 - \frac{\sum_k p_k q_k}{\sigma_t^2} \right] \quad (8)$$

where K is the number of items, σ_t^2 the total variance of test scores, and p_k and q_k are the proportions of individuals scoring 1 or 0 on the k th item.

With polytomous items, we just have to replace $p_k q_k$ with $\sigma_{k_c}^2$, the variance of the k th item [8].

Basically, it reads like an R^2 or η^2 coefficient since it is just a ratio of variances, with explained variance in the numerator (variance accounted by

items responses). But, it is not as simple and Cronbach's alpha can be negative!

Internal consistency: Numerical example

Consider the following example from [33]:

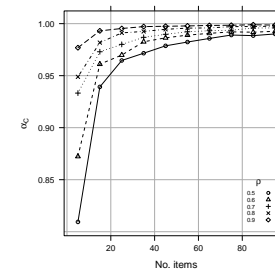
```
a <- matrix(scan('ex1.dat', sep=',', skip=1), nc=3, byrow=T)
cronbach.alpha(a)
cronbach.alpha.boot(a)
```

Here, we have $\alpha_C = 0.899$ with 95% CI [0.703;0.973] (gaussian) or [0.752;0.952] (bootstrap). Should we be interested in testing whether this value is different from 0.700, we could do a modified F -test as proposed by [5]:

```
cronbach.alpha.test(a)
```

which gives a p -value of 0.024 suggesting that we should reject the null.

Internal consistency: Illustrative simulations (1)



For a fixed sample size ($N = 300$), Cronbach's alpha increases when the number of items and inter-items correlation (ρ) are also increasing.

Even with modest (albeit perfect) correlation between items, e.g. $\rho = 0.35$, Cronbach's alpha would still be at 0.943 with 30 items (and 0.910 with 20 items).

Varying sample size does not modify this pattern of variation.

Internal consistency: Illustrative simulations (2)

We could also add random noise around pairwise correlation so as to better reflect real-life situations.

Internal consistency: Illustrative simulations (3)

Now, we use a band-diagonal structure for the correlation matrix, that is we simulate clusters of correlated items instead of assuming a perfect correlation between all items.

Exemple of reliability analysis

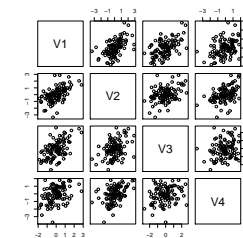
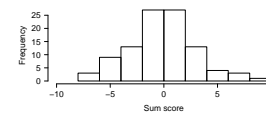
Let's consider scores generated according to a congeneric test (i.e., tests where scores may be expressed as $x_{ij} = \lambda_i F_i + \varepsilon_{ij}$, with both unequal factor loadings and error variances).

```
require(psych)
set.seed(100)
x <- sim.congeneric(N=100, short=FALSE)
pairs(x$observed)
alpha(x$observed)
my.summary(apply(x$observed,1,sum))
```

Mean sum score is -0.16 ± 3.06 (range, $-6.29-8.32$).

Exemple of reliability analysis (Con't)

	V1	V2	V3	V4
V1	1.00	0.53	0.54	0.36
V2	0.53	1.00	0.37	0.44
V3	0.54	0.37	1.00	0.22
V4	0.36	0.44	0.22	1.00



The items correlation matrix and distribution of items and test scores are shown above.

Exemple of reliability analysis (Con't)

Statistics for the four items and overall test reliability are reported in the Table below.

```
mean(cor(x$observed)[upper.tri(cor(x$observed))])
```

	raw α	std α	G6	\bar{r}	mean	SD
Overall	0.73	0.74	0.7	0.41	-0.16	3.1
V1	0.61	0.61	0.53	0.35	-	-
V2	0.63	0.64	0.58	0.37	-	-
V3	0.70	0.70	0.62	0.44	-	-
V4	0.73	0.74	0.66	0.48	-	-

What's wrong with Cronbach's alpha?

