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 Preface

Preface page still in progress

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 Effect Sizes and Notation

3.1 What are Effect Sizes?

Effect sizes are statistics that measure the magnitude of a relationship between two variables. It's important to remember that effect sizes are a valuable tool, enabling researchers to extract meaningful insights, rather than being the ultimate objective themselves. Effect sizes aide in researcher's ability to draw meaningful inferences from data and therefore it is important that they are accurate. Biased effect sizes can be likened to a foggy windshield. Just as condensation on glass obstructs a clear view of the road, biased effect sizes can obscure the true association between variables. Similar to how one must clean the windshield to drive safely, researchers must correct for biases in effect sizes to attain a clear and accurate perspective on their data. Correlation coefficients and standardized mean differences are two of the most common effect sizes and so they will be the primary focus of this book. To see how an effect size may look in practice, the example below will illustrate how calculating one may look in a clinical setting.

3.1.1 Applied Example

Lets say we want to test whether a new drug can alleviate anxiety, therefore we decide to conduct an experiment to see how well this drug performs. We first randomly assign each participant in the study to either a treatment group (T) or a control group (C). In our experiment we want test how well the experimental drug reduces anxiety, therefore we measure the subjects' self reported anxiety after under-going the treatment. To see if the drug actually worked in alleviating anxiety, we want to compare the scores from the treatment group and the control group. To do this we can estimate the average treatment effect (ATE), which is the difference in the mean value of self-reported anxiety scores between the treatment group and the control group such that, $ATE = \text{Mean}(X_T) - \text{Mean}(X_C)$. However, anxiety scores have no meaningful units, so if we obtain an ATE value of -3 there is no way to tell if this value is large or small, since it is entirely dependent on how the anxiety scores are scaled. Standardization can allow us to draw meaningful inferences about the size of the effect that can be comparable across scales. We can standardize the ATE by dividing by the standard deviation of scores in the control group, (SD): Effect Size $= \frac{ATE}{\text{SD}_C}$. The effect size is now on an interpretable scale (standard deviations). If we achieve an standardized effect size value of -0.50, we can interpret this as the treatment group exhibiting a reduction in anxiety equivalent to half a standard deviation compared to the control group.

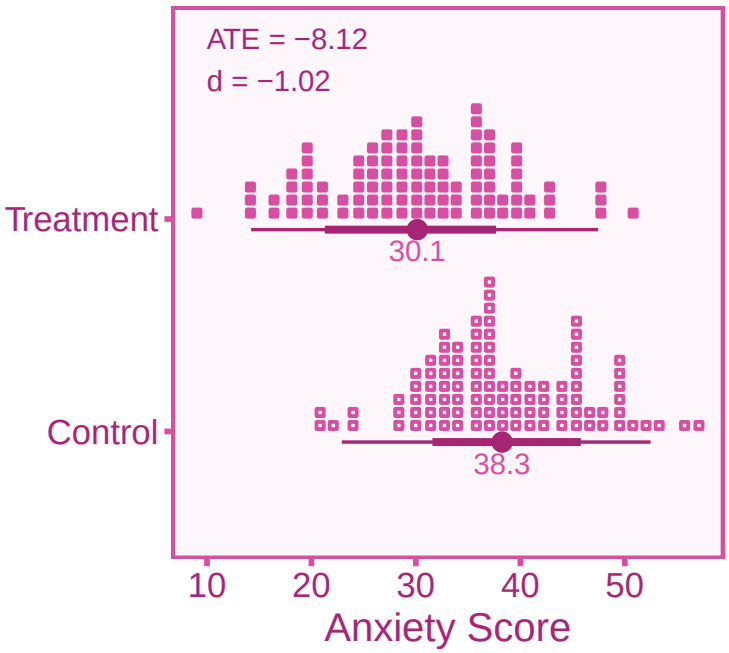


Figure 3.1: Simulated experimental data

3.1.2 Defining effect sizes

Lets say we have an effect size of interest that quantifies the relationship between an independent and dependent variable. The population effect size can be denoted as ϑ, however this population value is unknown. We can obtain an estimate the population effect size by conducting a study on a sample drawn from the population and then calculating a study effect size, θ. The study effect size is a function of the population effect size and sampling error (ε) such that,

$$\theta = \vartheta + \varepsilon \tag{3.1}$$

Effect sizes will differ from study to study, this can be due to two reasons: variance in population effect sizes (σ^2_ϑ) or variance in sampling error (σ_ε). Accordingly, we can express the variance in study effect sizes (σ_θ) as,

$$\sigma^2_\theta = \sigma^2_\vartheta + \sigma^2_\varepsilon$$

If studies were drawing samples from the same population, the variance in the population effect size would be zero ($\sigma^2_\vartheta = 0$) and the expected value (i.e., the mean) of study effect sizes would be equal to the population effect size, $\mathbb{E}[\theta] = \vartheta$.

3.2 Effect Sizes and Artifacts

In practice, *observed* effect size estimates are often biased relative to the *true* effect size of interest, that is, the observed population effect size (ϑ_o) is a product of the true population effect size (ϑ) and artifactual bias (a):

$$\vartheta_o = a\vartheta \tag{3.2}$$

Note that if $a = 1$ this would indicate that there is no artifactual bias ($\vartheta_o = \vartheta$), if $a > 1$ then it would indicate effect size inflation (i.e., biased away from zero), and if $a < 1$ that would indicate effect size attenuation (i.e., biased toward zero). It can be seen in Equation **??** that we can re-arrange the formula to obtain the true population effect size by dividing the observed population effect size by a,

$$\vartheta = \frac{\vartheta_o}{a}.$$

For a single study that computes an effect size from a sample drawn from the population, the observed study effect size (θ_o) would be expressed by

$$\theta_o = \vartheta_o + \varepsilon_o$$

Using Equation **??** we can express the observed effect size in terms of the true population effect size rather than the observed population effect size,

$$\theta_o = a\vartheta + \varepsilon_o$$

Then we can correct the observed effect size by dividing by the biasing factor, a, to obtain an unbiased estimate of the true effect size:

$$\theta_c = \frac{\theta_o}{a}$$

The sampling error and it's variance must also be corrected,

$$\varepsilon_c = \frac{\varepsilon_o}{a}$$
$$\sigma^2_{\varepsilon_c} = \frac{\sigma^2_{\varepsilon_o}}{a^2}.$$

The corrected effect size should be an unbiased estimate of the true population effect size as long as the systematic bias multiplier is accurately measured (which is not a trivial task). It is important to note that the corrected effect size will not yield additional statistical power, that is, test-statistics and p-values will remain unchanged. We can demonstrate this mathematically

that the z-statistic of the observed effect size (z_{θ_o}) is identical to the z-statistic of the corrected effecct size (z_{θ_c}),

$$z_{\theta_o} = \frac{\theta_o}{\sigma_{\varepsilon_o}} = \frac{\frac{\theta_o}{a}}{\frac{\sigma_{\theta_o}}{a}} = \frac{\theta_c}{\sigma_{\varepsilon_c}} = z_{\theta_c}$$

3.3 Defining an Effect Size Estimand

An effect size *estimand* is the theoretical quantity that we are trying to estimate. Before delving into the application of correction factors, it is important to clearly define the effect size estimand you aim to capture, including the summary statistic, relevant variables, and the target population. This preliminary step might appear trivial, but it is crucial, as it determines the accuracy and relevance of any subsequent artifact corrections. For instance, consider a scenario where we conduct a study involving a sample of college students with the aim of generalizing our findings to the broader general population. In this context, it is important to correct for range restriction, given the evident selection effects that exist in the college student populations. However, if our sole objective is to draw conclusions pertaining exclusively to the college student demographic, correcting for range restriction would be inappropriate. Furthermore, let's examine the variable of interest, such as grade-point average (GPA), within this population. Do we intend to focus solely on the raw GPA score, or is our goal to capture what GPA represents, namely, academic achievement? If our aim is to investigate the raw GPA score, then correcting for measurement error would be inappropriate. However, if our primary focus lies in assessing the student's academic achievement, then it may be relevant to correct GPA scores for measurement error. Defining our estimand guides our approach to artifact correction and ensures that these correction procedures align with the underlying research goals.

3.4 Effect Size Notation

Because of the nature of the topic, this book will cover a large amount of equations and computer code. Therefore to make it as straight-forward as possible the notation will follow a systematic framework to distinguish between types of effect sizes. This book will only be covering two main types of effect sizes: correlations (r) and standardized mean differences (d). Throughout the book variations of r and d will show up frequently, these variations will be differentiated with subscripts that are consistent with that section. Also, to distinguish between population-level values (i.e., the effect size across all potential observations) and effect sizes specific to a study or sample (i.e., the effect size observed within a single sample drawn from the population), we will use the following notation:

- Arbitrary Effect Size

  - Population value: ϑ
  - Study/sample value: θ
- Correlations
  - Population value: ρ
  - Study/sample value: r
- Standardized Mean Differences
  - Population value: δ
  - Study/sample value: d

In the most cases, continuous independent variables will be denoted with x and dependent variables with y (note that this notation may differ when referring to observed and true scores). Categorical (i.e., groupings) variables will be denoted with g (these will be used for standardized mean differences).

3.5 Correlations

A correlation describes the relationship between two continuous variables. The Pearson correlation coefficient was first introduced by Auguste Bravais (1844). Later developed by Karl Pearson, lending itself to the name.

