Correcting Effect Sizes for Statistical Artifacts

Application in Meta-Analysis and Implementation in R and Python

Matthew B. Jané

2023-06-13

Table of contents

1 Proposal and Outline

Results across scientific studies vary drastically even when studying the same underlying phenomena. Sometimes these can be accounted for by some study-level characteristic (i.e., methodology, population, etc.) or it can be accounted for by variations in statistical artifacts such as measurement error or selection effects. Not only does the heterogeneity increase in the presence of statistical artifacts, but artifacts also induce systematic biases that can cause inaccurate results. Artifacts restrict our ability to draw meaningful inferences from scientific results, therefore it is important to apply corrections in order to obtain unbiased estimates.

Overview. The goal of the general exam is to turn it into an online open source e-textbook similar to this one on meta-analysis in R by Mathias Harrer and colleagues. The e-text will be split into two parts: artifact corrections and their application to meta-analysis. Each type of artifact will have a section of its own that details the history, methodology, and its implementation in R and Python. Another section will be dedicated to applying artifact corrections to meta-analysis, which is the type of research where the corrections are most used. The meta-analysis section will also contain the four parts (history, methodology, assumptions, and implementation in R). See this webpage where I cataloged equations and code that will go into this general exam. I also have started writing the unreliability section so you can start get an idea of how this will look.

History. The history part will provide an overview of the literature for that artifact correction. It will note where the correction first was introduced, the adjustments people have made since then, as well as studies where the correction has been applied (most likely in a meta-analysis). Also if applicable, provide examples of where these meta-analyses have utilized such corrections (e.g., Roth et al., 2015) Methodology. The methodological part will review the correction equations (for point estimates and standard errors) and how they were derived for each artifact correction. Artifact corrections will be applied to both correlation coefficients and standardized mean differences (repeated measures and independent samples). The methodological part for the meta-analysis section will discuss how heterogeneity, credibility/confidence intervals, and averages are calculated in the context of artifact corrections. It will also touch on competing approaches (Raju et al., 1983) to the traditional artifact correction approaches (Hunter & Schmidt, 2004).

Assumptions. Each artifact correction contains assumptions that must be met in order to obtain unbiased estimate of the true population effect size. This part will discuss each of these assumptions, when each matter, and what simulation studies have found in regard to violation

of assumptions. Assumptions at the meta-analytic level of analysis (e.g., independence between artifacts and moderators) will also be discussed.

Implementation in R and Python. Artifact corrections are only useful if you can apply them. Since many of the corrections are quite complex, it is important that these can be implemented easily in an open-source software such as R and Python. Each correction will be supplemented with code using R and Python with some additional packages like psychmeta (Dahlke and Wiernik 2019). For statistical analyses and especially for meta-analysis, R has tremendous support in terms of useful packages and a large community that makes it a highly flexible and powerful language. Python is a general-purpose programming language that is a more popular than R. For an example of how this will be done see this section on estimating reliability in R and Python.

2 Dedication

In Loving Memory of Haley Jané

My companion whose unwavering presence and unconditional love provided me with stability and solace in life's ever-changing journey

3 Introduction

3.1 Effect Sizes

3.1.1 Correlations

A correlation describes the relationship between two continuous variables. The correlation coefficient was first introduced 1904 by Charles Spearman

Technical Overview

If we draw a sample of n observations from a population, we can estimate the population correlation between variables x and y using a reasonably unbiased estimator, r,

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

This formulation is commonly referred to as the Pearson correlation coefficient (**pe?**) . To see under the hood of this seemingly complex mathematical formulation. Since the correlation is simply the the standardized covariance between two variables, x and y, we can first define the covariance as the product of the squared errors between $(x_i - \bar{x})$ and y $(y_i - \bar{y})$,

$$\sigma_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Then we can find the variance for x and y by taking the average squared error from the mean for x and y,

$$\sigma_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})}$$

$$\sigma_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})}$$

Now with each of these components, we can standardized the covariance by dividing by the standard deviations of x and y (i.e., square root of the variance). It can be now seen that the $\frac{1}{n-1}$ term cancels out in the numerator and denominator and thus will give us the original formula for the sample correlation coefficient.