- Cronbach alpha is a measure of internal consistency, it is not a measure of unidimensionality and can't be used to infer unidimensionality [10].
- As we said before, reliability is sample-dependent. Therefore, α estimate should be based on observed responses, not on known property of instrument itself.
- Alpha is not items-invariant as it increases with items number. As a consequence, comparison of alpha levels between scales with differing numbers of items is not appropriate.

[25, 26]

The sole case where alpha will be essentially the same as reliability is the case of uniformly high factor loadings, no error covariances, and unidimensional instrument [24].

Finally, as highlighted by L. J. Cronbach himself [9],

Coefficients are a crude device that does not bring to the surface many subtleties implied by variance components. In particular, the interpretations being made in current assessments are best evaluated through use of a standard error of measurement.

Alternatives to Cronbach's alpha

Guttman's Lambda 6, or G6 [16], considers the amount of variance in each item that can be accounted for the linear regression of all of the other items (the squared multiple correlation or smc), or more precisely, the variance of the errors, e_j^2 , and is

$$G6 = 1 - \sum(e^2)/Vx = 1 - \sum(1 - r^2(smc))/Vx. \quad (9)$$

The squared multiple correlation is a lower bound for the item communality and as the number of items increases, becomes a better estimate.

Revelle's beta [27]

Ordinal reliability alpha. [37] use a polychoric correlation matrix input to calculate alpha parallel to Cronbach. Their simulation studies lead them to conclude that ordinal reliability alpha provides "consistently suitable estimates of the theoretical reliability, regardless of the magnitude of the theoretical reliability, the number of scale points, and the skewness of the scale point distributions. In contrast, coefficient alpha is in general a negatively biased estimate of reliability" for ordinal data (p. 21). Ordinal reliability alpha will normally be higher than the corresponding Cronbach's alpha.

Sijtsma's glb [30], but see also [36]

The nonequivalence of α , β , and ω_h suggests that important information about the psychometric properties of a scale may be missing when scale developers and users only report α as is almost always the case.

Zinbarg et al., [36]

The Kappa statistic

The Kappa (κ) statistic [6] is a quality index that compares observed agreement between 2 raters on a nominal or ordinal scale with agreement expected by chance alone (as if raters were tossing up). More formally, it reads

$$\kappa = \frac{\overbrace{\sum_i \pi_{ii}}^{\text{raw agreement}} - \sum_i \pi_{i\bullet} \pi_{\bullet i}}{1 - \underbrace{\sum_i \pi_{i\bullet} \pi_{\bullet i}}_{\text{random ratings}}} \quad (10)$$

Extensions for the case of multiple raters exist [29, pp. 284–291]. Fleiss [14] provided guidelines to interpret κ values but these are merely rules of thumbs.

Weighted Kappa

[7]

Use of κ in diagnostic agreement

| Reproduced from [11, p. 18]

Consider the following cross-classification into cases vs. non-cases of n patients assessed by two psychiatrists, A and B .

		A		Total
		Case	Non-case	
B	Case	p_{11}	p_{12}	$p_{1\bullet}$
	Non-case	p_{21}	p_{22}	$p_{2\bullet}$
	Total	$p_{\bullet 1}$	$p_{\bullet 2}$	1

Use of κ in diagnostic agreement (Con't)

