Artifact Corrections for Effect Sizes

Implementation in R and Application to Meta-Analysis

Matthew B. Jané

2023-10-08

Table of contents

Chapter 1

Greetings

Welcome to the living open source textbook ***Correcting Effect Sizes for Statistical Artifacts***. This textbook covers all the essential equations and code needed to correct for biases in our effect size estimates. It will also hopefully provide readers with a deeper understanding, appreciation, and intuition for these seemingly complex formulas. It also covers how to apply these corrections to meta-analysis.

What are Statistical Artifacts?

In this book statistical artifacts will be defined broadly as **any source of contamination that induces bias in research findings**. Artifacts are present in virtually every research study, so it is crucial that we address them.

Living Textbook

A living textbook is one that constantly updates with new features and is open to changes from others. This book will contain modern methods and cutting-edge techniques for artifact corrections, so in order to keep this book up-to-date it needs to grow as the research grows. New features, such as interactive figures will be added soon

Open Science Practices

The current maintainer of the site is Matthew B. Jané.

It is important that this book is both open-source and open-access. is also open source (CC-BY license). All the figures, code, and documents are available in my github repository.

Contributions

Please feel free to contribute to this textbook, if your contribution makes it to the published version of this book, your name will be included in the contributor list below with a description of your work.

Cite This Work

APA

Jané, M. B. (2023). Correcting Effect Sizes for Statistical Artifacts: Implementation in R and Application to Meta-Analysis. (n.p.). https://matthewbjane.quarto.pub/artifact-corrections-for-effect-sizes/

BibTeX

```
@book{MatthewBJane2023,
  title     = "Correcting Effect Sizes for Statistical Artifacts: Implementation in
  author    = "Jané, Matthew B.",
  year      = 2023,
  publisher = "(n.p.)",
  url      = {https://matthewbjane.quarto.pub/artifact-corrections-for-effect-sizes/
}
```

Chapter 2

Dedication

In Loving Memory of Haley Jané

My companion, whose love and presence have filled my life with joy and comfort.

Chapter 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 from data, rather than being the ultimate objective themselves. Effect sizes aide in researcher's ability to draw meaningful inferences from data and therefore it is crucial 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.

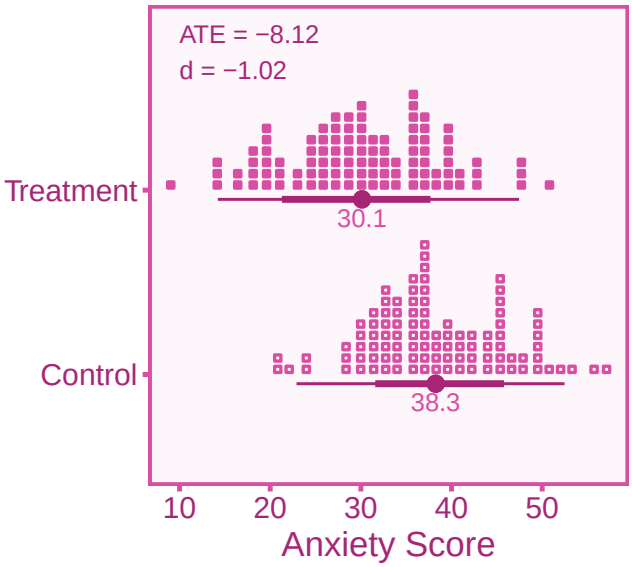


Figure 3.1: Simulated experimental data

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 3.2 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 3.2 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 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 (this notation will be used primarily 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 standard deviations (i.e., square root variance),

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

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 3.5, Equation 3.6, and Equation 3.7 into Equation 3.4 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 3.3,

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

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 chapter 4). We can express r similarly to Equation 3.1,

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

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,

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

In practice, the population correlation is unknown so the study correlation can be used instead (r) in the above formula as an estimate of the population correlation. Note that the sampling variance is the square of the standard error.

In the context of artifacts, if the observed correlation is biased relative to the true correlation, we can see model the observed population (ρ_o) correlation as a function of the true population correlation (ρ) and a artifact biasing factor, a,

$$\rho_o = a\rho.$$

The observed study (sample) correlation would then be defined as,

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

To obtain an unbiased estimate of the true population correlation, we can correct the correlation coefficient (r_c) by dividing by the biasing factor, a,

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

Note that the sampling error would also need to be adjusted and therefore it's sampling variance (σ_{ε_c}) would be corrected to be,

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

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 (e.g., $\text{df}_A = n_A - 1$),

$$\sigma_p = \sqrt{\frac{(n_A - 1)\sigma_A^2 + (n_B - 1)\sigma_B^2}{n_A + n_B - 2}}.$$

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 an artifactual biasing factor (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 and bias:

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

Thus the corrected standardized mean difference (d_c) and it's corresponding 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

Chapter 4

Small Samples

4.1 Introduction

The purpose of sample statistics such as correlations and standardized mean differences is to draw meaningful inferences about the population. However, effect size estimators such as Pearson's correlation coefficient and standardized mean difference are biased in small sample sizes. This small sample bias will impair our ability to draw accurate inferences about the population.

4.2 When Correcting alongside other Artifacts

The small sample bias should always be corrected for prior to applying any other artifact correction. It is independent of all other artifact corrections and therefore the corrected effect sizes in this section can be treated as the observed effect sizes (and observed sampling variance) in other sections.

4.3 Correcting for Small Sample Bias in Standardized Mean Differences

4.3.1 Defining our Estimand

Our quantity of interest is the population standardized mean difference, δ, between groups A and B on variable, y. We can model the relationship between the population standardized mean difference and our observed standardized mean difference (d_o) with,

$$d_o = a\delta + \varepsilon.$$

Where a is our small sample biasing factor and ε is our sampling error term. Ultimately, we can obtain an unbiased estimate of the population standardized mean difference by correcting the observed standardized mean difference as follows,

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

4.3.2 Artifact Correction for Small Sample Bias

As the sample size approaches infinity, Cohen's standardized mean difference estimator is an unbiased estimator of the population standardized mean difference (Hedges 1981; Cohen 2013). However, in small sample sizes Cohen's estimator is biased upward, that is, on average, it overestimates the population standardized mean difference. To see why this is the case, we can first define the standardized mean difference between group A and group B such that,

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

Where $\bar{y}_A$ and $\bar{y}_B$ are the observed arithmetic means of group A and group B, respectively. The raw difference between these two means will be unbiased, since the arithmetic means themselves are unbiased estimators at all sample sizes. However, the pooled standard deviation, σ_p, is biased when sample sizes are small. Statisticians ultimately had to choose whether the estimator of the standard deviation or the variance (the square of the standard deviation) would be unbiased. Since the variance has more utility in much of statistics, it was more important for the estimator of variance to be unbiased. Therefore the resulting bias in the standard deviation will bleed over into the equation for standardized mean differences. This bias can be visualized in the figure below. Notice that the sample standard deviation is under-estimated in small sample sizes (left plot), and the standardized mean difference is over-estimated in small sample sizes (right plot).

To obtain an unbiased estimate of the population standardized mean difference, we need to estimate a correction factor that can account for this bias. The small sample correction factor has been derived previously by Hedges (1989). When applied to a d value, it is common convention to refer to resulting corrected value as "Hedges' g", giving credit to the originator and to keep it similar in style to the conventionally termed "Cohen's d". However this convention will not be used here, instead we will denote it in this section as the small sample corrected d value (d_c). We can compute the artifact biasing factor, a, with the total sample size ($n = n_A + n_B$) and Gamma functions ($\Gamma(\cdot)$),

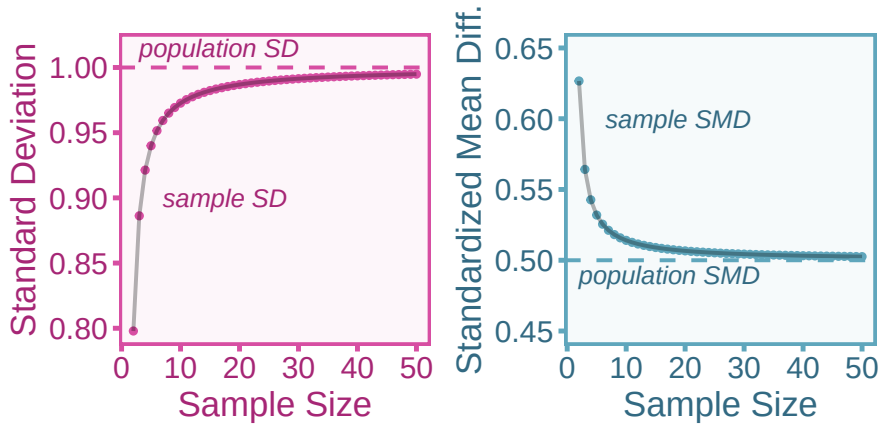


Figure 4.1: *Note.* dotted lines in both plots indicate the population value and the dots indicate the sample value at each sample size integer (n=1,2,3,...)