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{3.1}$$

Since r is the observed sample correlation, it is important to note that, in the absence of artifacts, r provides an unbiased estimate of the true population correlation ρ ($r = \hat{\rho}$). Therefore, in conditions uncontaminated by artifacts, differences between the observed sample correlation and the true population correlation are attributable to sampling error (ε) such that,

$$\rho = \hat{\rho} + \varepsilon = r + \varepsilon \tag{3.2}$$

Where se_r is the standard error of the observed correlation. The standard error can be calculated from the sample size (n) and the observed correlation,

$$se_r = \sqrt{\frac{1 - r^2}{n - 2}} \tag{3.3}$$

3.1.2 Standardized Mean Differences

Standardized mean differences are used to quantify the average difference between groups along some variable. Standardizing the mean difference allows researchers to compare results between

Technical Overview

If we draw a sample of n_A subjects from group A and n_B subjects from group B, the mean difference between groups (d) on variable y can be defined as,

$$d = \frac{\bar{y}_A - \bar{y}_B}{\sigma_y^*}$$

Where the standardizer, σ_y^* is the pooled standard deviation between the two groups. The pooled standard deviation is calculated by taking the square root of the average variance between the two groups weighted by the degrees of freedom.

$$\sigma_y^* = \sqrt{\frac{(n_A - 1)\sigma_{y,A}^2 + (n_B - 1)\sigma_{y,B}^2}{n_A + n_B - 2}}$$

Where $\sigma_{y,A}$ and $\sigma_{y,B}$ are the standard deviations of y within groups A and B respectively. This estimator is commonly referred to as Cohen's d, however to avoid the use of jargon labels, d will be referred to as the standardized mean difference (SMD). The standard error of d is,

$$s_d = \sqrt{\frac{n_A + n_B}{n_A n_B} + \frac{d^2}{2(n_A + n_B)}}$$

The standardized mean difference assumes equal variance between groups, therefore in cases with unequal variance, the standardizer can simply be the standard deviation of just one of the groups. This is mostly used when the group comparison is between treatment and control groups since the control group standard deviation tends to be a better estimate of the baseline population standard deviation.

$$d_\Delta = \frac{\bar{y}_A - \bar{y}_B}{\sigma_{\text{control}}}$$

Since this equation only utilizes the standard deviation from just one group, the sampling error will be slightly larger.

$$s_{d_\Delta} = \sqrt{\frac{n_{\text{control}} + n_{\text{treatment}}}{n_{\text{control}} n_{\text{treatment}}} + \frac{d_\Delta^2}{2(n_{\text{control}} - 1)}}$$

3.1.3 Standardized Mean Change

Standardized mean change quantifies the average within-person change between time-points (e.g., pre-treatment vs post-treatment).

Technical Overview

Standardized mean

3.2 Bias induced by Statistical Artifacts

(Roth 2015)

(Van Aarde, Meiring, and Wiernik 2017)

(John E. Hunter and Hunter, n.d.)

Part I

Artifact Corrections

4 Small Samples

(Hedges 1989)

(Lin 2018)

(Hedges 1981)

(Fisher 1915)

(Olkin and Pratt 1958)

5 Unreliability

5.1 Introduction

In general terms, measurement is the process of quantifying an attribute or characteristic of something. In scientific measurement, the measurand is the quantity or the attribute we intend to measure. In the psychological sciences, measurands usually take the form of constructs such as intelligence or anxiety. The goal of measurement is to produce quantities (i.e., scores) that accurately reflect the measurand. It is important to note that measures are not all created equal, some perform better than others. Ideally, measures should produce scores that are consistent and repeatable, this is referred to as the *reliability* of a measure. A high quality measure should produce highly reliable scores. This section will review what reliability is in theory, how to estimate reliability, and how to adjust effect sizes for measurement error.