3.5.1 Technical Overview Correlations (*r*)

If we draw a sample of n observations from a population, we can calculate the study correlation (r) between variables x and y using the following Pearson's product-moment estimator,

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}. \tag{3.3}$$

For digestibility, we can break down the formula into parts. The correlation coefficient can be defined as the covariance between x and y standardized by the product of their variances,

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3.4}$$

we can first define the covariance (σ_{xy}) as the average product of errors for x and y,

$$\sigma_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \tag{3.5}$$

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

$$\sigma_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.6}$$

$$\sigma_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{3.7}$$

Plugging in Equation **??**, Equation **??**, and Equation **??** into Equation **??** we can see that the term, $\frac{1}{n-1}$, will cancel out and we will be left with the original pearson correlation coefficient formula from Equation **??**,

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}. \tag{3.8}$$

In the absence of artifacts, the Pearson correlation r is an asymptotically (i.e., as n approaches infinity) unbiased estimator (in small sample sizes, it is biased see section on small samples). We can express r similarly to Equation **??**,

$$r = \rho + \varepsilon, \;\; \sigma_\varepsilon^2 = \mathrm{Var}(\varepsilon) \tag{3.9}$$

Where σ_ε^2 is the sampling variance of the observed correlation. The sampling variance can be calculated from the sample size (n) and the population correlation,

$$\sigma_\varepsilon^2 = \frac{1-\rho^2}{n-2} \tag{3.10}$$

In practice, the population correlation is unknown so the study correlation can be used instead (r) in the above formula. Note that the sampling variance is the square of the standard error. If the observed correlation is biased relative to the true correlation, we can see that the observed population correlation is systematically biased by an artifact factor, a,

$$\rho_o = a\rho$$

The observed study correlation would then be defined as,

$$r_o = \rho_o + \varepsilon_o = a\rho + \varepsilon_o$$

The corrected correlation coefficient (r_c) and it's sampling variance (σ_{ε_c}) can both be defined as:

$$r_c = \frac{r_o}{a}$$

$$\sigma^2_{\varepsilon_c} = \frac{\sigma^2_{\varepsilon_o}}{a^2}$$

3.6 Standardized Mean Differences

Standardized mean differences are used to quantify the average difference in some variable between groups. The most commonly used formulation is Cohen's d (Cohen 1988) which quantifies the average difference between groups (e.g., men vs. women) and standardizes by the pooled standard deviation. Note that the other most commonly used estimator is Hedges' g, but the difference between the two is a small sample correction factor that can be found in the chapter on small samples.

3.6.1 Technical Overview of Standardized Mean Difference (*d*)

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,

$$d = \frac{\bar{y}_A - \bar{y}_B}{\sigma_p}$$

Where the standardizer, σ_p 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.

$$\sigma_p = \sqrt{\frac{(n_A - 1)\sigma^2_A + (n_B - 1)\sigma^2_B}{n_A + n_B - 2}}$$

Where σ_A and σ_B are the standard deviations of y within groups A and B respectively. This SMD estimator is commonly referred to as Cohen's d. We can define the study/sample d value as a function of the population d value (δ):

$$d = \delta + \varepsilon$$

Similar to the previous section on correlation coefficients, the observed d value is a function of the true population value and artifactual bias (a),

$$\delta_o = a\delta.$$

Therefore the observed study/sample d value can be defined as a function of the observed population value *or* the true population value plus artifactual bias:

$$d = \delta_o + \varepsilon_o = a\delta + \varepsilon_o.$$

Thus the corrected standardized mean difference (d_c) and it's sampling variance ($\sigma^2_{\varepsilon_c}$) can both be defined as:

$$d_c = \frac{d_o}{a}$$

$$\sigma^2_{\varepsilon_c} = \frac{\sigma^2_{\varepsilon_o}}{a^2}$$

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 correct effect sizes for measurement error.

> Too long didn't read?
>
> Correction for correlations (r)
>
> $$r_c = \frac{r_o}{\sqrt{r_{xx'} r_{yy'}}}, \quad \sigma_{\varepsilon_c} = s_r \left(\frac{r_c}{r_o} \right)$$
>
> Correction for SMD (d)
>
> $$d_c = \frac{d_o}{\sqrt{r_{yy'_P}}}, \quad \sigma_{\varepsilon_c} = s_d \left(\frac{d_c}{d_o} \right)$$

5.2 Reliability in True Score Theory

True score theory (or classical test theory) is a mathematical formalization of observed scores obtained from measurements. Observed scores, x_{if}, is defined as a score obtained from individual i upon measurement m. The true score model assumes that each individual, has a true score, T_i, that stays constant over repeated measurements. Variation in observed scores over repeated measurements is due to measurement-specific error, e_{im},

$$x_{im} = T_i + e_{im}.$$

Here, measurements are *strictly parallel*. Strictly parallel measurements have the following four properties (p. 69, Haertel 2006):

1. Measurements have identical specifications. That is, each measurement is obtained with an identical format and procedure.
2. The distribution of observed scores for each measurement are identical: $f(x_1) = f(x_2) = ...$
3. Any set of two measurements are assumed to covary the same as any other set of two measurements: $\sigma_{x_1x_2} = \sigma_{x_2x_3} = \sigma_{x_1x_3} = ...$
4. Each measurement equally covaries with any other variable: $\sigma_{x_1y} = \sigma_{x_2y} = ...$

True scores can be defined as the expected value (i.e., the mean) of observed scores over repeated measurements such that, $\mathbb{E}_m[x_{im}] = T_i$. Given this assumption, it can be inferred that the average of the resultant errors is zero across repeated measurements, $\mathbb{E}_m[e_{im}] = 0$ and therefore the covariance between errors on repeated measurements is zero and the covariance between errors in parallel measurements is zero ($\sigma_{ee'} = 0$). It follows that the covariance between errors and true scores is also zero ($\sigma_{eT} = 0$). 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 measurement errors (σ_e^2),

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

Ultimately we desire to have observed scores that closely resemble true scores, therefore it is important to minimize measurement error variance (σ_e^2). If $\sigma_e^2 = 0$ then the scores can be said to have perfect reliability, that is, observed scores do not vary upon repeated measurements and thus are identical to true scores. In practice, this is virtually never the case. Since we know that the covariance between errors in parallel measurements is zero, it should be apparent that the covariance between observed scores in parallel measurements must solely be attributable to variance in true scores, $\sigma_{xx'} = \sigma_{TT'} + \sigma_{ee'} = \sigma_{TT'} = \sigma_T^2$. In true score theory, reliability can be defined as the proportion of true variance in the total observed variance ($\frac{\sigma_T^2}{\sigma_x^2}$) or the correlation between observed scores in parallel measurements ($r_{xx'}$).

$$r_{xx'} = \frac{\sigma_{xx'}}{\sigma_x \sigma_{x'}} = \frac{\sigma_T^2}{\sigma_x^2}$$

The reliability is also equivalent to the square of the correlation between observed scores and true scores. To understand why this is the case, note that the covariance between parallel forms of a measure is equivalent to the covariance between observed scores and true scores, $\sigma_{xT} = \sigma_{(T+e)T} = \sigma_T^2 + \sigma_{Te} = \sigma_T^2 = \sigma_{xx'}$.

$$r_{xx'} = \frac{\sigma_T^2}{\sigma_x^2} = \frac{(\sigma_T^2)^2}{\sigma_x^2 \sigma_T^2} = \frac{\sigma_{xT}^2}{\sigma_x^2 \sigma_T^2} = r_{xT}^2 \tag{5.2}$$

It is important to emphasize that true scores are expected values over repeated observations and they do not necessarily correspond to an actual, tangible quantity of interest (Borsboom and Mellenbergh 2002). As a result, every measurement has a true score, regardless of whether it gauges a concrete attribute or not. For example, if we construct a test by summing the responses to the items: "how many languages can you confidently hold a conversation in?" and "Estimate the number of photos you've taken in the last year across all devices". Even in such cases, the test's composite score retains a true score, but this true score does not mirror a tangible reality.

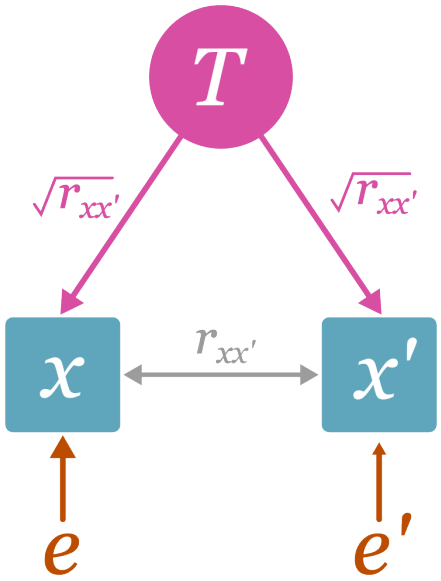


Figure 5.1: Structural diagram illustrating the relationship between true scores, observed scores, and error scores. The pink circle labeled t indicates the true scores, the blue squares labeled with x and x' represent observed scores on parallel measurements, and the red e denotes error. Correlations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

5.3 Reliability vs Validity

Reliability and validity are distinct properties in measurement theory. Validity pertains to whether a measure reflects the quantities it is intended to measure (p. 14 Kelley 1927). According to Borsboom, Mellenbergh, and Van Heerden (2004), a measure is *valid* if the following statements are true:

1. The attribute exists.

2. Variations in the attribute causally produce variations in the outcomes of the measurement procedure.

Borsboom's formulation of validity is simpler and more practical than other formulations such as Cronbach and Meehl's (1955) nomological network approach to validity. It is important to note that even if an attribute does not exist (statement 1), scores may still provide predictive utility. For example, socio-economic status (SES) is a formative quantity that is constructed from a composite of education, income, occupation status, etc. Although SES is not causal to these indicators, SES can still be used as a predictor of important life outcomes.

5.4 Estimating Reliability

In practice, reliability must be estimated through indirect methods, this is due to the fact that true scores and errors are unknown. There are many reliability estimators that can be used, however we will go over a selection of internal consistency estimators as well as test-retest stability estimators.