$$a = \frac{\Gamma\left(\frac{n-3}{2}\right)\sqrt{\frac{n-2}{2}}}{\Gamma\left(\frac{n-2}{2}\right)}.$$

Then to correct the observed standardized mean difference for small sample bias we can divide by a,

$$d_c = \frac{d_o}{a} = \frac{d_o}{\left[\frac{\Gamma\left(\frac{n-3}{2}\right)\sqrt{\frac{n-2}{2}}}{\Gamma\left(\frac{n-2}{2}\right)}\right]}.$$

It is important to point out that there is not a sample size threshold in which this correction does not apply, therefore common suggestions such as "small sample correction should be applied when n<30" is misguided since this correction can (and should) be applied at any sample size. Since the formula is quite complicated, there is also a simpler approximation of the this formula also given by Hedges (1989), $a \approx 1/\left(1 - \frac{3}{4n-3}\right)$. When this correction is made, we must also adjust the sampling variance (σ_{ε_o}),

$$\sigma^2_{\varepsilon_c} = \frac{\sigma^2_{\varepsilon_o}}{a^2} = \frac{\sigma^2_{\varepsilon_o}}{\left[\frac{\Gamma\left(\frac{n-3}{2}\right)\sqrt{\frac{n-2}{2}}}{\Gamma\left(\frac{n-2}{2}\right)}\right]^2}.$$

4.3.3 Correcting Standardized Mean Differences in R

To compute the correction in R, we can first simulate 20 data points, 10 in group A and 10 in group B. Then we can calculate the observed d value.

```
# set seed
set.seed(120)

# define parameters
nA <- 10
nB <- 10
n <- nA + nB
delta <- .5

# simulate data
yA <- rnorm(nA,delta,1)
yB <- rnorm(nB,0,1)

# calculate observed d value
SD_pooled = ((nA-1)*var(yA) + (nB-1)*var(yB)) / (nA+nB-2)
do = (mean(yA) - mean(yB)) / SD_pooled
```

Once we have obtained d_o, we can then correct it with the equations in the previous section.

```
# calculate bias factor
a <- (gamma((n-3)/2)*sqrt((n-2)/2)) / gamma((n-2)/2)

# correct d value
dc = do / a

# print results
rbind(paste0('Observed: do = ',round(do,3)),
      paste0('Corrected: dc = ',round(dc,3)))
```

```
     [,1]
[1,] "Observed: do = 0.889"
[2,] "Corrected: dc = 0.851"
```

You will notice that the small sample correction reduced the observed d value however it is still far away from the true value. This is simply due to the fact that there is a large amount of sampling error on top of the bias.

4.4 Correcting for Small Sample Bias in Correlations

4.4.1 Defining our Estimand

Our quantity of interest is the population correlation, ρ, between independent variable, x and dependent variable, y. We can model the relationship between the population correlation and our observed correlation (r_o) with,

$$r_o = a\rho + \varepsilon$$

Where a is our small sample biasing factor and ε is our sampling error term. Ultimately, we can obtain an unbiased estimate of the population correlation by correcting the observed correlation as follows,

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

4.4.2 Artifact Corrections

Correlation coefficients also are biased in small sample sizes (Olkin and Pratt 1958). As opposed to standardized mean differences, correlations are under-estimated, rather than over-estimated, in small samples. The bias is quite small, however we can apply a correction factor to obtain unbiased estimates of the

population correlation. Because the bias is so small and the exact formula is a hypergometric function using infinite power series, we will instead focus on the extremely close approximation provided in Olkin and Pratt (1958). Therefore, the biasing factor can be calculated such that,

$$a \approx \frac{1}{1 + \frac{1-r_o^2}{2(n-3)}}.$$

Then we can correct the point-estimate the sampling variance for small sample bias

$$r_c = r_o \times \left[1 + \frac{1 - r_o^2}{2(n-3)}\right]$$

$$\sigma^2_{\varepsilon_c} = \sigma^2_{\varepsilon_o} \times \left[1 + \frac{1 - r_o^2}{2(n-3)}\right]^2.$$

4.4.3 Correcting Correlations in R

To compute the correction in R, we can first simulate 10 correlated data points. Then we can calculate the observed correlation.

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

# set seed
set.seed(1)

# define parameters
n <- 10
rho <- .5

# simulate data
data = mvrnorm(n = 10,
               mu=c(0,0),
               Sigma = data.frame(x=c(1,rho),x=c(rho,1)))

# calculate observed d value
ro = cor(data[,1],data[,2])
```

Once we have obtained r_o, we can then correct it with the equations in the previous section.

```r
# calculate bias factor
a <- 1/(1 + (1-ro^2)/(2*(n-3)))

# correct d value
rc = ro / a

# print results
rbind(paste0('Observed: ro = ',round(ro,3)),
      paste0('Corrected: rc = ',round(rc,3)))
```

```
     [,1]
[1,] "Observed: ro = 0.247"
[2,] "Corrected: rc = 0.264"
```

You will notice that the small sample correction increases the observed correlation however it is still far away from the true value. This is simply due to the fact that there is a large amount of sampling error on top of the bias.

Chapter 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. Often the goal of measurement is to produce quantities (i.e., scores) that accurately reflect the measurand. However, quantities that do not reflect a *real* attribute can still have useful predictive value (e.g., socio-economic status). 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.

5.2 Reliability in True Score Theory

True score theory (or classical test theory) is a mathematical formalization of observed scores obtained from a measurement procedure. Observed scores, x_{im}, 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 assume that the covariance between errors in parallel measurements is zero, it becomes 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 nonsensical cases, the test's composite score retains a true score, but this true score does not mirror a tangible reality.

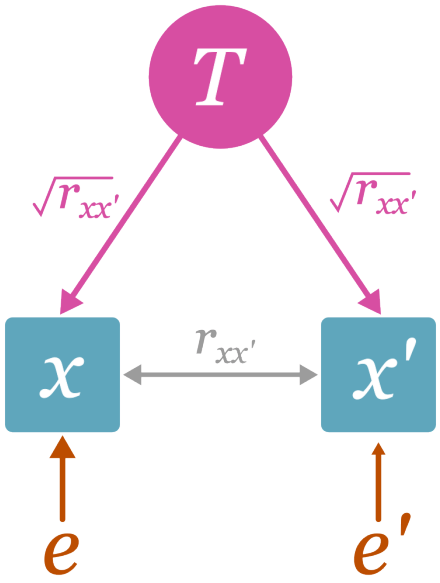


Figure 5.1: Structural diagram illustrating the relationship between true scores, observed scores, and error scores. The pink circle labeled t indicates the true scores, the blue squares labeled with x and x' represent observed scores on parallel measurements, and the red e denotes error. Correlations 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 the scores produced by 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 both of 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 items ($\aleph$) 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 item scores. Coefficient alpha only requires three parameters to calculate, the number of measurements (κ), the variances of each item ($\sigma^2_{\mathrm{i}_m}$), and the variance of the composite score (σ^2_x). Coefficient alpha will estimate the reliability of the composite observed score, x ($r_{xx'}$),

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

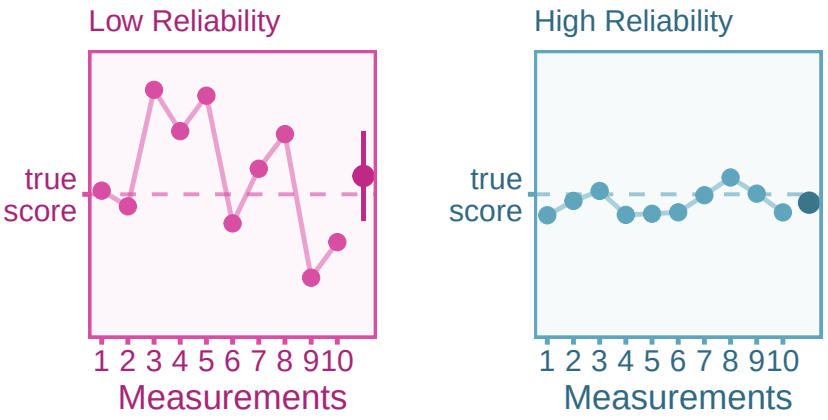


Figure 5.2: Figures showing the observed scores upon 10 repeated measurements and the composite observed score for a single person (the true score is denoted with the dashed line). The left panel shows 10 observed scores with a lot of variation (i.e., low reliability). The composite score (dark red dot 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 ($\bar{r}_{x_i x_j}$, where $i \neq j$). This formula is known as Spearman-Brown's prophecy,

$$_{\text{sb}}r_{xx'} = \frac{\kappa \bar{r}_{x_i x_j}}{1 + (\kappa - 1)\bar{r}_{x_i x_j}} \tag{5.4}$$

This can be simplified further if we have two observed item 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.

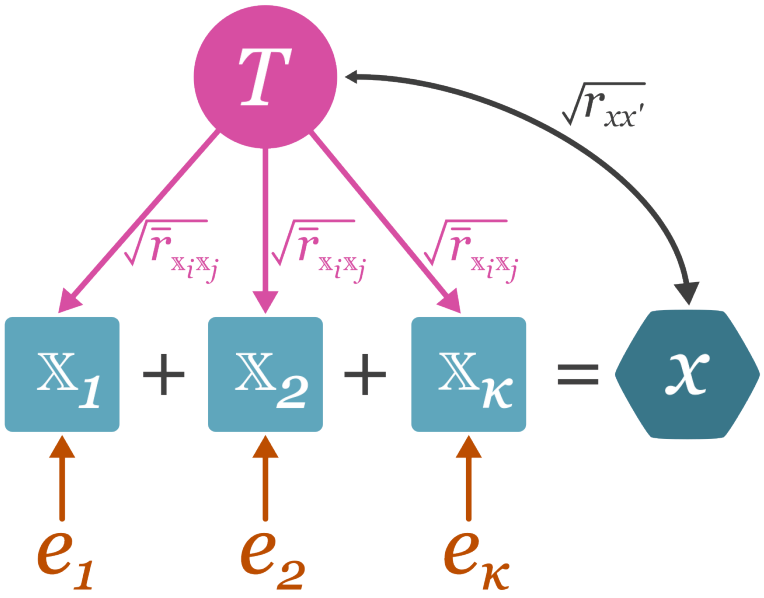


Figure 5.3: Structural model illustrating internal consistency. The pink circle labeled T indicates the true scores, the blue squares, $\mathbb{x}_{1...\kappa}$, represent the observed scores across multiple measurements, and the red e denotes error. The dark blue hexagon, x, indicates a composite score as a sum of the observed scores ($\mathbb{x}_{1...\kappa}$). 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.