5.2 Reliability in True Score Theory

True score theory (or classical test theory) is a mathematical formalization of scores obtained from measurements. The true score model assumes that each person (or animal), p has a true score, t_p, that stays constant over measurements. Observed scores, x_{pf}, can vary between forms (f) of the measure, m. This variation is due to measurement-specific error, e_{pf}.

$$x_{pf} = t_p + e_{pf}$$

The true score can be defined as the expected value (i.e., the mean) of observed scores over an infinite number of repeated measurements such that, $\mathbb{E}_{f\to\infty}[x_f] = t$. It is also assumed that the expectation of measurement-specific error is zero, $\mathbb{E}_{f\to\infty}[e_f] = 0$. It follows from these assumptions that the covariance between errors and true scores is zero ($\sigma_{et} = 0$) and the covariance between error scores in parallel measurements is zero ($\sigma_{ee'} = 0$). The independence between true scores and errors provide convenient parsing of the variance in observed scores (σ_x^2) into components of variance in true scores (σ_t^2) and errors (σ_e^2),

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2 \tag{5.1}$$

If $\sigma_e^2 > 0$ then the measurement has imperfect reliability, that is, observed scores are not identical to true scores. In practice, this is almost always the case. Reliability can be defined

as the square correlation between observed scores and true scores, r^2_{xt}, or the correlation between observed scores in parallel measurements, $r_{xx'} = r^2_{xt}$.

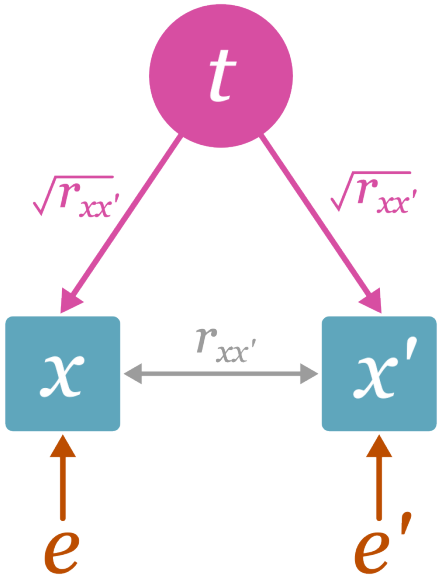


Figure 5.1: Structural model illustrating the relationship between true scores, observed scores, and error scores. The pink circle labeled t indicates the true scores, the blue squares labeled with x and x' represent observed scores on parallel measurements, and the red e denotes error. Correlations between t, x, and x' are in terms of reliability ($r_{xx'}$). Note that $\sqrt{r_{xx'}} = r_{xt}$.

Given that errors do not co-vary between parallel measurements and true scores are held constant over measurements, it becomes apparent that the covariance between observed scores produced from parallel measurements must solely be attributable to true score variance, $\sigma_{xx'} = \sigma^2_t$. The covariance in observed scores can be standardized to obtain the correlation coefficient between parallel measurements (i.e., the reliability), such that,

$$r_{xx'} = \frac{\sigma_{xx'}}{\sigma_x \sigma_{x'}} = \frac{\sigma^2_t}{\sigma^2_x}$$

Therefore reliability can be expressed in a few forms different forms

$$r_{xx'} = r^2_{xt} = \frac{\sigma^2_t}{\sigma^2_t + \sigma^2_e} = \frac{\sigma^2_t}{\sigma^2_x} \tag{5.2}$$

In the literature, the correlation between observed and true scores, r_{xt}, is often referred to as the "measure quality index" (John E. Hunter and Schmidt 1990), however measure quality

encompasses both reliability *and* validity. Validity A measure can demonstrate high reliability even though the scores produced by the measure do not accurately reflect the measurand (the quantity that we are intending to measure). For example, if an individual were to step on a weight scale with shoes on, the weight presented on the scale would be highly reliable, namely, if the individual were to repeat this process, they would achieve highly similar results. Nevertheless, the observed weight is systematically biased upward by the weight of the shoes. Therefore if a measure is reliable it does not logically follow that the measure is necessarily valid.

5.3 Estimating Reliability

In practice, reliability must be estimated through indirect methods, since true scores and errors are unknown. Their are many estimators that can be used however, we will go over three of the most common approaches: coefficient alpha, split-half, and test-retest reliability.