Let $P_o = p_{11} + p_{22}$ denotes the raw agreement and $P_c = p_{1\bullet}p_{\bullet 1} + p_{2\bullet}p_{\bullet 2}$ the chance-expected value. Then the sampling variance of $\kappa = (P_o - P_c)/(1 - P_c)$ is

$$\mathbb{V}(\kappa) = (A + B - C)/N(1 - P_c)^4,$$

where

$$A = p_{11}[(1 - P_o) - (p_{\bullet 1} + p_{1\bullet})(1 - P_o)]^2 + p_{11}[(1 - P_o) - (p_{\bullet 1} + p_{1\bullet})(1 - P_o)]^2$$

$$B = (1 - P_o)^2[p_{12}(p_{\bullet 1} + p_{2\bullet})^2 + p_{21}(p_{\bullet 2} + p_{1\bullet})^2]$$

and

$$C = (P_o P_c - 2P_c + P_o)^2$$

The Kappa statistic: Asymptotic distribution

The κ coefficient is asymptotically equivalent to the ICC estimated from a two-way random effects ANOVA (cf. Equation 4), but significance tests and SE coming from the usual ANOVA framework are not valid anymore with binary data.

As for Crobach's alpha, or more generally rank correlations, bootstrapping allows to obtain 95% CI for κ without making any distributional assumption.

Example of inter-rater reliability: The Kappa statistic

Two pathologists rated 118 tumors on a scale of 1 to 5 [1, p. 368].

```
data(pathologist.dat, package="exactLoglinTest")
aa <- xtabs(y~A+B,data=pathologist.dat)
```

		Rater B				
		1	2	3	4	5
Rater A	1	22	5	0	0	0
	2	2	7	2	1	0
	3	2	14	36	14	3
	4	0	0	0	7	0
	5	0	0	0	0	3

Raw agreement equals 0.636
 $(\text{sum}(\text{diag}(\text{aa}))/\text{sum}(\text{aa}))$.
 Ordinary $\kappa = 0.498$, with 95% CI (bootstrap) [0.386; 0.604].
 Weighted $\kappa = 0.779$, with 95% CI (bootstrap) [0.690; 0.848].

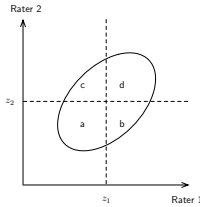
Tetrachoric and Polychoric Correlation

Tetrachoric (binary data, [23]) and polychoric (ordinal data) correlation coefficient may be used as a measure of inter-rater agreement. Indeed, they allow to

- estimate what would be the correlation if ratings were made on a continuous scale,
- test marginal homogeneity between raters.

According to J. Uebersax [34], tetrachoric and polychoric correlation models are special cases of latent trait modeling, which allows to relax distributional assumptions.

Illustration of the tetrachoric correlation



The picture on the left is a graphical summary of the proportions for the 2×2 cross-classification of the raters' ratings:

		Rater 1		
		-	+	
Rater 2	-	a	b	a + b
	+	c	d	c + d
		a + c	b + d	1

Here $a = \Pr(Y_1 < z_1 \text{ and } Y_2 < z_2)$ is the CDF of the bivariate normal distribution, specifically $\int_{-\infty}^{z_2} \int_{-\infty}^{z_1} \Phi(x, y, r) dx dy$, where $\Phi(\bullet)$ is the density of the bivariate normal distribution, and $\Phi(z_1) = a + c$ and $\Phi(z_2) = a + b$ (cut-off values).

Association statistics for 2×2 table

We shall assume a bivariate normal distribution with unknown parameters for the two ratings, say Y_1 and Y_2 , with mean $(0, 0)$ and covariance matrix $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

In this case, an approximation of tetrachoric ρ is Yule's $Q = (ad - bc)/(ad + bc)$, which is has an approximate normal distribution with mean 0 and variance estimated by

$$V(Q) = \frac{1}{4}(1 - Q^2)^2 \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \right). \quad (11)$$

Formal tests of independence in this case are provided by the Fisher exact test and the Yates-corrected chi-squared test.

Association statistics for 2×2 table (Con't)

The other approximation, initially proposed by [23], is

$$\sin \left\{ \frac{\pi}{2} \left(\frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} \right) \right\}. \quad (12)$$

The significance of ρ (where we consider $H_0: \rho - \rho_0 = 0$) is usually assessed using a Wald test statistic, $W = (r\rho_0)/se(r)$, where $se(r)$ is the standard error computed at $\rho = \rho_0$, but see also [21]. Now, r and $se(r)$ may be estimated using an exact or approximate way.