```
#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 5.3.

```
# 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 5.4.

```
# 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
```

If we simplify even further, we can calculate the Split-Half reliability formula provided in Equation 5.5,

```
# 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,

```
# 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)
```

```r
# 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 5.4.5). 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`).

```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
```

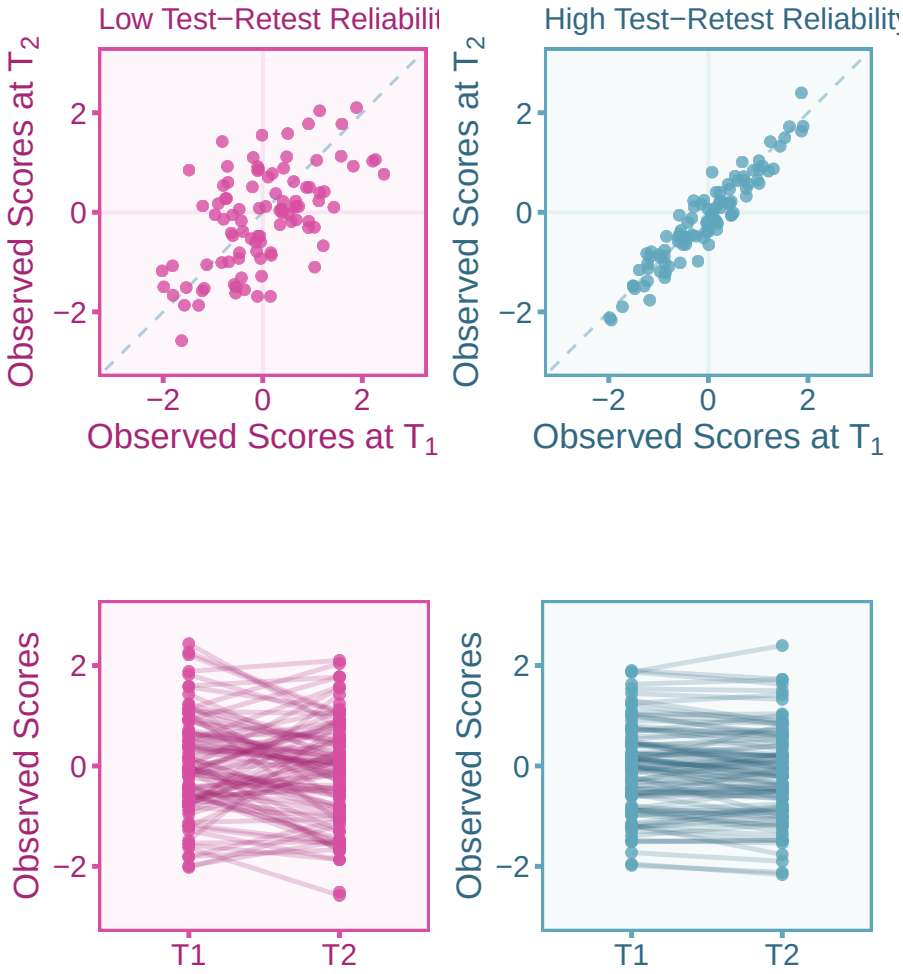


Figure 5.4: Illustrating test-retest reliability. Top-left and top-right panels show the correlation between observed scores at both time-points for a measure that has low and high reliability, respectively. Bottom-left and bottom-right panels show the within-person change from time-point 1 to time-point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

```
xTime2 = T_score + rnorm(n,0,.5)

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

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

```
[1] 0.755
```

```
# 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):

1. Random Response Error: Genuine randomness in responses. Examples include: motor errors and variation in response time.
2. 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.
3. 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.
4. 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: List of reliability coefficients and the sources of error they account for. The sources of error are denoted by the columns labeled 1-4, corresponding to each of the four sources of error.

Estimator	Description	1	2	3	4
Coefficient Alpha	Internal consistency coefficient for composite measures.	X		X	
Coefficient Omega	Internal consistency coefficient for composite measures with specified factor structure.	X		X	
Split-Half	Internal consistency coefficient for measurements that are split into two halves.	X		X	

Estimator	Description	1	2	3	4
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	1	2	3	4
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

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 observed study correlation is r_{xy}. The relationship between the observed correlation and the true population correlation is,

$$r_{xy} = a\rho_{TU} + \varepsilon.$$

Where a is the biasing factor. An unbiased estimate of the true score population correlation (ρ_{TU}) can then be calculated by dividing the obsereved study correlation by the biasing factor,

$$r_c = \frac{r_{xy}}{a}.$$

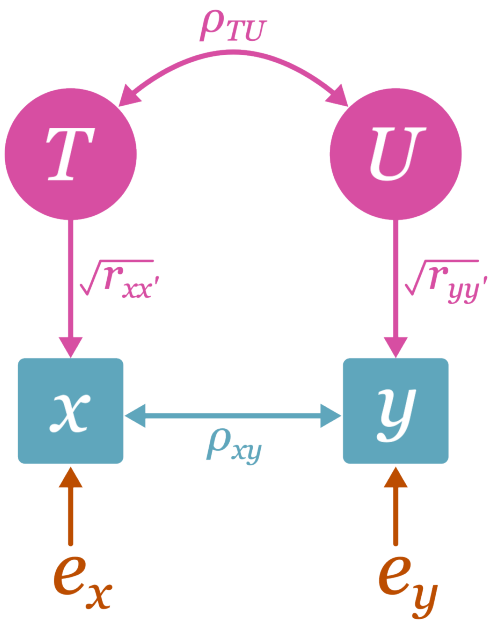


Figure 5.5: This figure shows the relationship between the true scores, observed scores, and error scores. The true score correlation is denoted by the curved arrow connecting the circles indicating true score variables, T and U.

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 5.2), 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 5.6), 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}),

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

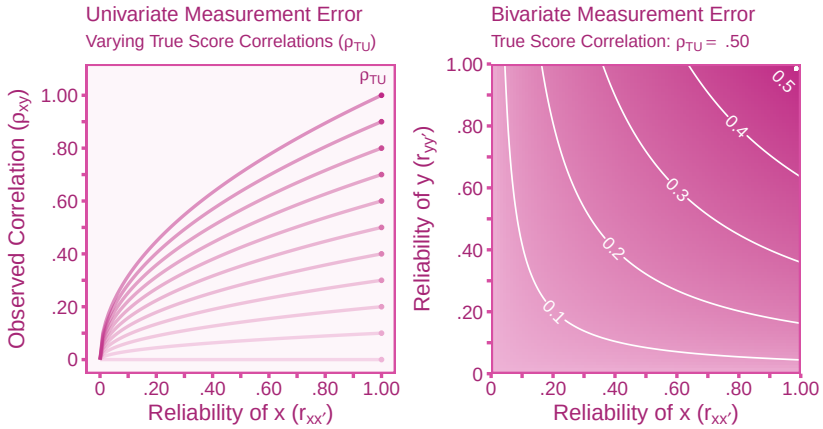


Figure 5.6: Visualizing the attenuation of observed correlation (ρ_{xy}) due to measurement error. The left panel shows a situation where only one variable (x) has measurement error. The observed correlation increases as a function of the true correlation ρ_{TU} (darker lines indicate a higher true score 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, $r_{xx'} = r_{yy'} = 1$).

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,

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

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
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`).