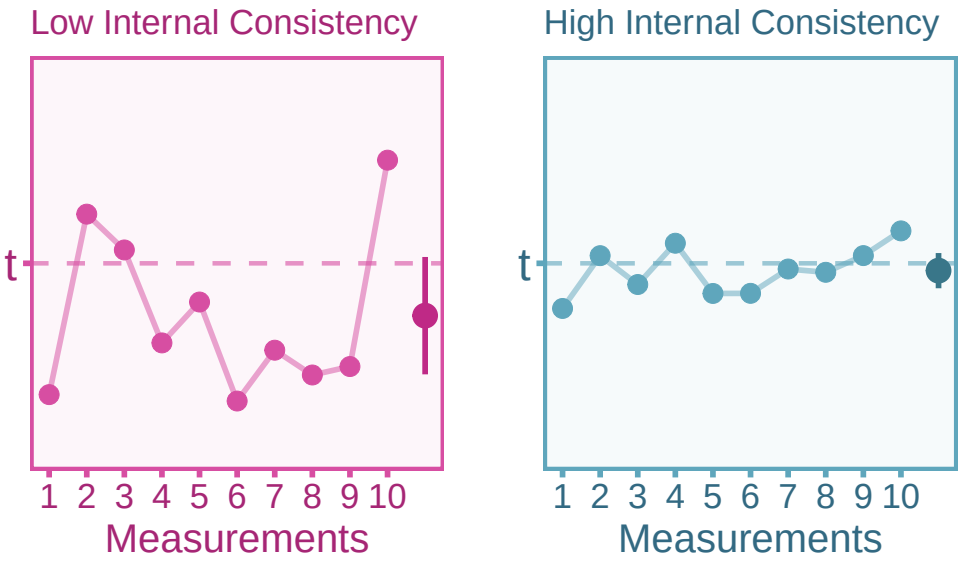
Internal Consistency Estimators

Maybe the most conventionally reported reliability estimator in the psychological sciences is coefficient alpha, also referred to as Cronbach's alpha or internal consistency. Alpha has the benefit of being computationally convenient, but it also brings along many assumptions that are often violated in practice (Haertel 2006; Sijtsma 2009). Cronbach's alpha, along with other internal consistency estimators, serves the purpose of assessing the reliability of composite measures comprising multiple components. Taking multiple measurements and then averaging tends to provide a better estimate of true values. For instance, let's consider the case of Francis Galton (Galton 1907), who conducted a study involving 787 individuals estimating the weight of an ox. On average, each person's estimate deviated by approximately 37 pounds from the actual weight of the ox, which was recorded as 1198 pounds. However, when all the guesses were averaged together, the combined estimate was 1207 pounds, just a 9 pound difference from the actual value. Averaging a number of noisy estimates provides a much more stable and reliable estimate. So to create a more stable composite score (X), we can take the score (x_m) from k measurements and average them such that,

$$X = \frac{1}{k}(x_1 + x_2 + ... + x_k) = \frac{1}{k}\sum_{m=1}^{k} x_m$$

Coefficient alpha is represents the reliability of this composite scores. Coefficient alpha only requires three parameters to calculate, the number of measurements (k), the variances of each items ($\sigma^2_{x_m}$), and the variance of the composite score (σ^2_X),

$$_{\alpha}r_{XX'} = \frac{k}{k-1}\left(1 - \frac{\sum_{m=1}^{k}\sigma^2_{x_m}}{\sigma^2_X}\right)$$



With tighter assumptions (i.e., strictly parallel forms, Haertel 2006), the formula for coefficient alpha can be simplified to just two parameters: the number of measurements and the average correlation between measured scores ($\bar{r}_{x_i x_j}$, where $i \neq j$). This formula is known as Spearman-Brown's prophecy,

$$_{\text{sb}}r_{XX'} = \frac{k\bar{r}_{x_i x_j}}{1 + (k-1)\bar{r}_{x_i x_j}}$$

This can be simplified further if we we have two observed scores. This formulation is traditionally called split-half reliability:

$$_{\text{sh}}r_{XX',} = \frac{2r_{x_1 x_2}}{1 + r_{x_1 x_2}}$$

All of these reliability estimators measure internal consistency, therefore they do not account for error outside of the measurement-specific error. There are other sources of error that internal consistency reliability estimates do not account for, such as transient error or rater-specific error.