5.4.1 Internal Consistency Estimators

Taking multiple measurements and then averaging tends to provide a more stable estimate of true values. For instance, let's consider the case of Francis 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 true value. This principle can be extended to broader applications, such as measuring psychological constructs. If we were to assess someone's level of extraversion using ratings from their mother, father, friend, and sibling, the average of their combined assessments would yield a more reliable score compared to relying solely on a single evaluator. So to create a more stable composite score (x), we can take the score from κ measures ($\aleph$) and sum them such that,

$$x = \aleph_1 + \aleph_2 + ... + \aleph_\kappa$$

The most commonly reported reliability estimator in the psychological sciences is coefficient alpha, also referred to as Cronbach's alpha. Coefficient alpha, along with other internal consistency estimators, serves the purpose of assessing the reliability of composite scores comprising multiple measurements. Coefficient alpha reflects an estimate of the reliability of the composite observed score, x ($r_{xx'}$). Coefficient alpha only requires three parameters to calculate, the number of measurements (κ), the variances of each items ($\sigma^2_{\mathrm{i}_m}$), and the variance of the composite score (σ^2_x),

$$_{\alpha}r_{xx'} = \frac{k}{k-1}\left(1 - \frac{\sum_{m=1}^{\kappa}\sigma^2_{\mathbb{x}_m}}{\sigma^2_x}\right) \tag{5.3}$$

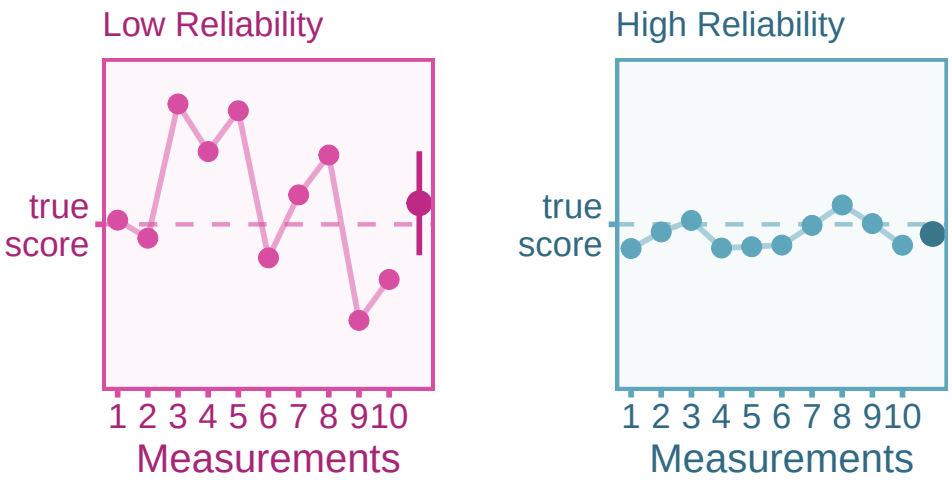


Figure 5.2: Figures showing the observed scores upon 10 repeated measurements and the composite observed score for a single person (the true score is denoted with the dashed line). The left panel shows 10 observed scores with a lot of variation (i.e., low reliability). The composite score (dark red dot with error bars), shows wide error bars illustrating the low precision of the observed score score. The right panel also shows 10 observed scores with little variation (i.e., high reliability). The composite score (dark blue dot with error bars), shows narrow error bars illustrating the high precision of the observed score.

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

$$_{\text{sb}}r_{xx'} = \frac{\kappa\bar{r}_{\mathbb{x}_i\mathbb{x}_j}}{1 + (\kappa - 1)\bar{r}_{\mathbb{x}_i\mathbb{x}_j}} \tag{5.4}$$

This can be simplified further if we have two observed scores. This formulation is a variation of split-half reliability:

$$_{\text{sh}}r_{xx'} = \frac{2r_{\mathbb{x}_1\mathbb{x}_2}}{1 + r_{\mathbb{x}_1\mathbb{x}_2}} \tag{5.5}$$

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.

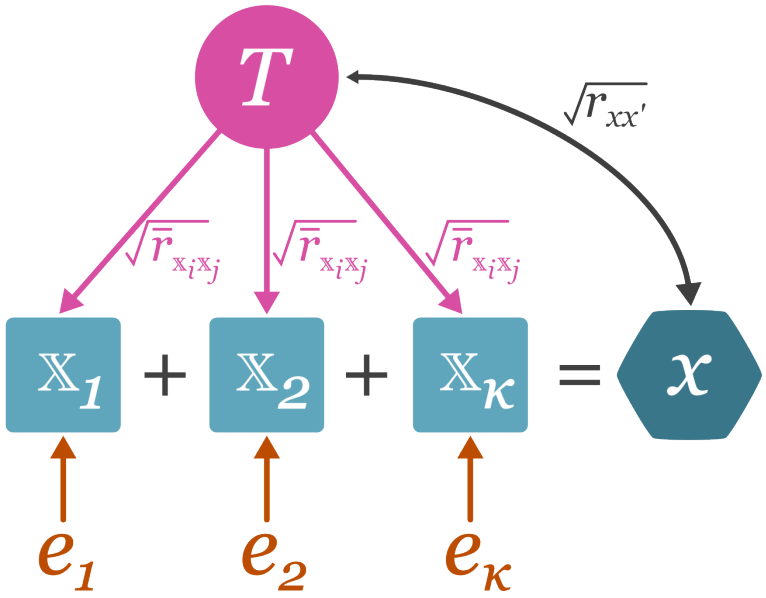


Figure 5.3: Structural model illustrating internal consistency. The pink circle labeled T indicates the true scores, the blue squares, $\mathbb{x}_{1\ldots\kappa}$, represent the observed scores across multiple measurements, and the red e denotes error. The dark blue hexagon, x, indicates a composite score as a sum of the observed scores ($\mathbb{x}_{1\ldots\kappa}$). Note that $\sqrt{r_{xx'}} = r_{xT}$.

5.4.2 Calculating Internal Consistency in R

Let us simulate a data set of 50 individuals that were measured four times resulting in four sets of scores (`x1,x2,x3,x4`) that have the same true score and error variance. Then let us calculate a composite score (`x`) from these sub-scores.

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

# set sample size
n <- 50

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

# calculate composite score
x <- x1 + x2 + x3 + x4
```

Now let us calculate coefficient alpha from the formula provided in Equation **??**.

```r
# step 1. calculate variance of observed (measured) scores
var_xm <- c(var(x1),var(x2),var(x3),var(x4))

# step 2. get number of items (k)
k <- length(var_xm)

# step 3. calculate variance of composite score
var_x <- var(x)

# step 4. calculate coefficient alpha reliability
rxx_alpha <- k / (k-1) * (1 - sum(var_xm)/(var_x))

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

```
[1] 0.775
```

With the simplification of Coefficient alpha's formula, let us calculate the reliability via Spearman-Brown's prophecy formula provided in Equation **??**.

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

```r
# 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
```

If we simplify even further, we can calculate the Split-Half reliability formula provided in Equation **??**,

```r
# step 1. make composite scores for each half of the observed scores
xh1 <- (x1 + x2)/2
xh2 <- (x3 + x4)/2

# step 2. calculate the correlation between the scores of both halves
rx1x2 <- cor(xh1,xh2)

# step 3. calculate the split-half reliability
rxx_SH <- 2*rx1x2 / (1 + rx1x2)

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

```
[1] 0.824
```

Lets see how the results compare to the actual reliability,

```r
# calculate true reliability, true scores must be re-scaled by number of items
rxx = var(k*T_score) / var_x

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

```
[1] 0.734
```

In this case, the reliability estimates do a fairly good job of estimating the true reliability of the observed scores. We can also use the `alpha` function from the `psych` package ("Psych: Procedures for Personality and Psychological Research" 2017) to estimate coefficient alpha too. It also provides additional item level information that is quite useful:

```
# load in package
# install.packages('psych')
library(psych)

# compute summary reliability (only need first table)
alpha(cbind(x1,x2,x3,x4))[[1]]
```

```
raw_alpha std.alpha   G6(smc) average_r      S/N        ase        mean
0.7749847 0.7751377 0.7337024 0.4628829 3.447166 0.05141467 -0.04386823
       sd  median_r
0.9567892 0.4571798
```

5.4.3 Test-Retest Stability Estimator

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. 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 (see Section **??**). To estimate test-retest reliability, we can compute the correlation 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).

5.4.4 Calculating Test-Retest Reliability in R

Lets calculate test-retest reliability in R. First, we can simulate observed scores at two time points, `xTime1` and `xTime2`. We can assume that the true scores remain constant between time points. Second, we can calculate the correlation between the observed scores at each time point (`rxx`).