```
# 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`).

```
# 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 5.5.2

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

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

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

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

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

```
# 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 dependent variables, y, therefore the observed score study standardized mean difference is d_y. The relationship between the population true score standardized mean difference (δ_U) can be related to the observed study standardized mean difference with the following formulation:

$$d_y = a\delta_U + \varepsilon.$$

Where a is the artifactual bias induced by measurement error and ε denotes sampling error. To obtain an unbiased estimate of δ_U, we can correct the observed standardized mean difference by dividing by a,

$$d_c = \frac{d_y}{a}.$$

5.6.2 Artifact Correction for Unreliability

We can calculate the standardized mean difference 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 rather than the means. Therefore, we can express the observed standardized mean difference as,

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

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 replace the observed score standard deviations with the attenuated true score standard deviation in Equation 5.10,

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

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 5.10,

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

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 study standardized mean

difference (d_y) is a function of the observed population standardized mean (δ_y), artifact bias (a), and sampling error (ε),

$$d_y = \delta_y + \varepsilon_o.$$

replacing the observed population standardized mean difference, 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 standardized mean difference (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_o) 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).

$$r_o = \frac{d_y}{\sqrt{\frac{1}{p(1-p)} + d_y^2}}.$$

Then correct r_o for the total sample reliability,

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

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 can be done for the sampling variance as well, but instead we can put it all into one equation,

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

5.6.3 Correcting Standardized Mean Differences 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 standardized mean difference with the true score standardized mean difference (`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 standardized mean difference 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 standardized mean difference for measurement error variance using the equations in Section 5.6.2.

```
# 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
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.

```
# 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 the study's test, with a re-arrangement of the Spearman-Brown prophecy formula,

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

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

Chapter 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. We can also assume 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 individuals 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}. 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 group assignments (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 5, 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 Group 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)

## Strategy 2: calculate from contingency table values
numerator <- con_table['A','A']*con_table['B','B'] - con_table['A','B']*con_table['B
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

## Strategy 3: calculate from chi-square test
```

```r
chi2 <- as.numeric(chisq.test(con_table)$statistic)
rgg <- sqrt(chi2/n)

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

```
[1] 0.64
```

6.5 Correcting for Group Misclassification in Standardized Mean Difference

6.5.1 Defining our Estimand

Our quantity of interest is the true population standardized mean difference, δ, between members of group A and group B on the scores of the dependent variable, y. However, the observed sample standardized mean difference (d_o) is estimating the difference between individuals who are assigned group to A and group B. Error in the assignment of groups (i.e., group misclassification) will bias the observed correlation by a factor we will label as a. The relationship between the observed study standardized mean difference and the true population standardized mean difference

$$d_o = a\delta + \varepsilon.$$

Where ε denotes the sampling error. Therefore an unbiased (corrected) estimate of the true population correlation would be:

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

6.5.2 Artifact Correction for Standardized Mean Difference

The 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 $\bar{y}_A^{\text{true}}$ and $\bar{y}_B^{\text{true}}$, 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,

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

$$\bar{y}_A^{\text{obs}} = \left(\frac{n_{BB}}{n_{BB} + n_{AB}}\right) \bar{y}_B^{\text{true}} + \left(\frac{n_{AB}}{n_{BB} + n_{AB}}\right) \bar{y}_A^{\text{true}}.$$

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, the figure below shows the distributions for groups A and B without any misclassification. In this case, there is no attenuation of the standardized mean difference.

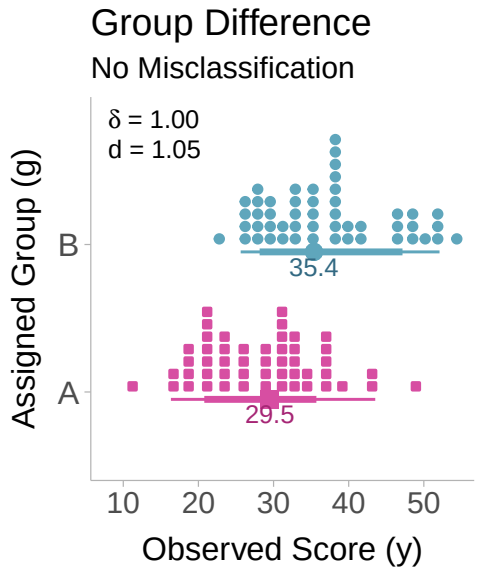


Figure 6.1: Distributions of scores without misclassification. True mean difference and observed mean differ only due to sampling error.

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

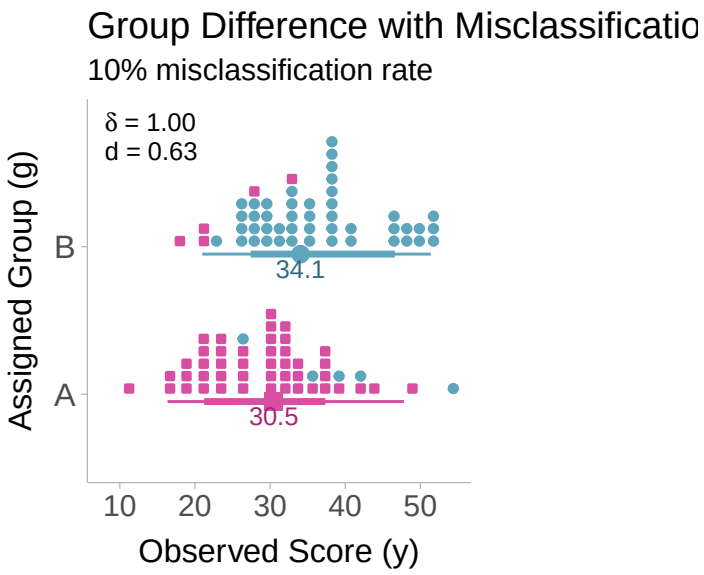


Figure 6.2: Distributions of scores with a 10% misclassification rate. Observed standardized mean differences are biased toward the null (i.e., SMD = 0). Note that a few members of group A (red squares) are within observed group B and vice versa (indicative of misclassification).

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 $(1-p)$,

$$\rho = \frac{\delta}{\sqrt{\frac{1}{p(1-p)} - \delta^2}}.$$

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 observed standardized mean difference:

$$\delta_o = \frac{\rho\sqrt{r_{gg'}}}{\sqrt{p(p-1)\left(1 - r_{gg'}\rho^2\right)}}.$$

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. 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:

$$r_o = \frac{d_o}{\sqrt{\frac{1}{p(1-p)} - d_o^2}}.$$

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

$$r_c = \frac{r_o}{\sqrt{r_{gg'}}}.$$

If we also wanted to correct for measurement error in the dependent variable y, we can use the correction formula used in the chapter 4

$$r_c = \frac{r_o}{\sqrt{r_{gg'}}\sqrt{r_{yy'}}}.$$

Now we can convert the corrected point-biserial correlation into a corrected standardized mean difference (d_c). 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:

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

Instead of doing the converting, correcting, and back-converting for the sampling variance, we can correct the sampling variance in a single step,

$$\sigma^2_{\varepsilon_c} = \frac{\sigma^2_{\varepsilon_o}\left(\frac{r_c}{r_o}\right)^2}{\left(1+d_o^2 p[1-p]\right)^2\left(d_o^2+\frac{1}{p(1-p)}\right)p^*(1-p^*)(1-r_c^2)^3}.$$

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

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

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 have an observed standardized mean difference of $d_o = 0.40$ 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 $1 - p = 1 - p^* = .60$.

```
do = .40
rgg = .70
nA = 40
nB = 60
```

The `psychmeta` package (Dahlke and Wiernik 2019) 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 = do,
          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 0.492 -0.000766     1.03 100        67.9
```

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`).

To calculate the corrected standardized mean difference, we can also use the equations in the previous section.