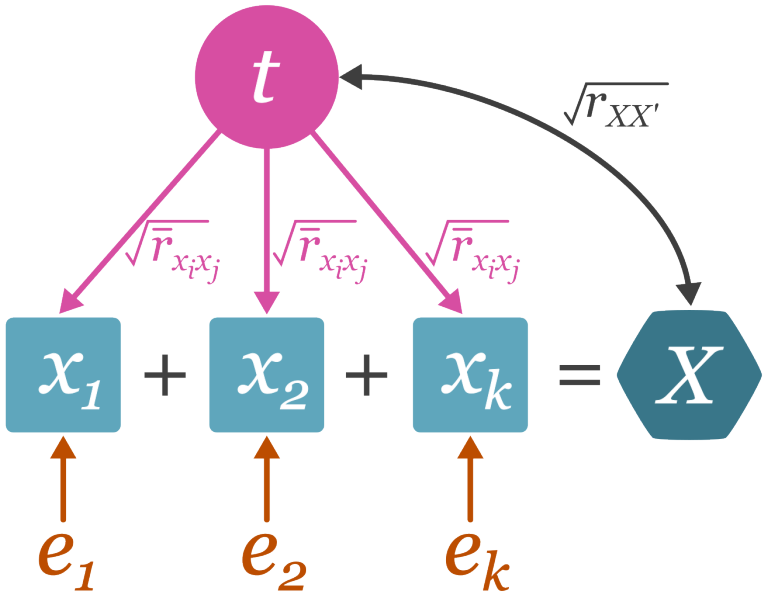


Figure 5.2: Structural model illustrating internal consistency. The pink circle labeled t indicates the true scores, the blue squares, $x_{1\ldots k}$, represent the observed scores across multiple measurements, and the red e denotes error. The dark blue hexagon, X, indicates a composite score as a sum of the observed scores ($x_{1\ldots k}$). Note that $\sqrt{r_{XX'}} = r_{Xt}$.

Calculating Internal Consistency in R and Python

5.3.0.1 R

Let us simulate a data set of 50 individuals where each observed score has the same true score with some error.

```
#set seed
set.seed(343)

# set sample size
n = 50

# simulate data
t = rnorm(n, 0, 1) # simulate true scores
x1 = t + rnorm(n, 0, 1) # simulate observed scores for measurement 1
x2 = t + rnorm(n, 0, 1) # simulate observed scores for measurement 2
x3 = t + rnorm(n, 0, 1) # simulate observed scores for measurement 3
x4 = t + rnorm(n, 0, 1) # simulate observed scores for measurement 4

# calculate composite score
X = x1 + x2 + x3 + x4
```

Calculate Coefficient Alpha Reliability:

```
# step 1. calculate variance of observed (measured) scores
var_xi = c(var(x1),var(x2),var(x3),var(x4))

# step 2. calculate variance of composite score
var_X = var(X)

# step 3. get number of items (k)
k = length(var_xi)

# step 4. calculate coefficient alpha reliability
rXX_alpha = k / (k-1) * (1 - sum(var_xi)/var_X)

# display reliability
print(round(rXX_alpha,3))
```

```
[1] 0.775
```

Calculate Reliability via Spearman-Brown's Prophecy:

```
# step 1. get correlation matrix between all observed scores
corr_mat = cor(cbind(x1,x2,x3,x4))

# step 2. average off-diagonal elements of matrix
diag(corr_mat) <- NA
rxixj = mean(corr_mat, na.rm = TRUE)

# step 3. get number of items (k)
k = dim(corr_mat)[1]

# step 4. calculate Spearman-Brown reliability
rXX_SB = k * rxixj / (1 + (k-1) * rxixj)

# display reliability
print(round(rXX_SB,3))
```

```
[1] 0.775
```

Calculate the Split-Half Reliability:

```
# step 1. make composite scores for each half of the observed scores
X1 = x1 + x2
X2 = x3 + x4

# step 2. calculate the correlation between the scores of both halves
rX1X2 = cor(X1,X2)

# step 3. calculate the split-half reliability
rXX_SH = 2*rX1X2 / (1 + rX1X2)

# display reliability
print(round(rXX_SH,3))
```

```
[1] 0.824
```

True Reliability: Lets see how the results compare to the squared correlation of the true scores and our composite score (true reliability).

```r
# calculate true reliability
rXt = cor(X,t)

# display true reliability
print(round(rXt^2,3))
```

```
[1] 0.753
```

In this case, the reliability estimates do a fairly good job of estimating the true reliability of the observed scores.