Computation of tetrachoric correlation coefficient are readily available in several statistical software, and is discussed in [4, 3] (Algorithm AS 116) and [12] (the latter being used in Stata).

Exact vs. approximate estimation for tetrachoric ρ

Exact approach:

Following [23], [3] proposed to compute $se(r)$ as

$$se(r) = \frac{1}{N^{3/2}} \Phi(z_1, z_2, r) \left\{ (a+d)(b+d)/4 + (a+c)(b+d)\Phi_2^2 + (a+b)(c+d)\Phi_1^2 + 2(ad-bc)\Phi_1\Phi_2 - (ab-cd)\Phi_2 - (ac-bd)\Phi_1 \right\}^{1/2},$$

$$\text{with } \Phi_1 = \Phi \left(\frac{z_1 - r z_2}{\sqrt{1-r^2}} \right) - 0.5, \quad \Phi_2 = \Phi \left(\frac{z_2 - r z_1}{\sqrt{1-r^2}} \right) - 0.5, \quad \text{and } \Phi(z_1, z_2, r) = \frac{1}{2\pi(1-r^2)^{1/2}} \exp \left[-\frac{z_1^2 - 2r z_1 z_2 + z_2^2}{2(1-r^2)} \right].$$

Exact vs. approximate estimation for tetrachoric ρ

Approximation approach:

Bonett and Price [2] proposed to estimate ρ by $\rho^* = \cos(\pi/(1 + \omega c))$, where

$$c = (1 - |(a + b) - (a + c)|/5 - (0.5 - p_m)^2) / 2$$

with p_m the smallest marginal proportion, and $\omega = ad/bc$.

✎ The exact computation of the tetrachoric correlation coefficient is difficult, mainly due to computational burden. Moreover, owing to the discrete nature of frequency data, the estimation of a cell probability can be no more accurate than $1/(2N)$ which may yield inaccurate estimates of ρ for cell frequencies < 5 [4, 2].

Non-parametric testing for tetrachoric ρ

A permutation test has been developed for the case of small sample sizes and/or disproportionate marginal frequency totals [19].

Summary for tetrachoric ρ

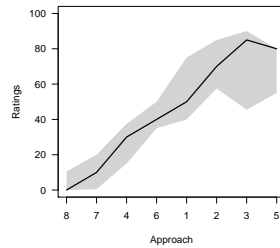
Finally, it should be noted that tetrachoric correlation matrices in SEM often provide very inflated chi-square values and underestimated standard errors of estimates due to larger variability than Pearson's r . Moreover, tetrachoric correlation can yield a nonpositive definite correlation matrix because eigenvalues may be negative (reflecting violation of normality, sampling error, outliers, or multicollinearity of variables) [15].

Example of inter-rater reliability: Use of ICC

The Table below summarizes ratings on a 0–100 scale of eight approaches to resume HRQL by seven North American experts [17].

Approach	Rater							Mean	SD
	1	2	3	4	5	6	7		
1	90	00	50	95	30	60	50	53.57	33.00
2	90	00	70	100	60	55	80	65.00	32.79
3	90	51	40	90	25	100	85	68.71	29.44
4	30	52	05	30	–	10	40	27.93	17.78
5	80	50	80	60	80	50	100	71.43	18.64
6	30	100	05	50	50	40	40	45.00	28.72
7	20	70	00	20	10	00	01	17.29	24.86
8	20	90	00	00	00	00	01	16.57	33.18

Example of inter-rater reliability



Median±IQR is shown on the left. Interestingly, approaches rated as the most adequate ones are those who are associated with larger variability between raters, although the rank (Spearman) correlation between mean and SD is not significant ($\rho = -0.167, p = 0.703$).