```
## correct using three step procedure
# step 1: convert d to r
ro <- do / sqrt( 1/(p*(1-p)) + do^2)

# step 2: correct r
rc <- ro / sqrt(rgg)

# step 3: convert r to d
dc <- rc / sqrt( p*(1-p)*(1-rc^2) )

## correct sampling variance
# step 1: compute sampling variance for do
var_e_o <- (nA+nB)/(nA*nB) + do^2 / (2*(nA+nB))

# step 2: adjust sampling variance for correction
var_e_c <- (var_e_o * (rc/ro)^2) / ((1 + do^2)^3 * p*(1-p) * (1-rc^2)^3)

# print corrected values
rbind(paste0('corrected: do = ',round(do,3)),
      paste0('corrected: var = ', round(var_e_o,3)),
      paste0('corrected: dc = ',round(dc,3)),
```

```
        paste0('corrected: var = ', round(var_e_c,3)))
```

```
     [,1]
[1,] "corrected: do = 0.4"
[2,] "corrected: var = 0.042"
[3,] "corrected: dc = 0.482"
[4,] "corrected: var = 0.191"
```

Chapter 7

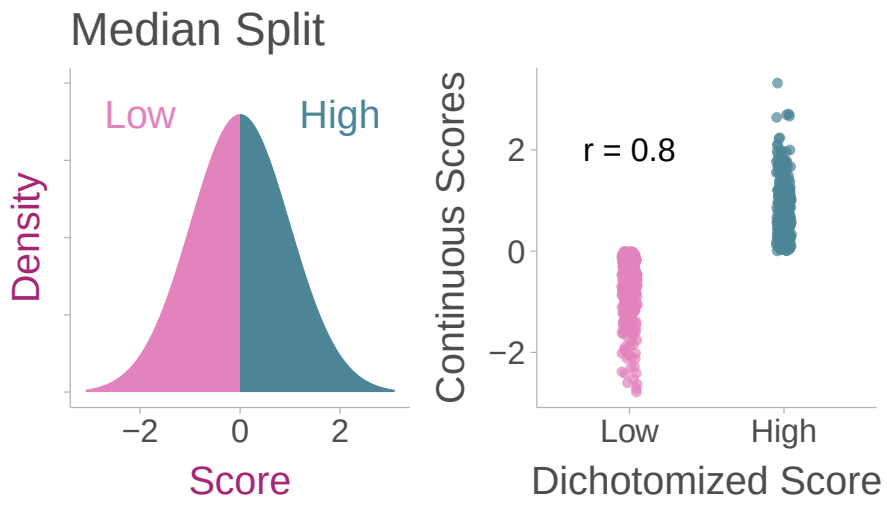
Artificial Dichotomization

7.1 Introduction

Researchers occasionally split 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). The obvious solution to this problem is to simply not dichotomize variables, however if only summary data is available to us, then we may not have this luxury. Dichotomization can also be practical in some instances. For example, 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 their underlying continuous scores. To demonstrate this, we can draw a sample of scores and then split the data into high and low scorers and then calculate 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 (assume x is standardized such that Mean = 0 and Variance = 1). 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_x and C_y is the cut-score where the split took place for each variable. 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. In both cases, dichotomization will have a biasing effect on the observed study correlation coefficient. The observed study correlation can be modeled as a function of the population correlation on continuous scores (ρ_{xy}; i.e., the estimand), an artifact biasing factor a, and sampling error, ε,

$$r_{x_{‡}y_{‡}} = a\rho_{xy} + \varepsilon.$$

Where unbiased estimate of ρ_{xy} can be calculated by dividing the observed correlation by the biasing factor, a,

$$r_c = \frac{r_{x_{‡}y_{‡}}}{a}.$$

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 artifact biasing factor, a,

$$\rho_{x_{‡}y} = a\rho_{xy}.$$

The first step in estimating the bias in the correlation is to first identify the cut-point, C_x, of the 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

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

Where n indicates the sample size. 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, $r_c = \frac{r_{x_{\ddagger}y}}{a}$. 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 ($\sigma^2_{\varepsilon_c}$) must also be adjusted using the artifact biasing factor (a), and the observed sampling variance ($\sigma^2_{\varepsilon_o}$),

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

The Bivariate Case

In some cases, both independent and dependent variables are dichotomized inducing measurement error in both variables. A pearson correlation calculated on these two dichotomized variables would be equal to the phi coefficient (or Matthew's correlation coefficient) and we can denote it with our notation for

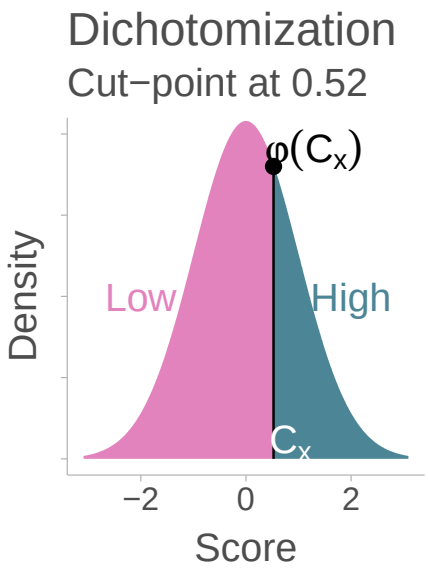
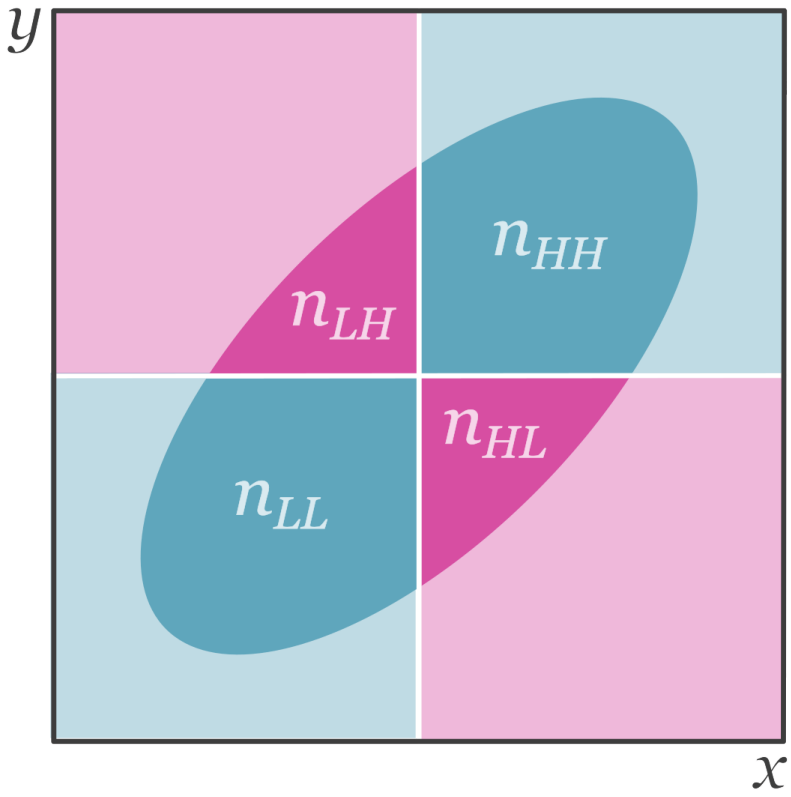


Figure 7.1: This figure shows the continuous distribution of scores split into a high scoring and low scoring group. The symbols relate to the notation in the equations shown in this section.

dichotomized variables, $r_{x_\ddagger y_\ddagger}$. The data can be structured in a contingency table.

	$x_\ddagger$ = Low	$x_\ddagger$ = High
$y_\ddagger$ = Low	n_{LL}	n_{HL}
$y_\ddagger$ = High	n_{LH}	n_{HH}

We can also visualize how this contingency table relates to an underlying continuous bivariate normal distribution. The ellipse indicates the correlation between x and y. If x and y are positively correlated then we should more individuals populating the high-high and low-low cells rather than the high-low and low-high cells.



The proper correction is to calculate the tetrachoric correlation coefficient. The tetrachoric correlation is specifically meant for the relationship between two dichotomous scores that represent continuous underlying normal distribution. To calculate the tetrachoric correlation coefficient, the contingency table or the odds ratio must be available. To estimate the correlation of continuous variables (r_{xy}) we can approximate the tetrachoric correlation with the following formulation,

$$r_c = \cos\left(\frac{\pi}{1+\sqrt{\frac{n_{HH}n_{LL}}{n_{HL}n_{LH}}}}\right) \tag{7.2}$$

If the contingency table is not provided, but the odds ratio ($OR = \frac{n_{HH}n_{LL}}{n_{HL}n_{LH}}$) is, then we can calculate r_{xy} in terms of the odds ratio,

$$r_c = \cos\left(\frac{\pi}{1+\sqrt{OR}}\right)$$

The sampling variance must be calculated from the contingency table as well. A sampling variance approximation can be obtained from Pearson (1913), however due to the complexity of the formulation and because it is only an approximation, instead I recommend that researchers use a bootstrap procedure to obtain approximate confidence intervals. To do this, we must resample the contingency table (>10,000 iterations), calculating the tetrachoric correlation using Equation 7.2 upon each iteration. Once you obtain a tetrachoric correlation from each iteration, the standard deviation of all correlations can be used as an estimate of the standard error and the square of the standard error is the sampling variance.

Unfortunately, studies may not report the full contingency table. Instead they may report summary statistics like a chi-squared value or a phi coefficient (i.e., the pearson correlation on binary variables). If the χ^2-statistic is reported, we can first convert that to a phi coefficient by using,

$$r_{x_\ddagger y_\ddagger} = \sqrt{\frac{\chi^2}{n}}$$

Where n is the total sample size. From the phi coefficient, we can estimate the correlation of the continuous variables with a formula similar to Equation 7.1,

$$r_c = \frac{r_{x_\ddagger y_\ddagger}}{\left[\frac{\varphi(C_x)}{\sqrt{p_x(1-p_x)}}\right]\left[\frac{\varphi(C_y)}{\sqrt{p_y(1-p_y)}}\right]} \tag{7.3}$$

This formula was introduced by J. Hunter and Schmidt (1990) and is a rough approximation of the correlation between the continuous independent and dependent variables (r_{xy}). The corresponding sampling variance of the corrected correlation coefficient is,

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

7.3.3 Correcting Correlations for Dichotomization in R

To correct correlations for dichotomization in R, we can use the `correct_r_dich` in the `psychmeta` package. First let us simulate a data set of 100 individuals that are then dichotomized on both x and y. We will use a median (50-50) split for each dichotomization.

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

# set seed
set.seed(343)

# define parameters
rho = .50
n <- 500

# simulate data
data <- mvrnorm(n = n,
                mu = c(0,0),
                Sigma = data.frame(x=c(1,rho),
                                   y=c(rho,1)),
                empirical = TRUE)

# obtain scores
x = data[,1]
y = data[,2]

# dichotomize data
x_dich <- as.numeric(x>0)
y_dich <- as.numeric(y>0)

# compute observed correlation
ro <- cor(x_dich,y_dich)

# compute observed sampling variance
var_e_o <- (1 - ro^2)^2 / (n-1)

# compute split proportions
px = sum(x_dich==1)/length(x_dich)
py = sum(y_dich==1)/length(y_dich)
```