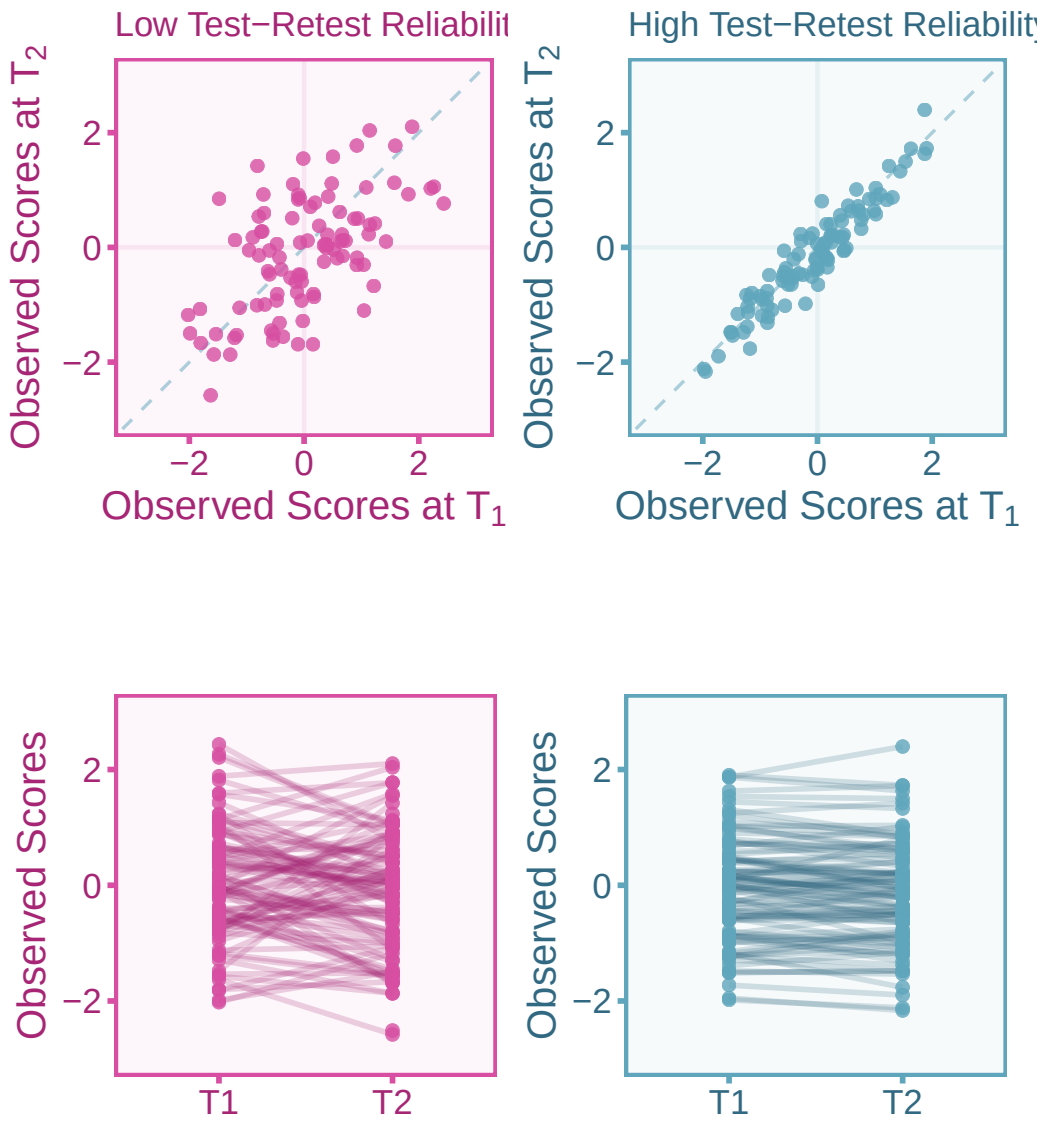


Figure 5.4: Illustrating test-retest reliability. Top-left and top-right panels show the correlation between observed scores at both time-points for a measure that has low and high reliability, respectively. Bottom-left and bottom-right panels show the within-person change from time-point 1 to time-point 2 for scores with low and high reliability, respectively.

```r
# set seed
set.seed(1)

# set sample size
n = 100

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

# simulate scores at time 1
xTime1 = T_score + rnorm(n,0,.5)

# simulate scores at time 2
xTime2 = T_score + rnorm(n,0,.5)

# calculate test-retest reliability
rxx = cor(xTime1,xTime2)

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

```
[1] 0.755
```

```r
# compare with true reliability
rxx_true = var(T_score) / var(xTime1)

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

```
[1] 0.779
```

5.4.5 Sources of Measurement Error

Measurement error variance can itself be broken down into multiple sources of error (e.g., transient,). Depending on the study, different sources of error may be more relevant than others. It is important for a researcher to choose the right reliability estimator for their study since they account for different sources of measurement error. A description of four of the most common sources of error is 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 may show 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.	X		X	
Coefficient Omega	Internal consistency coefficient for composite measures with specified factor structure.	X		X	

Estimator	Description	Random Response Error	Transient Error	Instrument-Specific Error	Rater-Specific Error
Split-Half	Internal consistency coefficient for measurements that are split into two halves.	X		X	
Kuder-Richardson 20	Internal consistency when observed scores are binary (special case of coefficient alpha).	X		X	
Item Response Theory Reliability	Reliability coefficient derived from item response theory (as opposed to classical test theory)	X		X	
Inter-Rater/Inter-Observer Reliability	Consistency in scoring between raters/observers.	X			X
Test-Retest	Stability coefficient for repeated measurements across time	X	X		
Delayed Coefficient Alpha	Average of all possible split-half reliabilities	X	X	X	

Estimator	Description	Random Response Error	Transient Error	Instrument-Specific Error	Rater-Specific Error
G-Coefficient	Reliability coefficient derived from generalizability theory (G-theory). Can incorporate any source of error if enough data is present.	X	X	X	X

5.5 Correction for Bias in Correlations (*r*)

5.5.1 Defining the Estimand

Continuing with our emphasis on clearly defining our quantity of interest (i.e., the estimand) prior to applying any corrections, let us define it. Our estimand here is the population correlation between true scores of our independent and dependent variables. We can define the observed scores of the independent and dependent variables x and y as,

$$x = T + e_x$$

$$y = U + e_y$$

Where T and U are the true scores for the independent and dependent variables, respectively. The true score correlation can thus be be denoted by, ρ_{TU}, and can be defined as the standardized covariance,

$$\rho_{TU} = \frac{\sigma_{TU}}{\sigma_T \sigma_U}$$

In a given study, we will only have knowledge of the observed scores of the independent and dependent variables, x and y, therefore the population observed score correlation is ρ_{xy}. To obtain an unbiased estimate of the true score correlation, we must correct the observed score correlation.

5.5.2 Artifactual Bias and Correction

Measurement error induces systematic bias in effect size estimates such as correlation coefficients Spearman (1904). In the population, let us assume there is some factor a that accounts for the systematic bias in observed score correlations (ρ_{xy}) relative to true score correlations (ρ_{TU}), such that

$$\rho_{xy} = a\rho_{TU}.$$

Since the correlation is defined as the covariance standardized by the standard deviations, the population correlation between true scores, T and U, is defined as

$$\rho_{TU} = \frac{\sigma_{TU}}{\sigma_T \sigma_U}$$

Likewise the correlation between the observed scores, x and y, would be the observed covariance divided by the observed standard deviations.

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

However, if we assume that there is no covariance between errors in x and y ($\sigma_{e_x e_y} = 0$), then the covariance between observed scores is only attributable to the covariance between true scores, therefore $\sigma_{xy} = \sigma_{TU}$. This means that the observed score correlation can be expressed as

$$\rho_{xy} = \frac{\sigma_{TU}}{\sigma_x \sigma_y} \tag{5.6}$$

Now the only difference between the observed score correlation and the true score correlation is the standard deviations in the denominator. In the presence of measurement error, the observed score standard deviations (σ_x and σ_y) will be larger than the true score standard deviations (σ_T and σ_U). Using the definition of reliability, we can show how the observed variance is inflated compared to the true variance as a function of reliability. Since the reliability is defined as the ratio of true variance to total observed variance (see Equation **??**), we can see how reliability inflates the observed variance

$$\sigma_x^2 = \sigma_T^2 \left(\frac{\sigma_x^2}{\sigma_T^2} \right) \tag{5.7}$$

$$= \sigma_T^2 \left(\frac{1}{r_{xx'}} \right) \tag{5.8}$$

$$= \frac{\sigma_T^2}{r_{xx'}}, \tag{5.9}$$

therefore the observed standard deviation is,

$$\sigma_x = \frac{\sigma_T}{\sqrt{r_{xx'}}}. \tag{5.10}$$

If we use the definition of an observed score correlation (Equation **??**), then we can replace σ_x and σ_y with $\frac{\sigma_T}{\sqrt{r_{xx'}}}$ and $\frac{\sigma_U}{\sqrt{r_{yy'}}}$, respectively. Now we can see how the observed score correlation differs from the true score correlation:

$$\rho_{xy} = \frac{\sigma_{TU}}{\left[\frac{\sigma_T}{\sqrt{r_{xx'}}}\right]\left[\frac{\sigma_U}{\sqrt{r_{yy'}}}\right]} \tag{5.11}$$

$$= \frac{\sigma_{TU}}{\sigma_T\sigma_U} \cdot \sqrt{r_{yy'}}\sqrt{r_{xx'}} \tag{5.12}$$

$$= \rho_{TU}\sqrt{r_{yy'}}\sqrt{r_{xx'}} \tag{5.13}$$

This attenuation formula was first derived by Spearman (1904). Note that this formulation requires that there is no correlation between e_x and e_y ($r_{e_x e_y} = 0$). The study observed correlation will also contain sampling error and thus can be expressed by,

$$r_{xy} = \rho_{xy} + \varepsilon_o$$

We can also express it in terms of our estimand, the population true score correlation (ρ_{TU}),

$$r_{xy} = \rho_{TU}\sqrt{r_{xx'}r_{xx'}} + \varepsilon_o$$

It becomes apparent that if we have the reliability of x and y, we can obtain an unbiased estimate of ρ_{TU} by dividing both sides of the above equation by $\sqrt{r_{xx'}r_{xx'}}$ such that,

$$\frac{r_{xy}}{\sqrt{r_{xx'}r_{xx'}}} = \rho_{TU} + \frac{\varepsilon_o}{\sqrt{r_{xx'}r_{xx'}}}$$