Analysis in R

The next commands allow to estimate the one-way model ❶, the two-way mixed effects (fixed rater effect) ❷, and the two-way random effects model ❸.

```
data(hays05)
require(lme4)
summary(aov(score~approach, hays05)) ❶
summary(aov(score~rater+approach, hays05)) ❷
summary(aov(score~approach*rater, hays05)) ❸
summary(lmer(score~approach+(1|rater), hays05)) ❹
```

Summary of ANOVAs

Source	Df	MS
Approach	7	3597.8
Within	47	792.6
Rater	6	844.7
Approach×Rater	41	785.0
Total	54	

Values for MS slightly differ from those of [17] in their Table 1.3.2.

Summary of ANOVAs

Model	Reliability	ICC
❶	$\frac{MS_B - MS_W}{MS_B}$	$\frac{MS_B - MS_W}{MS_B + (K-1)MS_W}$
❷	$\frac{MS_B - MS_E}{MS_B}$	$\frac{MS_B - MS_E}{MS_B + (K-1)MS_E}$
❸	$\frac{N(MS_B - MS_E)}{NMS_B + MS_J - MS_E}$	$\frac{MS_B - MS_E}{MS_B + (K-1)MS_E + K(MS_J - MS_E)/N}$

MS, mean square for between-subject (B), within-subject (W) effects; K, number of replications; N, number of rates.

Model ❷ is widely used in practice and gives rise to so-called “consistency” and “agreement” ICCs [28], although the “consistency” ICC is generally computed without considering the Approach×Rater interaction.

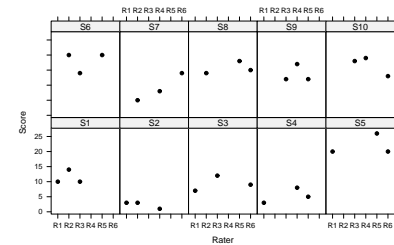
Example with a BIBD

The following results come from a BIBD study of depression by [13]. Each of ten subjects is rated by three raters.

Rater	Subject									
	1	2	3	4	5	6	7	8	9	10
1	10	3	7	3	20					
2	14	3				20	5	14		
3	10		12			14		12	18	
4		1		8			8		17	19
5				5	26	20		18	12	
6			9		20		14	15		13

What is the reliability of the ratings?

Example with a BIBD (Con't)



This plot shows that both Subject and Rater factors carry on a large part of variance in observed test scores.

Example with a BIBD (Con't)

Testing for overall mean difference between raters (Table 4 in [?]) yields:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Subject	9	982.00	109.11	11.76	0.0000
Rater	5	35.44	7.09	0.76	0.5898
Residuals	15	139.22	9.28		

To obtain such a Table, the Subject effect should be accounted first in the ANOVA. Indeed, as the design is not completely balanced, sum of squares depend on the order of model terms (unless you rely on Type III SS, which is not useful here).

Example with a BIBD (Con't)

Another approach consists in estimating the *intraclass correlation coefficient*, which quantifies the proportion of variance attributable to the between-subject variation, in this case $MS(rater)/(MS(rater) + MS(residuals))$.

Here, we obtain the relevant MS when Rater effect is entered first in the ANOVA model:

```
anova(lm(Score ~ Rater + Subject, a.df))
```

This gives an ICC of 0.770, which is rather high and suggests that most of the variability in test scores is due to heterogeneity of ratings.

Reliability of true score

Back to the fundamentals of TCT (but see the chapter on *Generalizability Theory*), let's make the assumption that the set of items we are working on is a random sample of a larger set of such items (i.e. items universe), and that inter-item correlation is not a parameter but a fixed constant.

Any individual sum score computed on such a random set of items can be seen as a *true score*, or domain/universe score [20, p. 217]. But, what about the reliability of the measure (of variance σ^2)?