We can correct the correlation using base R. In order to correct for dichotomization, we can use the equations presented in the previous section.

```r
# get cut-point
Cy <- qnorm(py)
Cx <- qnorm(px)
```

```r
# calculate attenuation factors
ax <- dnorm(Cx)/sqrt(px*(1-px))
ay <- dnorm(Cy)/sqrt(py*(1-py))

# correct r
rc <- ro / (ax*ay)

# adjust standard error for rc
var_e_c <- var_e_o * (rc/ro)^2

# print results
rbind(paste0('Observed: ro = ',round(ro,3)),
      paste0('Observed: var = ',round(var_e_o,5)),
      paste0('Corrected: rc = ',round(rc,3)),
      paste0('Corrected var = ',round(var_e_c,5)))
```

```
     [,1]
[1,] "Observed: ro = 0.321"
[2,] "Observed: var = 0.00161"
[3,] "Corrected: rc = 0.504"
[4,] "Corrected var = 0.00398"
```

We can also correct for dichotomization using the `correct_r_dich` function in the `psychmeta` package (Dahlke and Wiernik 2019).

```r
correct_r_dich(ro,px=px,py=py,n=n)
```

```
  r_corrected var_e_corrected    n_adj
1   0.5035871     0.003982818 140.8791
```

7.4 Correcting Standardized Mean Differences for Artificial Dichotomization

7.4.1 Defining our estimand

We would like to know the group difference between scores of a naturally continuous variable. Our estimand can thus be defined as the population standardized mean difference between groups A and B on continuous scores of the dependent variable (y), δ_y. Where dichotomized scores can be defined as

$$y_{A\ddagger} = \begin{cases} 1, & \text{if } y_A > C_y \\ 0, & \text{if } y_A < C_y \end{cases}$$

$$y_{B\ddagger} = \begin{cases} 1, & \text{if } y_B > C_y \\ 0, & \text{if } y_B < C_y \end{cases}$$

In studies of group differences, since the independent variable is already dichotomous, the only dichotomization that can occur is on the dependent variable. The observed study standardized mean difference can be modeled as a function of the population standardized mean difference on continuous scores (δ_y; i.e., the estimand), an artifact biasing factor a, and sampling error, ε,

$$d_{y_\ddagger} = a\delta_y + \varepsilon.$$

Where unbiased estimate of δ_y can be calculated by dividing the observed standardized mean difference by the biasing factor, a,

$$d_c = \frac{d_{y_\ddagger}}{a}.$$

7.4.2 Artifact Correction for Standardized Mean Differences

The simplest way to correct for dichotomization in a standardized mean difference is to first convert the observed d value of the dichotomized dependent variable to a correlation coefficient. When converting to a correlation coefficient, it's important to note the binary nature of both variables, leading us to estimate the phi coefficient rather than the point-biserial correlation that we would be estimating if the dependent variable was continuous. To calculate the phi coefficient from a d value we can use the proportion of group membership in group A or group B (p; it does not matter which one is chosen, as long as it is consistent for every instance of p),

$$r_o = \frac{d_{y_\ddagger}}{\sqrt{d_{y_\ddagger}^2 + \frac{1}{p(1-p)}}}$$

We can then correct the phi coefficient similar to how we correct the point-biserial correlation in Section 7.3.2,

$$r_c = \frac{r_o}{\left[\frac{\varphi(C_y)}{\sqrt{p_y(1-p_y)}}\right]}.$$

Then we can convert the corrected correlation back into a standardized mean difference,

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

Where d_c is our corrected correlation. The observed sampling variance (σ_{ε_o}) must also be corrected using the same three step procedure. For simplicity, we will consolidate this into one formula,

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

Obtaining Standardized Mean Difference from Odds Ratio

In most cases, difference in dichotomized outcomes between two groups is unlikely to be reported as a standardized mean difference, instead it will be more commonly reported as an odds ratio (OR). The odds ratio is the difference in the ratio of high and low scorers of group A and group B, such that,

$$OR = \frac{\left(\frac{n_{A,\text{high}}}{n_{A,\text{low}}}\right)}{\left(\frac{n_{B,\text{high}}}{n_{B,\text{low}}}\right)}$$

Where $n_{Group,\text{high/low}}$ is the corresponding sample size. The odds ratio is asymmetric about 1 (null), but we can make it symmetric by log transforming it ($\log(OR)$). The sampling variance of the log odds ratio $\sigma^2_{\log(OR)}$ can be defined as,

$$\sigma^2_{\log(OR)} = \frac{1}{n_{A,\text{low}}} + \frac{1}{n_{A,\text{high}}} + \frac{1}{n_{B,\text{low}}} + \frac{1}{n_{B,\text{high}}}$$

The equation above requires the full contingency table to compute. From there we can use the cox-logit method to convert the odds ratio to a d value (**cox1989?**; **haddock1998?**). Them method is quite simple as it just divides the log odds ratio by 1.65,

$$d_c = \frac{\log(OR)}{1.65}$$

and the corresponding sampling variance of the d value is,

$$\sigma^2_{\varepsilon_c} = \frac{\sigma^2_{\log(OR)}}{1.65^2}.$$

7.4.3 Correcting Standardized Mean Differences in R

To correct standardized mean differences for dichotomization in R . At the moment the `psychmeta` package does not have a `correct_d_dich` function. In order to correct for dichotomization, we can use the three step process from Section 7.4.2. Lets simulate data based on a true standardized mean difference of $\delta_y = 0.50$ and y scores that have been cut in half (i.e., a 50-50 median split).

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

# define parameters
delta = .50
nA <- nB <-  1000
n <- nA + nB
p <- nA / n

# simulate data
yA <- rnorm(nA,.50,1)
yB <- rnorm(nB,0,1)
y_dich <- as.numeric(c(yA,yB)>0)
yA_dich <- y_dich[1:nA]
yB_dich <- y_dich[(nA+1):n]

# calculate observed standardized mean difference
do = (mean(yA_dich) - mean(yB_dich)) / sqrt((var(yA_dich) + var(yB_dich))/2)

# calculate observed sampling variance
var_e_o = (n / (nA * nB) + do^2 / (2 * n))

# get cut-point
py = sum(y_dich==1)/length(y_dich)
Cy = qnorm(py)

# calculate attenuation factor of y
ay <- dnorm(Cy)/sqrt(py*(1-py)) # attenuation factor for dichotomization in y

# convert d to r
ro <- do / sqrt(do^2 + (1 / (p*(1-p))))

# correct r
rc <- ro / ay

# convert r to d
dc <- rc / sqrt(p*(1-p)*(1-rc^2))
```

```
# correct sampling variance
var_e_c <- (var_e_o * (rc/ro)^2) / ((1+do^2*p*(1-p))^3 * (1-rc^2)^3)