Therefore the corrected study correlation, r_c, is defined as,

$$r_c = \frac{r_{xy}}{\sqrt{r_{xx'}r_{xx'}}}.$$

The sampling error of the corrected study correlation is,

$$\varepsilon_c = \frac{\varepsilon_o}{\sqrt{r_{xx'}r_{xx'}}}$$

and thus the sampling variance would be,

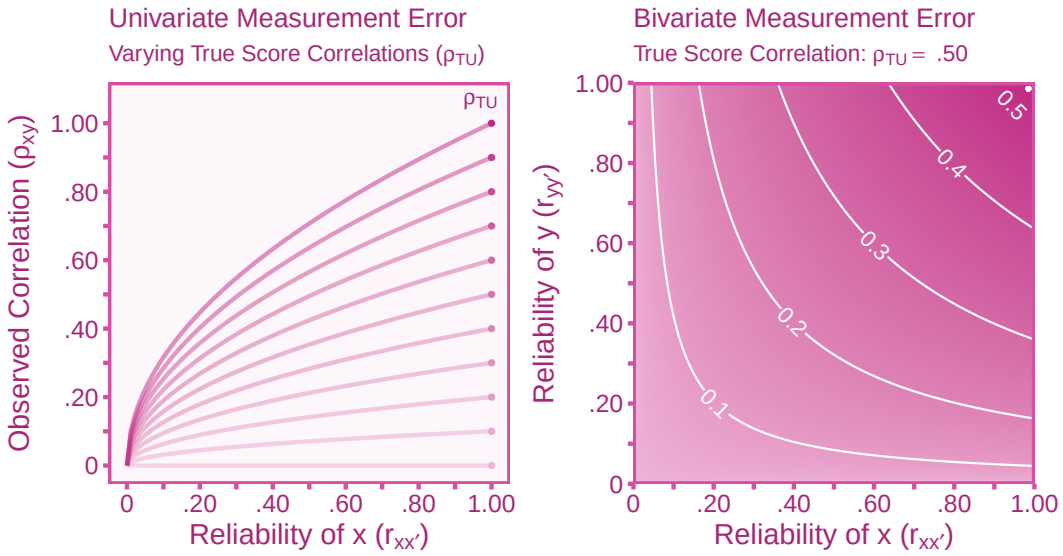


Figure 5.5: Visualizing the attenuation of observed correlation (ρ_{xy}) due to measurement error. The left panel shows a situation where only one variable (x) has measurement error. The observed correlation increases as a function of the true correlation ρ_{TU} (darker lines indicate a higher true score correlation) and the reliability of x (x-axis). The right panel shows the attenuation of the correlation when both x and y variables are affected by measurement error. The darker end of the gradient shows a higher correlation, while the lighter end represents a smaller correlation (the true score correlation sits on the top where no measurement error is present, $r_{xx'} = r_{yy'} = 1$).

$$\sigma^2_{\varepsilon_c} = \frac{\sigma^2_{\varepsilon_o}}{r_{xx'} r_{xx'}}.$$

5.5.3 Correcting Correlations in R

We can simulate continuous data that contains measurement error by using the `simulate_r_sample` function in the `psychmeta` package. Below we will simulate observed scores (`x_score` and `y_score`) and true scores (`T_score` and `U_score`).

```
# load packages
# install.packages('psychmeta')
library(psychmeta)


set.seed(1)

# define parameters
rhoTU <- .5
rxx <- .8
ryy <- .7
n <- 500

# simulate data
data <- simulate_r_sample(n = n,
                          rho_mat = reshape_vec2mat(rhoTU),
                          rel_vec = c(rxx, ryy),
                          sr_vec = c(1, 1),
                          var_names = c("x","y"))

# obtain observed scrores
x_score <- data$data$observed$x
y_score <- data$data$observed$y

# obtain true scores
T_score <- data$data$true$x
U_score <- data$data$true$y
```

Then we can compute observed score (`rxy`).

```r
# compute observed score correlation and standard error
rxy <- cor(x_score,y_score)

# compute sampling variance of observed score correlation
var_e_o <- (1-rxy^2) / (n-2)

# print results
print(paste0('rxy = ',round(rxy,3),'  var_e_o = ',round(var_e_o,4)))
```

```
[1] "rxy = 0.351  var_e_o = 0.0018"
```

Let us now compare the observed correlation with the true score correlation (`rTU`).

```r
# compute observed score correlation and standard error
rTU <- cor(T_score,U_score)

# compute sampling variance of observed score correlation
var_e <- (1-rTU^2) / (n-2)

# print results
print(paste0('rTU = ',round(rTU,3),'  var_e = ',round(var_e,4)))
```

```
[1] "rTU = 0.463  var_e = 0.0016"
```

The observed correlation is substantially lower than the true score correlation. In order to correct the observed score correlation, we can calculate it by hand or use the `correct_r()` function. Lets first correct by hand using the equations in Section **??**

```r
# correct correlation coefficient
rc <- rxy / sqrt(rxx*ryy)

# correct sampling variance
var_e_c <- var_e_o / sqrt(rxx*ryy)

# print results
print(paste0('rc = ',round(rc,3),'  var_e_c = ',round(var_e_c,4)))
```

```
[1] "rc = 0.47  var_e_c = 0.0024"
```

Now lets correct the correlation with the `correct_r()` function,

```r
# correct correlation
correct_r(rxyi = rxy,
          rxx = rxx,
          ryy = ryy,
          n = n)
```

```
Correlations Corrected for Measurement Error:
--------------------------------------------------------------------------------
  value CI_LL_95 CI_UL_95   n n_effective
1  0.47    0.364    0.569 500         222
```

As we can see, the corrected correlation ($r_c = .470$) is a more accurate estimate of the true score population correlation $\rho_{TU} = .500$, than the observed score correlation ($r_{xy} = .351$).

5.6 Correction for Bias in Standardized Mean Differences (*d*)

5.6.1 Defining the Estimand

Prior to correcting for measurement error let us define our estimand. Our estimand here is the difference in the means of group A and B with respect to the true scores of our dependent variable. We can define the observed scores of the independent and dependent variables x and y as,

$$y_A = U_A + e_A$$

$$y_B = U_B + e_B$$

Where U_A and U_B are the true scores for group A and group B, respectively. The true score standardized mean difference can thus be be denoted by, δ_U, and can be defined as

$$\delta_U = \frac{\overline{U}_A - \overline{U}_B}{\sigma_{U_P}}$$

Where $\overline{U}$ is the mean of true scores for the respective group. In a given study, we only have access to the observed scores of the independent and dependent variables, x and y, therefore the population observed score correlation is δ_y. To obtain an unbiased estimate of δ_U, we must apply a correction to the observed score standardized mean difference.

5.6.2 Artifactual Bias and Correction

We can calculate the standardized mean difference (SMD) of the observed scores by dividing the mean difference in observed scores ($\bar{y}_A - \bar{y}_B$) by the pooled standard deviation (σ_p). It is important to note that the mean of true scores and the mean of observed scores will be identical due to the fact that measurement error only affects variance in scores. Therefore, we can express the observed standardized mean difference as,

$$d = \frac{\bar{y}_A - \bar{y}_B}{\sigma_{y_P}} = \frac{\overline{U}_A - \overline{U}_B}{\sigma_{y_P}}$$

The pooled standard deviation is a weighted average of the observed score standard deviations,

$$\sigma_{y_P} = \sqrt{\frac{(n_A + 1)\sigma^2_{y_A} + (n_B + 1)\sigma^2_{y_B}}{n_A + n_B - 2}}.$$

To express σ_{y_P} in terms of the true score standard deviations, we can place the observed score standard deviations with the attenuated true score standard deviation in Equation **??**,

$$\sigma_{y_P} = \sqrt{\frac{(n_A + 1)\left(\frac{\sigma^2_{U_A}}{r_{yy'_A}}\right) + (n_B + 1)\left(\frac{\sigma^2_{U_B}}{r_{yy'_B}}\right)}{n_A + n_B - 2}}.$$

Alternatively, we can pool the reliability and the true score standard deviations separately so that we can obtain a simplified version of the above equation,

$$\sigma_{U_P} = \sqrt{\frac{(n_A + 1)\sigma^2_{U_A} + (n_B + 1)\sigma^2_{U_B}}{n_A + n_B - 2}}.$$

$$r_{yy'_P} = \sqrt{\frac{(n_A + 1)r^2_{yy'_A} + (n_B + 1)r^2_{yy'_B}}{n_A + n_B - 2}}.$$

Then we can express σ_{y_P} similarly to Equation **??**,

$$\sigma_{y_P} = \frac{\sigma_{U_P}}{\sqrt{r_{yy'_P}}}$$

Now we can put it all together and see how the observed score standardized mean difference (δ_y) is biased relative to the true score standardized mean difference (δ_U),

$$\delta_y = \frac{\bar{y}_A - \bar{y}_B}{\sigma_{y_P}} \tag{5.14}$$

$$= \frac{\overline{U}_A - \overline{U}_B}{\sigma_{y_P}} \tag{5.15}$$

$$= \frac{\overline{U}_A - \overline{U}_B}{\frac{\sigma_{U_P}}{\sqrt{r_{yy'_P}}}} \tag{5.16}$$

$$= \frac{\overline{U}_A - \overline{U}_B}{\sigma_{U_P}} \sqrt{r_{yy'_P}} \tag{5.17}$$

$$= \delta_U \sqrt{r_{yy'_P}} \tag{5.18}$$

This attenuation bias is very similar to the one we saw in the correlation, with the only difference being that the pooled reliability is used here instead of the total sample reliability. Within a study, the observed score standardized mean difference (d_y) will not only be attenuated by measurement error, but it will also contain sampling error such that,

$$d_y = \delta_y + \varepsilon_o$$

replacing the observed population SMD, δ_y, with $\delta_U \sqrt{r_{yy'_P}}$, gives us,

$$d_y = \delta_U \sqrt{r_{yy'_P}} + \varepsilon$$

Therefore to obtain the corrected study SMD (d_c) we can divide d_y by the attenuation factor,

$$d_c = \frac{d_y}{\sqrt{r_{yy'_P}}}$$

Where the sampling variance of the corrected SMD must also be similarly adjusted,

$$\sigma^2_{\varepsilon_c} = \frac{\sigma^2_{\varepsilon_o}}{r_{yy'_P}}$$

Although the attenuation factor is quite simple, in more complex formulations (e.g., bivariate direct range restriction), it will be easier to apply a simplified correction for the sampling variance using the corrected correlation coefficient:

$$\sigma^2_{\varepsilon_c} = \sigma^2_{\varepsilon_o} \left(\frac{d_c}{d_y} \right)^2$$

It is important to point out that this correction can only be done if when estimates of the within-group reliability are available. It is common that studies will only report the full sample reliability. If there are differences between groups on the variable, the total sample reliability will over-estimate the within-group reliability. When the total sample reliability is all that is available, to correct d_y, we must first convert it to a point-biserial correlation coefficient (r_{pb}) using the observed proportion of subjects in either group A or B (p; it does not matter which one is chosen, as long as it is consistent throughout).