Relation between the true score and its measure

Any empirical measurement is entangled of a measurement error (with $\mathbb{V} = \sigma_e^2$, see Eq. (2)), hence we would not expect a perfect correlation between the true score and its measure. In fact, [20, p. 218] shows that

$$r = \frac{k\bar{r}_{ij}}{1 + (k-1)\bar{r}_{ij}} \quad (13)$$

where r_m stands for the (fixed) correlation between the k items.

Relation between the true score and its measure (Con't)

Now, what about the variance of the measurement error, σ_e^2 ? It is related to σ^2 by the following relation:

$$\sigma_e = \sigma\sqrt{1 - \alpha_c} \quad (14)$$

where α_c is the Cronbach's alpha, or equivalently the r^2 used in Eq. 13.

A 95% CI can easily be derived as $x \pm 1.96\hat{\sigma}_e$.

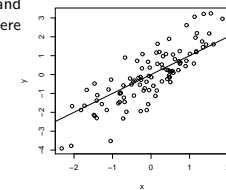
In the case of Parallel tests (see Chapter 2), $r = \frac{\text{cov}(x_1, x_2)}{\sqrt{\mathbb{V}(x_1) \times \mathbb{V}(x_2)}} = \mathbb{V}(t)/\mathbb{V}(x)$

Back to our Measurement model

Formula 3 parallels that for the Linear Model, $Y = X\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_e)$ (where y stands for the observed test score and x for the *true score*). The correlation between x_i and y_i where Y is RV distributed as $\mathcal{N}(0, \sigma_y)$ is:

$$r_{xy} = \text{cov}(x, y) / \hat{\sigma}_x \hat{\sigma}_y.$$

However, in the linear regression framework, we postulate that $\mathbb{E}(Y|X) = \beta_0 + \beta_1 X$ (X is fixed). What about r_{XY} when both X and Y are RV? This is a special case of LM called *Model II Linear regression* [31, chap. 14].



```
x <- rnorm(100)
y <- 1.2*x + rnorm(100, sd=.8)
cov(x,y)/(sd(x)*sd(y)) == cor(x,y)
```

References

- [1] Alan Agresti. *Categorical Data Analysis*. Wiley-Interscience, 1990.
- [2] D G Bonett and R M Price. Inferential methods for the tetrachoric correlation coefficient. *Journal of Educational and Behavioral Statistics*, 30:213–225, 2005.
- [3] M B Brown. Algorithm as 116: The tetrachoric correlation and its asymptotic standard error. *Journal of the Royal Statistical Society. Series C*, 26(3):343–351, 1977.
- [4] M B Brown and J K Benedetti. On the mean and variance of the tetrachoric correlation coefficient. *Psychometrika*, 42(3):347–355, 1977.
- [5] R A Charter and L S Feldt. Testing the equality of two alpha coefficients. *Perceptual and Motor Skills*, 82:763–768, 1996.
- [6] J Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

- [7] J Cohen. Weighted kappa: Nominal scale agreement with provision for scales disagreement of partial credit. *Psychological Bulletin*, 70:213–220, 1968.
- [8] L J Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334, 1951.
- [9] L J Cronbach and R J Shavelson. My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, 64(3):391–418, 2004.
- [10] J E Danes and O K Mann. Unidimensional measurement and structural equation models with latent variables. *Journal of Business Research*, 12:337–352, 1984.
- [11] Graham Dunn. *Statistics in Psychiatry*. Hodder Arnold, 2000.
- [12] J H Edwards and A W F Edwards. Approximating the tetrachoric correlation coefficient. *Biometrics*, 40:563, 1984.
- [13] J L Fleiss. Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement*, 5:105–112, 1981.
- [14] J L Fleiss. *Statistical Methods for Rates and Proportions*. New York: Wiley, Second edition, 1981.
- [15] D Garson. Correlation. Statnotes: Topics in Multivariate Analysis, 2008. Available

at: <http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm>. Accessed February 24, 2010.