# print results
rbind(paste0('Observed: do = ',round(do,3)),
      paste0('Observed: var = ',round(var_e_o,6)),
      paste0('Corrected: dc = ',round(dc,3)),
      paste0('Corrected: var = ',round(var_e_c,6)))
```

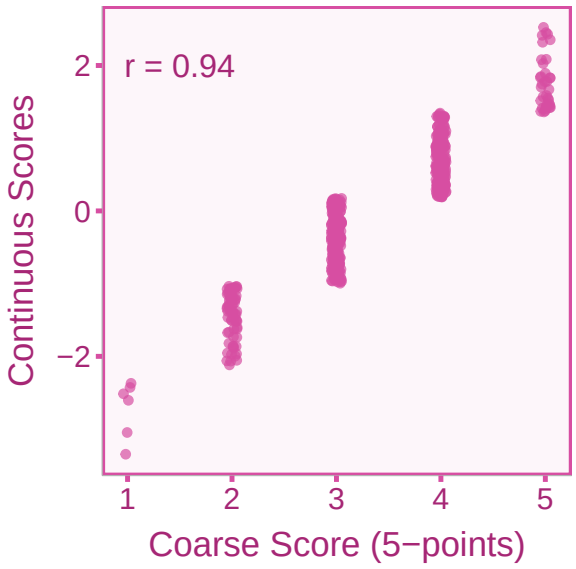
```
     [,1]
[1,] "Observed: do = 0.43"
[2,] "Observed: var = 0.002046"
[3,] "Corrected: dc = 0.553"
[4,] "Corrected: var = 0.003586"
```

Chapter 8

Scale Coarseness

8.1 Introduction

Scale coarseness describes a situation where a variable that is naturally continuous (e.g., happiness) is binned into discrete values (e.g., happiness measured on a scale of 1-10). This situation is quite common in the social and psychological sciences where Likert items or dichotomous yes/no responses are aggregated to form a coarse total score for a naturally continuous construct. When coarseness is present, measurement error is introduced into the observed scores and those scores lose information. Unlike dichotomization, coarseness is an artifact that occurs due to the design of the study rather than during the analysis phase (Aguinis, Pierce, and Culpepper 2009). Particularly, dichotomization occurs after scores are obtained (e.g., splitting a group into high scorers and low scorers), whereas coarseness occurs as an artifact of the measurement instrument itself. The primary issue with coarseness is that it limits the set of possible values a score can be which introduces error when the variable is naturally continuous. Lets visualize how this occurs by sampling 500 data points from a normal distribution and binning the scores into 5 equal-interval scale points. Now the figure below shows the relationship between the coarse scores and the true underlying continuous scores.



Notice that the correlation between coarse and continuous scores is not perfect, indicating that the coarse scores do not perfectly capture the underlying continuous scores.

8.2 Correcting for Coarseness in Correlations

8.2.1 Defining our Estimand

Our quantity of interest is the population correlation, ρ, between continuous independent variable, x, and continuous dependent variable, y. We can model the relationship between the observed sample correlation on coarse scores and the true population correlation,

$$r_o = a\rho + \varepsilon.$$

Where a is our coarseness biasing factor and ε is our sampling error term. Ultimately, we can obtain an unbiased estimate of the true (continuous) population correlation by correcting the observed correlation as follows,

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

8.2.2 Artifact Correction for Coarseness

To correct the correlation between coarse scores for x and y, we need to know the correlation between coarse scores and their underlying continuous scores.

These correlations between coarse scores and their underlying continuous scores do not have a specific closed form equation as of now (Aguinis, Pierce, and Culpepper 2009). To find this we need to make a couple of assumptions:

1) The shape of the underlying distribution (i.e., normal or uniform).
2) The intervals between scale-points are equal.

Peters and Voorhis (1940) constructed a table based on these assumptions that is also reported more recently by Aguinis, Pierce, and Culpepper (2009). The table below shows the correction factor for a given number of scale points and underlying distribution shape (either normal or uniform). The correction factor is equal to the correlation between coarse scores and the underlying continuous normal (or uniform) distribution. For the normal distribution correction, its been shown that even in cases of extreme skew, these correction factors perform well Wylie (1976).

Scale Points	Correction Factor (normal)	Correction Factor (uniform)
2	.816	.866
3	.859	.943
4	.916	.968
5	.943	.980
6	.960	.986
7	.970	.990
8	.977	.992
9	.982	.994
10	.985	.995
11	.988	.996
12	.990	.997
13	.991	.997
14	.992	.997
15	.994	.998

These correction factors can be applied similarly to the correction formula for reliability (see chapter 5),

$$r_c = \frac{r_o}{a_x a_y}.$$

Where a_x and a_y are the appropriate correction factors from the table for x and y. We must also adjust the sampling variance as well,

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

8.2.3 Correcting Correlations in R

To correct scale coarseness in R, we can use the table in Section 8.2.2. Lets start by simulating a coarse data (5 scale-points for x and 7 scale-points for y) set of 500 individuals with a true population correlation of $\rho = .50$.

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

# load packages
# install.packages('MASS')
library(MASS)

# simulate data
n <- 500
rho <- .50
data <- mvrnorm(n = n,
                mu = c(0,0),
                Sigma = data.frame(x=c(1,rho),y=c(rho,1)),
                empirical = TRUE)

# obtain simulated continuous data
x <- data[,1]
y <- data[,2]

# get coarse data
x_coarse <- as.numeric(cut(x,breaks=5))
y_coarse <- as.numeric(cut(y,breaks=7))
```

Now we can calculate the observed standardized mean difference and apply the correction factor.

```
# calculate observed d value
ro <- cor(x_coarse,y_coarse)

# correct observed do using table
ax <- .943
ay <- .970

# correct observed d value
rc <- ro/(ax*ay)

# print results
rbind(paste0('Observed: ro = ', round(ro,3)),
      paste0('Corrected: rc = ', round(rc,3)))
```

```
     [,1]
[1,] "Observed: ro = 0.457"
[2,] "Corrected: rc = 0.5"
```

8.3 Correcting for Coarseness in *d* values

8.3.1 Defining our Estimand

Our quantity of interest is the population standardized mean difference, δ, between groups A and B on variable, y. We can model the relationship between the observed sample standardized mean difference on coarse scores and the true population standardized mean difference,

$$d_o = a\delta + \varepsilon.$$

Where a is our coarseness biasing factor and ε is our sampling error term. Ultimately, we can obtain an unbiased estimate of the true (continuous) population standardized mean difference by correcting the observed standardized mean difference as follows,

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

8.3.2 Artifact Correction for Coarseness

To correct a standardized mean difference for coarseness in the outcome variable, y, we can use the correction factors from the table in Section 8.2.2,

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

Where a is the appropriate correction factor from the table. We must also adjust the sampling variance as well,

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

8.3.3 Correcting Standardized Mean Differences in R

To correct scale coarseness in R, we can use the table in Section 8.2.2. Lets start by simulating a coarse data (5 scale-points) set of 1000 individuals with a true population standardized mean difference of $\delta = .50$.

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

# simulate data
nA <- nB <- 1000
yA <- rnorm(nA,.5,1)
yB <- rnorm(nB,0,1)

# concatenate data
group <- c(rep('A',nA),rep('B',nB))
y <- c(yA,yB)

# get coarse data
y_coarse <- as.numeric(cut(y,breaks=5))
```

Now we can calculate the observed standardized mean difference and apply the correction factor.

```
# calculate observed d value
do <- ( mean(y_coarse[group=='A']) - mean(y_coarse[group=='B']) ) / sqrt((var(y_coar