$$r_{pb} = \frac{d}{\sqrt{\frac{1}{p(1-p)} + d^2}}.$$

Then correct r_{pb} for the total sample reliability,

$$r_c = \frac{r_{pb}}{\sqrt{r_{yy'}}}$$

Then we can convert r_c back into d_c,

$$d_c = \frac{r_c}{\sqrt{p(1-p)(1-r_c^2)}}$$

The same process of converting to a point-biserial correlation and back to a standardized mean difference can be done for the sampling variance as well, but instead we can put it all into one equation,

$$\sigma^2_{\varepsilon_c} = \frac{\sigma^2_{\varepsilon_o} \left(\frac{r_c}{r_{pb}}\right)^2}{\left(1 + d^2 p[1-p]\right)^3 (1-r_c^2)^3}$$

5.6.3 Correcting *d* values in R

We can simulate data that contains measurement error by using the `simulate_d_sample` function in the `psychmeta` package. Below we will simulate observed scores (`y_score`) and true scores (`U_score`).

```
# load packages
# install.packages('psychmeta')
library(psychmeta)

set.seed(123)

# define parameters
```

```
Means = c()
ryyA <- .75
ryyB <- .70
nA <- 100
nB <- 100
n <- nA + nB

# simulate data
data<- simulate_d_sample(n_vec = c(nA, nB),
                         rho_mat_list = list(reshape_vec2mat(1),reshape_vec2mat(1)),
                         mu_mat = rbind(c(.5,0),
                                        c(0,0)),
                         sigma_mat = rbind(c(1,1),
                                           c(1,1)),
                         rel_mat = rbind(c(ryyA,1),
                                         c(ryyB,1)),
                         sr_vec = c(1,1),
                         group_names = c("A", "B"))

# obtain observed scores
y_score <- data$data$observed$y1
group <- data$data$observed$group

# obtain true scores
U_score <- data$data$true$y1
```

Then we can compute observed score standardized mean difference (`dy`).

```
# compute observed score means and standard deviations
Mean_A <- mean(y_score[group=='A'])
Mean_B <- mean(y_score[group=='B'])
SD_A <- sd(y_score[group=='A'])
SD_B <- sd(y_score[group=='B'])

# compute pooled standard deviation
SD_P <- sqrt( ((nA-1)*SD_A^2 + (nB-1)*SD_B^2) / (nA+nB-2) )

# compute standardized mean difference
dy <- (Mean_A - Mean_B) / SD_P

# compute sampling variance of observed score correlation
```

```
var_e_o <- n/(nA*nB) + dy^2 / (2*n)

# print results
print(paste0('dy = ',round(dy,3),'  var_e_o = ',round(var_e_o,4)))
```

```
[1] "dy = 0.273  var_e_o = 0.0202"
```

Let us now compare the observed score SMD with the true score SMD (`dU`).

```
# compute true score means and standard deviations
Mean_A <- mean(U_score[group=='A'])
Mean_B <- mean(U_score[group=='B'])
SD_A <- sd(U_score[group=='A'])
SD_B <- sd(U_score[group=='B'])

# compute pooled standard deviation
SD_P <- sqrt( ((nA-1)*SD_A^2 + (nB-1)*SD_B^2) / (nA+nB-2) )

# compute standardized mean difference
dU <- (Mean_A - Mean_B) / SD_P

# compute sampling variance of the true score SMD
var_e <- n/(nA*nB) + dy^2 / (2*n)

# print results
print(paste0('dU = ',round(dU,3),'  var_e = ',round(var_e,4)))
```

```
[1] "dU = 0.509  var_e = 0.0202"
```

The observed score SMD is substantially lower than the true score SMD (.286 vs .509). In order to correct the observed score correlation for attenuation, we can calculate it by hand. Lets correct the observed SMD for measurement error variance using the equations in Section **??**.

```
# calculate the pooled reliability
ryy_P <- sqrt(((nA-1)*ryyA^2 + (nB-1)*ryyB^2) / (nA+nB-2))

# correct correlation coefficient
dc <- dy / sqrt(ryy_P)

# correct sampling variance
```

```r
var_e_c <- var_e_o / sqrt(ryyA)

# print results
print(paste0('rc = ',round(dc,3),'  var_e_c = ',round(var_e_c,4)))
```

```
[1] "rc = 0.32  var_e_c = 0.0233"
```

Now lets correct the correlation with the `correct_r()` function. The `correct_d()` function only takes in the total sample reliability, therefore we can extract the total sample reliability from the simulated dataset and then use the resulting reliability coefficient in the `ryy` argument.

```r
# total sample reliability
ryy = data$overall_results$observed$parallel_ryyi_total[1]

# correct correlation
correct_d(d = dy,
          ryy = ryy,
          n1 = n)
```

```
d Values Corrected for Measurement Error:
--------------------------------------------------------------------------------
  value CI_LL_95 CI_UL_95   n n_effective
1 0.322 -0.00878    0.667 200         142
```

As we can see, the corrected correlation ($d_c = .32$) is a more accurate estimate of the true score population SMD $\delta_U = .500$, than the observed score correlation ($r_{xy} = .273$).

5.7 Estimating Reliability with Limited Information

Reliability estimates should preferably be calculated from within the study's sample, however there are a couple of ways to estimate reliability when this information is not provided. A common way to obtain an estimate of the reliability is to look in meta-analyses or a test manuals. If the number of items in a study differs from the test manual, you can approximate the reliability of a study's test, with a re-arrangement of the Spearman-Brown prophecy formula,

$$r_{xx'_{study}} \approx \frac{1}{\frac{k_{\text{ref}}}{k_{\text{study}}}\left(\frac{1}{r_{xx'_{study}}} - 1\right) + 1}$$

Where k_{ref} and k_{study} denote the number of items in the reference test and the test used in the study, respectively.

(Haertel 2006)

(F. L. Schmidt, Le, and Ilies 2003)

(Gliem and Gliem 2003)

(Bobko, Roth, and Bobko 2001)

(Mendoza and Mumford 1987)

(Brennan 2010)

(Viswanathan 2005)

(Viswesvaran et al. 2014)

(Sijtsma 2009)

(Charles 2005)

(Spearman 1904)

6 Group Misclassification

6.1 Introduction

Group misclassification describes a situation where true group membership (e.g., people with a disorder) does not perfectly match the observed group membership (e.g., people *diagnosed* with a disorder). Group misclassification can be considered a type of measurement error where instead of accounting for errors in continuous variables (i.e., unreliability), group misclassification accounts for errors in categorical variables.

6.2 Defining Group Misclassification

Misclassification can be defined as any deviations between true group membership and observed group membership. Let us imagine two arbitrary groups, group A and group B. In order to identify members of group A and group B, we have to use some measurement instrument. Also let us assume that this measurement instrument produces imperfect group classifications, that is, people who are actually in group A are sometimes assigned group B and vice versa. We can visualize the performance of the classification procedure with a contingency table between actual group membership (G) and observed group membership (g):

	$G = A$	$G = B$
$g = A$	AA	BA
$g = B$	AB	BB

We can see from the contingency table that subjects who were correctly classified, would be labeled in the cell block AA or BB and those who were misclassified would belong to cells BA and AB. Therefore we can define the proportion of people that are accurately classified as $p_{\text{acc}} = P(AA) + P(BB)$ whereas the proportion of people misclassified can be defined as $p_{\text{mis}} = P(AB) + P(BA)$. A high-quality classifier would would minimize p_{mis} and maximize p_{acc}. Additionally, note that the proportion of people misclassified is inversely proportional to the proportion of people accurately classified such that, $p_{\text{mis}} = 1 - p_{\text{acc}}$.

6.3 Classification Reliability

Similar to quantifying reliability in continuous variables by calculating the correlation in parallel sets of observed scores, the same can be done in categorical variables. Instead of a contingency table between observed (g) and true (G) group membership, we will instead create a contingency table of two measurements producing two sets of observed classifications (g and g'). Measurements often will take the form of inter-rater assessments, for example, two clinician's diagnosis of Major Depressive Disorder (MDD) in the same sample of patients.

	$g = A$	$g = B$
$g' = A$	AA	BA
$g' = B$	AB	BB

To obtain the reliability of the group assignments, we can calculate the correlation coefficient between G and G'. Since both variables are categorical, a Pearson correlation coefficient would not be an appropriate correlation estimator, instead, we must compute the phi coefficient. The phi coefficient is often referred to as Matthew's correlation coefficient and is most frequently used as an index of performance of a binary classifier in machine learning. For the sake of consistency, the phi coefficient will be denoted with the letter r, and thus the reliability (i.e., the correlation between G and G') is denoted with $r_{GG'}$.

There are a few ways we can calculate the phi coefficient. The first way is to calculate phi directly from the contingency table,

$$r_{gg'} = \frac{n_{AA}n_{BB} - n_{AB}n_{BA}}{\sqrt{(n_{AA} + n_{BA})(n_{AB} + n_{BB})(n_{AA} + n_{AB})(n_{BA} + n_{BB})}}.$$

Where n_{AA}, n_{BB}, n_{AB}, and n_{BA} are the number of subjects within their respective cells of the contingency table. If the values of the contingency table are not available, we can calculate the phi coefficient from the χ^2-statistic,

$$r_{gg'} = \sqrt{\frac{\chi^2}{n}}.$$

Where n is the total sample size. If the χ^2-statistic is unavailable, we can approximate the phi coefficient from the accuracy (p_{acc}) or the proportion of people misclassified (p_{mis}),

$$r_{gg'} = (2p_{\text{acc}} - 1)^2 = (1 - 2p_{\text{mis}})^2$$

This approximation assumes that the group sizes are approximately equal *and* the misclassification rates are approximately equal between groups. Otherwise, $r_{gg'}$ will be overestimated (Wiernik and Dahlke 2020).