5.3.0.2 Python

Simulate Data: Let us simulate a data set of 50 individuals where each observed score has the same true score with some error. To calculate the necessary statistics, we will import the `numpy` package.

```python
#import numpy
import numpy as np

# set seed
np.random.seed(343)

# set sample size
n = 50

# simulate data
t = np.random.normal(0, 1, n) # simulate true scores
x1 = t + np.random.normal(0, 1, n) # simulate observed scores for measurement 1
x2 = t + np.random.normal(0, 1, n) # simulate observed scores for measurement 2
x3 = t + np.random.normal(0, 1, n) # simulate observed scores for measurement 3
x4 = t + np.random.normal(0, 1, n) # simulate observed scores for measurement 4

# calculate sum score
X = x1 + x2 + x3 + x4
```

Calculate Coefficient Alpha Reliability:

```python
# step 1. calculate variance of observed (measured) scores
var_xm = [np.var(x1),np.var(x2),np.var(x3),np.var(x4)]
```

```
# step 2. calculate variance of composite score
var_X = np.var(X)

# step 3. get number of items (k)
k = len(var_xm)

# step 4. calculate coefficient alpha reliability
rXX_alpha = k / (k-1) * (1 - sum(var_xm)/var_X)

print(round(rXX_alpha,3))
```

```
0.769
```

Calculate Reliability from Spearman-Brown's prophecy formula:

```
# step 1. get correlation matrix between all observed scores
corr_mat = np.corrcoef([x1,x2,x3,x4])

# step 2. average off-diagonal elements of matrix
rxixj = np.mean(corr_mat[~np.eye(k,dtype=bool)])

# step 3. get number of items (k)
k = len(corr_mat)

# step 4. calculate Spearman-Brown reliability
rXX_SB = k * rxixj / (1 + (k-1) * rxixj)

print(round(rXX_SB,3))
```

```
0.772
```

Calculate Split-Half Reliability:

```
# step 1. make composite scores for each half of the observed scores
X1 = x1 + x2
X2 = x3 + x4

# step 2. calculate the correlation between the scores of both halves
rX1X2 = np.corrcoef(X1,X2)[0,1]
```

```
# step 3. calculate the split-half reliability
rXX_SH = 2*rX1X2 / (1 + rX1X2)

# display reliability
print(round(rXX_SH,3))
```

```
0.772
```

Lets see how the results compare to the squared correlation of the true scores and our composite score (true reliability).

```
# calculate true reliability
rXt = np.corrcoef(X,t)[0,1]

# display reliability
print(round(rXt**2,3))
```

```
0.791
```