- [16] L Guttman. A basis for analyzing test-retest reliability. *Psychometrika*, 10(4):255–282, 1945.
- [17] Ron D Hays and Dennis Revicki. Reliability and validity (including responsiveness). In Peter Fayers and Ron Hays, editors, *Assessing Quality of Life in Clinical Trials*, chapter 1.3, pages 25–39. Oxford University Press, Second edition, 2005.
- [18] G F Kuder and M W Richardson. The theory of the estimation of test reliability. *Psychometrika*, 2:151–160, 1937.
- [19] M A Long, K J Berry, and P W Mielke. Tetrachoric correlation: A permutation alternative. *Educational and psychological measurement*, 69(3):429–437, 2009.
- [20] Jum C Nunnally and Ira H Bernstein. *Psychometric Theory*. McGraw-Hill Series in Psychology, Third edition, 1994.
- [21] H Ogasawara. Accurate distribution and its asymptotic expansion for the tetrachoric correlation coefficient. *Journal of Multivariate Analysis*, 101(4):936–948, 2010.
- [22] L Wilkinson & APA Task Force on Statistical Inference. Statistical methods in psychology journals: Guidelines and explanations. *American*

Psychologist, 54:594–604, 1999. reprint available through the APA Home Page: <http://www.apa.org/journals/amp/amp548594.html>.

- [23] K Pearson. Mathematical contributions to the theory of evolution. vii. on the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London, Series A*, 195:1–47, 1900.
- [24] T Raykov. Scale reliability, cronbach's coefficient alpha, and violations of essential tau-equivalence for fixed congeneric components. *Multivariate Behavioral Research*, 32:329–354, 1997.
- [25] T Raykov. Reliability if deleted, not “alpha if deleted”: evaluation of scale reliability following component deletion. *British Journal of Mathematical and Statistical Psychology*, 60:201–216, 2007.
- [26] T Raykov. “alpha if item deleted”: A note on loss of criterion validity in scale development if maximizing coefficient alpha. *British Journal of Mathematical and Statistical Psychology*, 61:275–285, 2008.
- [27] W Revelle. Hierarchical cluster-analysis and the internal structure of tests. *Multivariate Behavioral Research*, 14(1):57–74, 1979.
- [28] P E Shrout and J L Fleiss. Intraclass correlation: Uses in assessing rater reliability.

Psychological Bulletin, 86:420–428, 1979.

- [29] Sidney Siegel and Jr N John Castellan. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Second edition, 1988.
- [30] K Sijtsma. Reliability beyond theory and into practice. *Psychometrika*, 74(1):169–173, 2009.
- [31] R R Sokal and F J Rohlf. *Biometry: The principles and practice of statistics in biological research*. W. H. Freeman, third edition, 1995.
- [32] W Sun, C-P Chou, A W Stacy, H Ma, J Unger, and P Gallaher. SAS and SPSS macros to calculate standardized cronbachs alpha using the upper bound of the phi coefficient for dichotomous items. *Behavior Research Methods*, 39(1):71–81, 2007.
- [33] Bruce Thompson, editor. *Score Reliability. Contemporary Thinking on Reliability issues*. Sage Publications, 2003.
- [34] J S Uebersax. The tetrachoric and polychoric correlation coefficients. Statistical Methods for Rater Agreement web site, 2006. Available at: <http://john-uebersax.com/stat/tetra.htm>. Accessed February 24, 2010.
- [35] D De Vaus. *Analyzing social science data*. London: Sage Publications, 2002.
- [36] R E Zinbarg, W Revelle, I Yovel, and W Li. Cronbach's α , Revelle's β , and McDonald's

ω_h : Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70(1):123–133, 2005.

- [37] B D Zumbo, A M Gadermann, and C Zeisser. Ordinal versions of coefficients alpha and theta for likert rating scales. *Journal of Modern Applied Statistical Methods*, 6:21–29, 2007.