In the chapter on unreliability, we discussed the relationship between reliability and the correlation between observed and true scores. The classification reliability will also be related similarly to the correlation between observed group membership and true group membership (r_{gG}) such that,

$$r_{gG} = \sqrt{r_{gg'}}$$

6.4 Calculating Classification Reliability in R

To calculate classification reliability we will first need data. We can simulate 100 subjects with a group value for three variables: a true group membership and two sets of assigned (observed) group membership. We will set the misclassification rate to 10%.

```
# set seed
set.seed(17)

# 10% misclassification rate
p_mis <- .10

# sample size of 100
nA <- 50
nB <- 50
n = nA + nB

# create a vector of true group values
true_A <- rep('A',nA)
true_B <- rep('B',nB)
true_group <- c(true_A,true_B)

# initialize vectors of observed group membership from true group membership
obs_1_A <- true_A
obs_1_B <- true_B
obs_2_A <- true_A
obs_2_B <- true_B

# add misclassified values to observed group membership
obs_1_A[sample(1:nA,nA*p_mis)] <- 'B'
obs_1_B[sample(1:nB,nB*p_mis)] <- 'A'
obs_2_A[sample(1:nA,nA*p_mis)] <- 'B'
obs_2_B[sample(1:nB,nB*p_mis)] <- 'A'
```

```
obs_1_group <- c(obs_1_A,obs_1_B)
obs_2_group <- c(obs_2_A,obs_2_B)
```

Then we can generate a contingency table of the two sets of observed group assignments.

```
# create contingency table of the two observed group memberships
con_table <- table(data.frame(obs_1=obs_1_group,obs_2=obs_2_group))
print(con_table)
```

```
     obs_2
obs_1  A  B
    A 40 10
    B 10 40
```

Now we can calculate the reliability of the group assignments by extracting the phi coefficient from the contingency table. We can compute it by hand or by using the `psych` package by William Revelle (2017).

```
## Strategy 1: Using the {psych} package
# load in psych package (make sure it is installed first: install.packages('psych'))
library(psych)
rgg = phi(con_table,digits = 3)
print(rgg)
```

```
[1] 0.6
```

```
## Strategy 2: calculate from contingency table values
numerator <- con_table['A','A']*con_table['B','B'] - con_table['A','B']*con_table['B','A']
denominator <- sqrt(con_table['A','A']+con_table['A','B']) *
               sqrt(con_table['B','A']+con_table['B','B']) *
               sqrt(con_table['A','A']+con_table['B','A']) *
               sqrt(con_table['A','B']+con_table['B','B'])

rgg <- numerator / denominator
print(rgg)
```

```
[1] 0.6
```

```r
## Strategy 3: calculate from chi-square test
chi2 <- as.numeric(chisq.test(con_table)$statistic)
rgg <- sqrt(chi2/n)
print(rgg)
```

```
[1] 0.58
```

```r
## Strategy 4: calculate from proportion of people misclassified
rgg <- (1-2*p_mis)^2
print(rgg)
```

```
[1] 0.64
```

6.5 Bias in Standardized Mean Difference

Standardized mean differences will become biased when subject's assigned groups differ from their actual group. This is largely due to the fact that the means of each group are driven closer to one another. Let us suppose that, on average, group A and group B score differently on some outcome, y. The true mean of y for groups A and B can be denoted as $\bar{y}_A^*$ and $\bar{y}_B^*$, respectively. Nonetheless, when some subjects are erroneously assigned to the wrong group, the *observed* mean within each group will reflect a weighted average of the respective means. This is due to the fact that the misclassified individuals are being drawn from a population with a different mean. To calculate the mean of the observed groups we must incorporate the true mean of the correctly classified subjects and the misclassified subjects

$$\bar{y}_A = \left(\frac{n_{AA}}{n_{AA} + n_{BA}}\right)\bar{y}_A^* + \left(\frac{n_{BA}}{n_{AA} + n_{BA}}\right)\bar{y}_B^*$$

$$\bar{y}_A = \left(\frac{n_{BB}}{n_{BB} + n_{AB}}\right)\bar{y}_B^* + \left(\frac{n_{BA}}{n_{BB} + n_{BA}}\right)\bar{y}_A^*$$

From the above equations, it becomes evident that as the number of misclassified individuals increases (n_{AB} and n_{BA}), the observed means of each group gradually converge towards each other. As the means converge, the standardized mean difference will correspondingly shift toward zero. To illustrate this phenomenon, Figure **??** shows the distributions for groups A and B without any misclassification. In this case, there is no attenuation of the standardized mean difference.

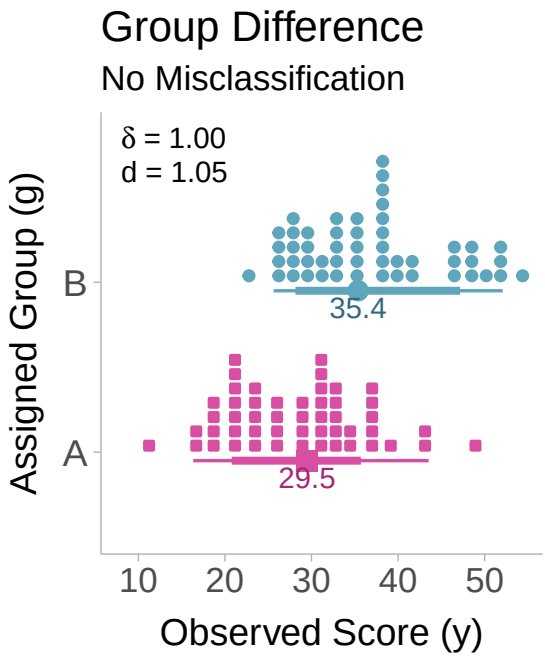


Figure 6.1: Distributions of scores without misclassification. True mean difference and observed mean differ only due to sampling error.

If some individual's are assigned to the incorrect group, then we will see attenuation in the standardized mean difference as the means converge. Figure **??** shows what happens when the misclassification rate is 10%. A misclassification rate of 10% is equivalent to a classification reliability of $r_{GG'} = .60$.

The bias in the standardized mean difference can be expressed as a function of the classification reliability ($r_{gg'}$). To illuminate this bias, we must first convert the true SMD to a point-biserial correlation coefficient (ρ) using the proportion of individuals in group A (p_A) and group B (p_B),

$$\rho = \frac{\delta}{\sqrt{\frac{1}{p_A p_B} - \delta^2}}$$

Then attenuation of the correlation is similar to the attenuation of correlation coefficients in the section on unreliability ($r = \rho\sqrt{r_{xx'}}$). However in this case, we also need to convert the point-biserial correlation to the SMD:

$$d = \frac{\rho\sqrt{r_{gg'}}}{\sqrt{p_A p_B \left(1 - r_{gg'} r^2\right)}}.$$

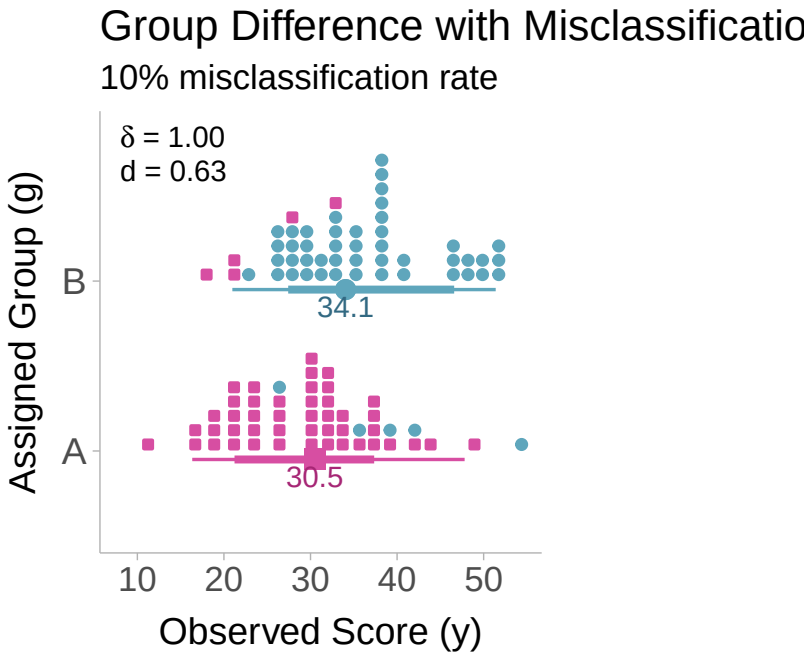


Figure 6.2: Distributions of scores with a 10% misclassification rate. Observed standardized mean differences are biased toward the null (i.e., SMD = 0). Note that a few members of group A (red squares) are within observed group B and vice versa (indicative of misclassification).

It is important to note that for many of the biasing effects and corrections, converting the standardized mean difference to a point-biserial correlation is often a necessary step. However once the corrected point-biserial correlation is obtained, the correlation can then be converted back into a standardized mean difference like we see in the last equation.