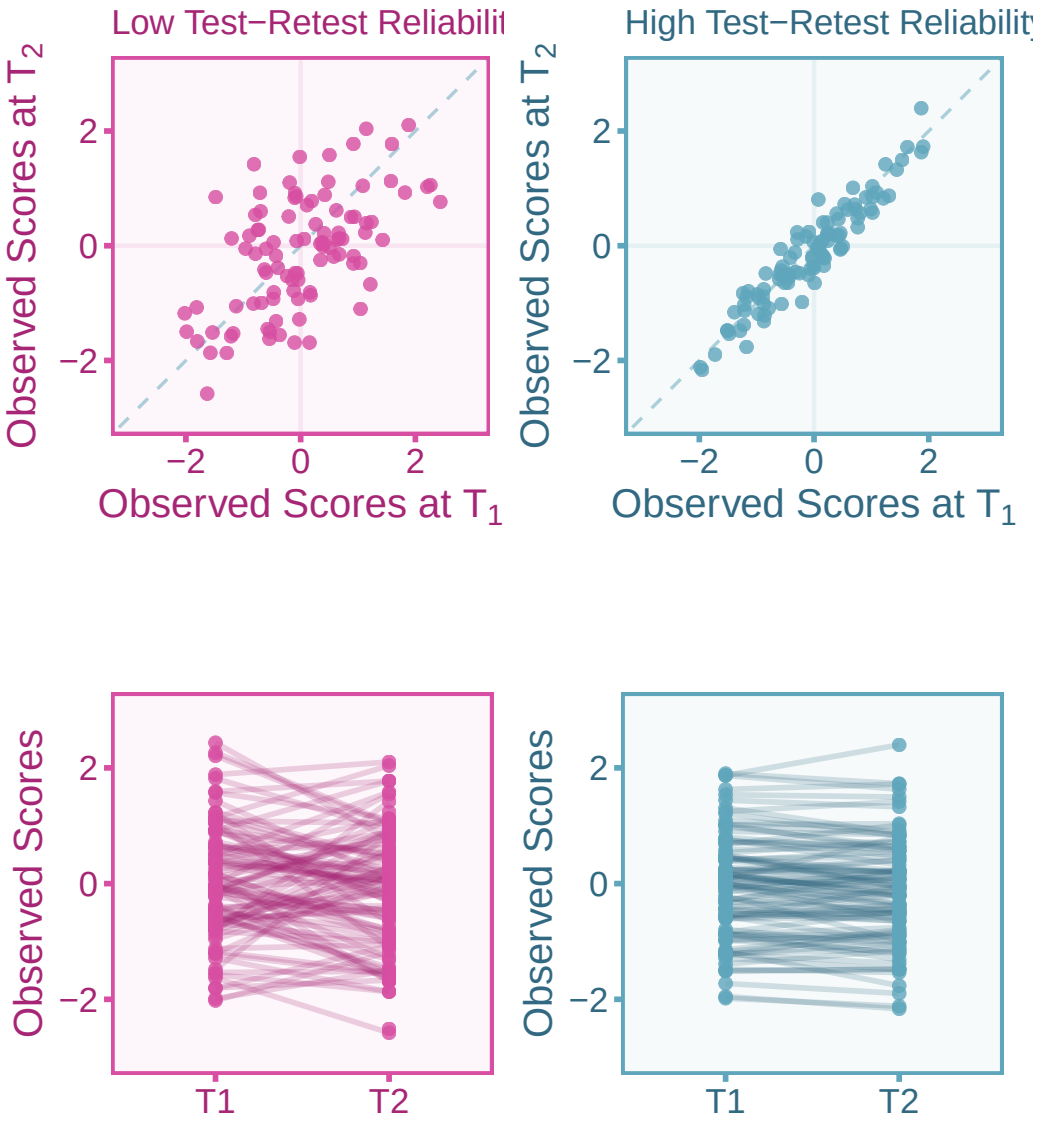
In this case, the reliability estimates do a fairly good job of estimating the true reliability of the observed scores. There are also functions within the `psych` package that allow you to easily calculate coefficient alpha among other reliability estimators

5.3.1 Test-Retest Stability Estimator

There measurement errors that exist outside of the measurement instrument itself. Transient errors represent fluctuations in observed scores over time. These fluctuations, even if they are systematic (e.g., fatigue over the course of a single day), add extraneous within-person variance that can mask true scores (i.e., expectation of observed scores). For example, if a researcher wants to investigate how individuals differ in processing speed, then variation within an individual's scores across multiple testing sessions would be considered error since the goal of the study is to investigate between-person variation. Considering transient fluctuations as error depends on the research goal, so it is important for researchers to take care in considering which variance components should be considered error in their study. To estimate test-retest reliability, we can compute the pearson correlation coefficient between the measurement at time 1 (x_{T_1}) and the second measurement at time 2 (x_{T_2}).

$$_{\text{tr}}r_{xx'} = r_{x_{T_1}x_{T_2}}$$

Note that calculating the pearson correlation coefficient between time-points ignores systematic changes (e.g., practice effects).



Calculating Test-Retest Reliability in R and Python

5.3.1.1 R

Lets calculate test-retest reliability in R. First, we can simulate observed scores at two time points, T_1 and T_2. We can assume that the true scores remain constant between T_1 and T_2. Second, we can calculate the correlation between the observed scores at each time point ($r_{x_{T_1}x_{T_2}}$).

```
# set seed
set.seed(343)

# set sample size
n = 70

# simulate true scores
t = rnorm(n,0,1)

# simulate scores at time 1
xT1 = t + rnorm(n,0,.5)

# simulate scores at time 2
xT2 = t + rnorm(n,0,.5)

# calculate test-retest reliability
rxx = cor(xT1,xT2)

# display reliability
print(round(rxx,3))
```

```
[1] 0.803
```

```
# compare with true reliability
rxx_true = cor(xT1,t)^2

# display true reliability
print(round(rxx_true,3))
```

```
[1] 0.768
```