6.5.1 Correction for Bias in Standardized Mean Difference

To correct for bias induced by misclassification we first need to convert the observed standardized mean difference to a point-biserial correlation coefficient by using the observed proportion of the sample that has been assigned to either group A or group B (p). The group proportion p in the following equations will only show up in the term $p(1-p)$ so it will not matter which group is used. Converting d to r:

$$r = \frac{d}{\sqrt{\frac{1}{p(1-p)} - d^2}}.$$

We can then correct the point-biserial correlation for group misclassification with the classification reliability:

$$\hat{\rho} = \frac{r}{\sqrt{r_{gg'}}}$$

Now we can convert the corrected point-biserial correlation ($\hat{\rho}$) into a corrected standardized mean difference ($\hat{\delta}$). When converting back to a standardized mean difference, we need to use the true group proportions, p^*. Although if we are to assume equal misclassification rates between groups, then the observed proportion can be used p:

$$\hat{\delta} = \frac{\hat{\rho}}{\sqrt{p^* \left(1 - p^*\right) \left(1 - \hat{\rho}^2\right)}}$$

The sampling variance of $\hat{\delta}$ ($s_{\hat{\delta}}$) will need to be adjusted accordingly. The three-step process for converting to a correlation, correcting, and converting back to a standardized mean difference can instead be done in a single step. Therefore the adjusted sampling variance (squared standard error) can be calculated as,

$$s^2_{\hat{\delta}} = \frac{s_d \left(\frac{\hat{\rho}}{r}\right)^2}{\left(1 + d^2 p[1-p]\right)^2 \left(d^2 + \frac{1}{p(1-p)}\right) p^*(1-p^*)(1-\hat{\rho}^2)^3}$$

This can be simplified if we assume that misclassification rates are equal between groups,

$$s^2_{\hat{\delta}} = \frac{s_d \left(\frac{\hat{\rho}}{r}\right)^2}{\left(1 + d^2 p[1-p]\right)^3 (1-\hat{\rho}^2)^3}$$

6.6 Correcting for Misclassification in R

We can correct for group misclassification in R by hand or by using the `psychmeta` package (Dahlke and Wiernik 2019). For our correction, say we got an observed standardized mean difference of $d = 0.50$ and we calculated the classification reliability to be $r_{gg'} = .80$. Let us also say that the observed *and* the true proportion of individuals in one of the groups is $p = p^* = .40$, therefore the other group would be $1 - p = 1 - p^* = .60$.

```
d = .50
rgg = .70
nA = 40
nB = 60
```

Using the *psychmeta* package

The `psychmeta` package has a function, `correct_d`, that is dedicated to correcting standardized mean differences multiple types of artifacts including group misclassification.

```
# step 1: install and load in psychmeta
# install.packages{'psychmeta'}
library(psychmeta)

# step 2: calculate proportion of group membership
p = nA / (nA + nB)
# p = nB / (nA + nB) # alternative calculation

# step 3: correct d for group misclassification
correct_d(d = d,
          rGg = sqrt(rgg), # square root of rgg = rGg
          correction = "meas",
          pi = p,
          pa = p,
          n1 = nA+nB,
          correct_bias = FALSE)
```

```
d Values Corrected for Measurement Error:
------------------------------------------------------------------------------------------
  value CI_LL_95 CI_UL_95   n n_effective
1 0.618    0.118     1.18 100        66.6
```

The output provides the corrected standardized mean difference (`value`), the upper and lower 95% confidence intervals (`CI_LL_95` and `CI_UL_95`), the sample size (`n`), and the effective sample size (`n_effective`).

Correcting by hand

To calculate the corrected standardized mean difference, we can use the equations in Section **??**.

```
## Calculate point estimate
# step 1: convert d to r
r = d / sqrt( 1/(p*(1-p)) + d^2)

# step 2: correct r
rho = r / sqrt(rgg)

# step 3: convert r to d
delta = rho / sqrt( p*(1-p)*(1-rho^2) )

## Calculate sampling variance
# step 1: compute sampling variance for r
v_d = (nA+nB)/(nA*nB) + d^2 / (2*(nA+nB))

# step 2: adjust sampling variance for correction
v_delta = (v_d * (rho/r)^2) / ((1 + d^2)^3 * p*(1-p) * (1-rho^2)^3)

# print results
print(paste0('delta hat = ',round(delta,3),',  var = ', round(v_delta,3)))
```

```
[1] "delta hat = 0.605,  var = 0.168"
```

(Chyou 2007)

(Wiernik and Dahlke 2020)

(J. E. Hunter and Schmidt 1990)

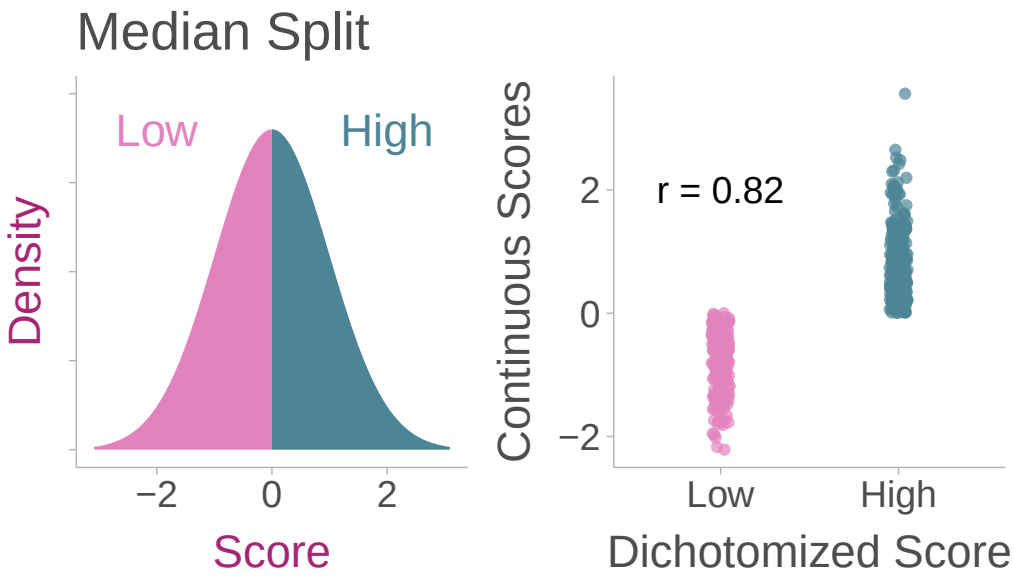
7 Artificial Dichotomization

7.1 Introduction

Primary studies sometimes will splitting naturally continuous variables into two discrete groups to increase interpretability or conduct specific analyses (e.g., t-tests). However, artificially dichotomizing variables introduces measurement error variance thus attenuating effect size estimates Maxwell and Delaney (1993). Clinical disorder diagnoses, such as generalized anxiety disorder, are examples of dichotomization where individuals are separated into either having the disorder or not even though individual differences in anxiety exist as a continuum.

7.2 Artificial Dichotomization Induced Measurement Error

Variables that are dichotomized contain measurement error. This can be demonstrated by the simple fact that dichotomized scores are not perfectly correlated with continuous scores. To demonstrate this, we can draw a sample of scores and then split the data into high and low scorers and then find the correlation coefficient between the two (see figure below). It becomes apparent that the dichotomized scores leave a lot of the variation in scores unaccounted for.



Even with a perfectly reliable measure, dichotomization will introduce measurement error variance. We can define naturally continuous scores (‡) that have been artificially dichotomized as,

$$x_{‡} = \begin{cases} 1, & \text{if } x > C_x \\ 0, & \text{if } x < C_x \end{cases}$$

Where C_x is the cut-score on the standard normal distribution. The reliability can be defined as the correlation between dichotomized scores and the underlying continuous scores ($r_{x_{‡}x}$).

7.3 Correcting Correlations for Artificial Dichotomization

7.3.1 Defining our estimand

Ultimately, we would like to know the correlation coefficient between two naturally continuous variables. Sticking with our notation for true scores, our estimand can be defined as the population correlation between continuous observed scores of the independent (x) and dependent variable (y), ρ_{xy}. Where dichotomized scores can be defined as,

$$x_{‡} = \begin{cases} 1, & \text{if } x > C_x \\ 0, & \text{if } x < C_x \end{cases}$$

$$y_{‡} = \begin{cases} 1, & \text{if } y > C_y \\ 0, & \text{if } y < C_y \end{cases}$$

Where C_y is the cut-score where the split took place. There are two cases of dichotomization that may occur in a given study: the univariate case where only one variable (either dependent or independent) is dichotomized and the bivariate case where both variables are dichotomized. Both of these situations will be addressed in the next section.

7.3.2 Artifact Correction for Correlations

The Univariate Case

In the simplest case of dichotomization, only one variable is dichotomized and the other is left continuous. In this case, a Pearson product-moment correlation is equivalent to the *point-biserial* correlation coefficient, however for dichotomized data, the *biserial* correlation is a relatively unbiased estimate of the pearson correlation on the underlying continuous data (assuming normality). Therefore in the population, the observed correlation $\rho_{x_{‡}y}$ is biased by some attenuation factor a,

$$\rho_{x_{‡}y} = a\rho_{xy}$$

The first step in estimating the attenuation of the correlation is to first identify the cut-point, C_x, of standard normal distribution where the split of the data occurred. This can be calculated by first obtaining the percent of the of the individuals in the low or high scoring group:

$$p_x = \frac{n_{\text{high}}}{n_{\text{high}} + n_{\text{low}}}$$

or

$$p_x = \frac{n_{\text{low}}}{n_{\text{high}} + n_{\text{low}}}.$$

Then we can use the quantile function (ϕ^{-1}; i.e., the inverse of the cumulative density of the standard normal distribution) to obtain the cut-point on the standard normal distribution,

$$C_x = \phi^{-1}(p_x)$$

Using the cut-point and the proportion of group membership in either the low or high scoring group (p_x), the attenuation factor can be defined as (J. Hunter and Schmidt 1990),

$$a = \frac{\varphi(C_x)}{\sqrt{p_x(1-p_x)}}$$

Where φ is the normal ordinate function (i.e., probability density function of a standard normal distribution). Since a standard normal distribution is symmetric, the sign of C_x does not matter. In the case of a median split, where the cut-point would be placed at zero of a standard normal (splitting the distribution in equal halves), the attenuation factor would simplify to $a = \frac{\varphi(0)}{\sqrt{.5(.5)}} = \frac{2}{\sqrt{2\pi}}$. To correct the pearson correlation when one of the variables is dichotomized, we can divide the observed correlation by the attenuation factor such that, $r_c = \frac{r_{x_D y}}{a}$. Therefore the full correction equation is,

$$r_c = \frac{r_{x_\ddagger y}}{\left[\frac{\varphi(C_x)}{\sqrt{p_x(1-p_x)}}\right]} \tag{7.1}$$

Where the sampling variance of the corrected correlation must also be adjusted using the compound attenuation factor,

$$\sigma^2_{\varepsilon_c} = \frac{\sigma^2_{\varepsilon_o}}{a^2} = \frac{\sigma^2_{\varepsilon_o}}{\left[\frac{\varphi(C_x)^2}{p_x(1-p_x)}\right]}$$