5.3.1.2 Python

Lets calculate test-retest reliability in R. First, we can simulate observed scores at two time points, T_1 and T_2. We can assume that the true scores remain constant between T_1 and T_2. Second, we can calculate the correlation between the observed scores at each time point ($r_{x_{T_1}x_{T_2}}$).

```
# import numpy
import numpy as np

# set seed
np.random.seed(343)

# set sample size
n = 70

# simulate 70 true scores
t = np.random.normal(0,1,n)

# simulate scores at time 1
xT1 = t + np.random.normal(0,.5,n)

# simulate scores at time 2
xT2 = t + np.random.normal(0,.5,n)

# calculate test-retest reliability
rxx = np.corrcoef(xT1,xT2)[0,1]

# display reliability
print(round(rxx,3))
```

```
0.768
```

```
# compare with true reliability
rxx_true = np.corrcoef(xT1,t)[0,1]**2

# display true reliability
print(round(rxx_true,3))
```

```
0.731
```

5.3.2 Sources of Measurement Error

There are many estimators of reliability beyond internal consistency and test-retest that account for different sources of error and hold different assumptions. There are many sources of measurement error that different estimators of reliability account for adapted from table 1 of Wiernik and Dahlke (2020) :

- Random Response Error: Genuine randomness in responses. Examples include: motor errors and variation in response time.
- Time/Environment-Specific (Transient) Error: Fluctuations in scores as a result of the specific time or environment of the measurement. For instance, if researchers administered an ability test to a sample of undergraduate students throughout the course of a day, the student's who complete the test at the end of the day will likely perform worse than participant's who completed due to fatigue rather than ability. Errors due to illness, mood, hunger, environmental distractors, etc. all fall under the umbrella of transient errors.
- Instrument-Specific Error: Error due to the specific content or make-up of the measurement instrument. For example, a psychological scale using likert items participant's idiosyncratic interpretations of questions and response options rather than their standing on the latent construct.
- Rater/Observer-Specific Error: Errors induced by idiosyncratic biases of individual raters and rater by ratee interactions (e.g., Teacher A gives higher grades to students who stay after class).

Different estimators of reliability account for different sources of measurement error therefore depending on the research design, it is important to carefully choose which reliability is most relevant for your use case. Note that even if two estimators account for the same types of measurement error, they likely hold different assumptions that may be violated in a given research context.

Table 5.1: Table 1. List of reliability coefficients and the sources of error they account for.

Estimator	Description	Random Response Error	Transient Error	Instrument-Specific Error	Rater-Specific Error
Coefficient Alpha	Internal consistency coefficient for composite measures.				

Estimator	Description	Random Response Error	Transient Error	Instrument-Specific Error	Rater-Specific Error
Coefficient Omega	Internal consistency coefficient for composite measures with specified factor structure.				
Split-Half	Internal consistency coefficient for measurements that are split into two halves.				
Kuder-Richardson 20	Internal consistency when observed scores are binary (special case of coefficient alpha).				
Item Response Theory Reliability	Reliability coefficient derived from item response theory (as opposed to classical test theory)				
Inter-Rater/Inter-Observer Reliability	Consistency in scoring between raters/observers.				

Estimator	Description	Random Response Error	Transient Error	Instrument-Specific Error	Rater-Specific Error
Test-Retest	Stability coefficient for repeated measurements across time				
Delayed Coefficient Alpha	Average of all possible split-half reliabilities				
G-Coefficient	Reliability coefficient derived from generalizability theory (G-theory). Can incorporate any source of error if enough data is present.				

5.4 Bias in Correlation Coefficients

Unreliability induces systematic bias in effect size estimates such as correlation coefficients Spearman (1904). Lets say we have two observed scores x and y,

$$x = x_t + e$$

$$y = y_t + e$$

In most research contexts, we would like to estimate the correlation between true scores where the correlation between true scores, x_t and y_t is

$$\rho = \frac{\sigma_{x_t y_t}}{\sigma_{x_t} \sigma_{y_t}}$$