# correct observed do using table
ay <- .943

# correct observed d value
dc <- do/ay

# print results
rbind(paste0('Observed: do = ', round(do,3)),
      paste0('Corrected: dc = ', round(dc,3)))
```

```
     [,1]
[1,] "Observed: do = 0.499"
[2,] "Corrected: dc = 0.529"
```

Chapter 9

Direct Selection

9.1 Introduction

Direct selection occurs when subjects are explicitly selected based on some eligibility criterion on the variables of interest (rather than a third variable). Range restriction is a form of selection bias that describes a situation where there is less variation in our sample then there is in the population. Whereas range enhancement indicates that there is *more* variation in a sample then there is in the population. Direct range restriction/enhancement biases the variances and effect size estimates.

9.2 An Applied Example of Direct Range Restriction

Imagine a tech company that wants to assess the correlation between years of experience and programming proficiency for their software engineers. They have two primary divisions: Division A and Division B. Division A primarily hires entry-level software engineers, with less than 3 years of experience. Division B, on the other hand, hires experienced software engineers with more than 3 years of experience. The company decides to conduct a study to assess the correlation between years of experience and programming proficiency. However, they only collect data from Division A due to logistical reasons, assuming that the relationship found there would be represent the entire company. In this scenario, direct range restriction occurs because the sample used for the study (Division A) represents a narrow range of years of experience (0-3 years) compared to the broader range present in the entire company (0+ years). Consequently, the standard deviation will be smaller in the sample then it would if we had sampled from the entire company. As we will see in later sections of this chapter, the observed correlation between years of experience and programming proficiency

would be attenuated, underestimating the true correlation.

9.3 Indexing Range Restriction with the *u*-ratio

The distribution of scores in the unrestricted pool of individuals will exhibit a greater (or lesser) degree of variability compared to the sample that has been selected into the study. Therefore the standard deviation of scores in the unrestricted population (σ_x) will differ from that of the selected (restricted/enhanced) sample (σ_{x_S}). To index the difference between the two standard deviations, we can calculate the u-ratio Wiernik and Dahlke (2020). The u-ratio is the ratio between the standard deviations of the selected sample to the unrestricted sample such that,

$$u_x = \frac{\sigma_{x_S}}{\sigma_x}$$

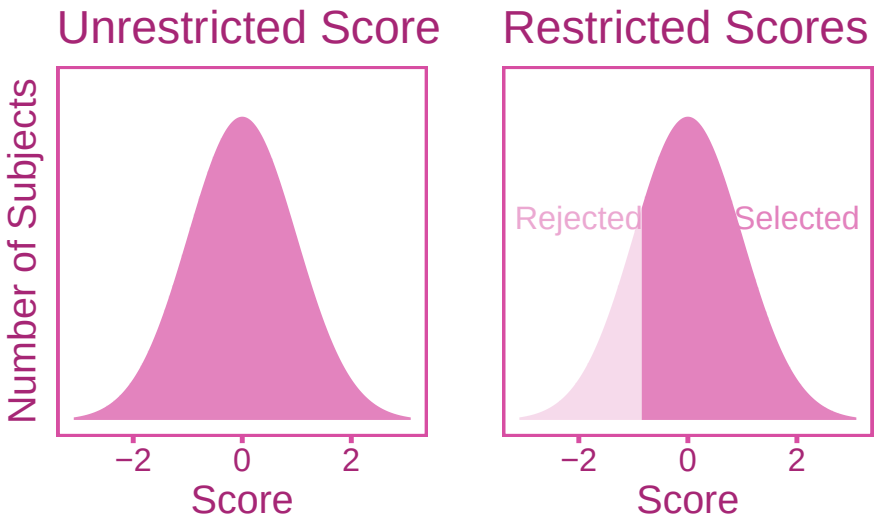
The u-ratio in cases of range restriction will exist in the interval (0–1). Conversely, when the u-ratio is greater than 1 it is indicative of range enhancement. The unrestricted standard deviation is often quite difficult to acquire since we do not usually have access to the unrestricted group. However, the unrestricted standard deviation can be estimated from some reference study that has been conducted on the unrestricted group. This often comes in the form of standardization samples or norm samples (obtained from test manuals) if the unrestricted group is the general population. For example, the distribution full-scale IQ scores derived from the Wechsler Adult Intelligence Test has a standard deviation of 15 in the US population (Wechsler 2008). We can use this estimate as the standard deviation for the unrestricted population. Lets say we select a sample from members of Mensa, a high IQ society, who are specifically selected on the basis high IQ scores. If the standard deviation of Mensa members is 5, then the u-ratio would be,

$$u_x = \frac{\sigma_{x_S}}{\sigma_x} = \frac{5}{15} = .33\,.$$

However it is not always the case that an estimate of the unrestricted standard deviation is readily available. Therefore if the reliability coefficient from the unrestricted and selected sample can be used to estimate the u-ratio,

$$u_x = \sqrt{\frac{1 - r_{xx'}}{1 - r_{xx'_S}}}.$$

Where $r_{xx'_S}$ and $r_{xx'}$ are the reliability estimates within the selected and unrestricted groups respectively.



9.4 Correcting Correlations for Direct Range Restriction

9.4.1 Defining our Estimand

For our study we want to estimate the population correlation of the unrestricted scores of the independent (x) and dependent variable (y). We can denote this correlation as ρ_{xy_S}. Within a study that suffers from range restriction, the correlation (r_{xy_S}) will be biased relative to our estimand, ρ_{xy}. This bias can be denoted by a such that,

$$r_{xy} = a\rho_{xy} + \varepsilon.$$

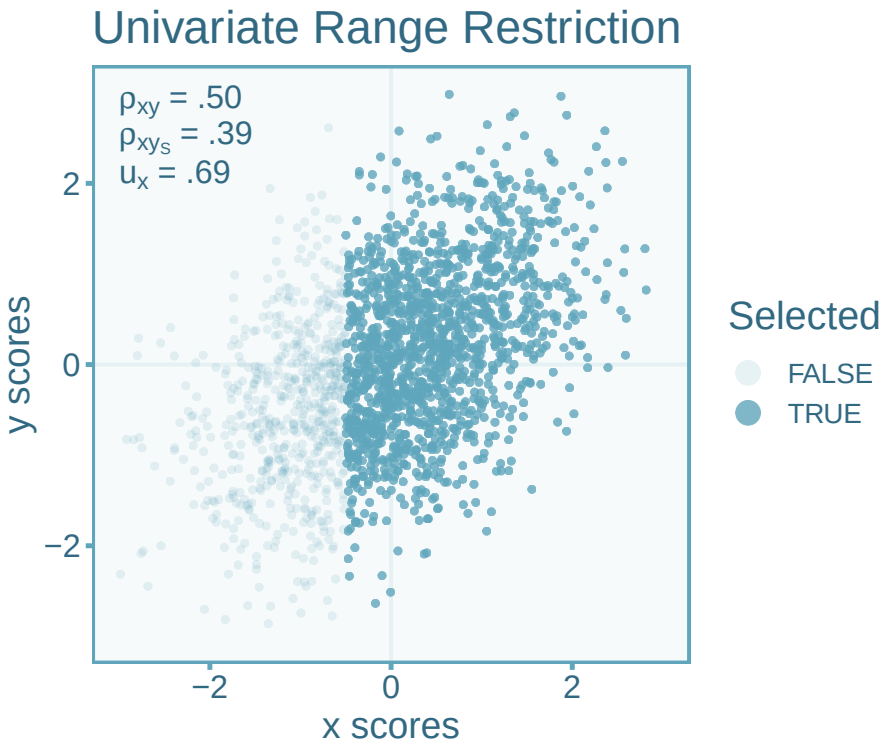
Therefore an unbiased estimate of the unrestricted population correlation would be

$$r_c = \frac{r_{xy_S}}{a}.$$

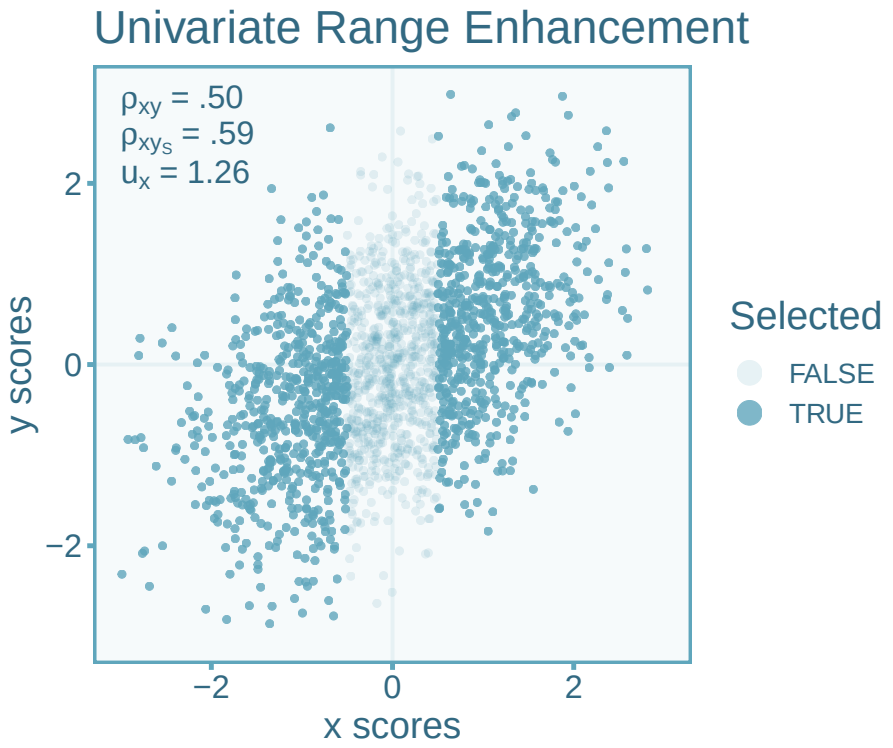
9.4.2 Artifact Correction for Correlations

The Univariate Case

Range restriction (or enhancement) in either the independent or dependent variable will induce bias into the correlation coefficient. Let us consider a case where just the independent variable is restricted (or enhanced) such that $u_x \neq 1$, but the dependent variable is not restricted (directly). It is important to note, that if there is direct selection one of the two variables, then there will be indirect selection in the other variable too if the two are correlated. This would suggest that if $u_x \neq 1$ and $\rho_{xy} \neq 0$ then $u_y \neq 1$. Lets visualize the correlation between independent (x) and dependent (y) variables under this range restriction by only selecting individuals above some cut off. The scores of individuals that have been selected will show less variance than the entire pool of individuals. Specifically, the scenario below shows a u-ratio of about 0.69 in the independent variable. We see in the figure that the correlation in the restricted scores (ρ_{xy_S}) is attenuated relative to the unrestricted (true) correlation (ρ_{xy}).



We can also visualize what happens to the correlation when the range is enhanced. Enhancement can be accomplished by selecting individuals at the ends of the distribution (Taylor and Griess 1976). In the visualization below, we see an opposite effect on the correlation, that is, an over-estimate of the unrestricted correlation rather than an attenuation like we see under range restriction. The scenario below has a u-ratio of about 1.26 in the independent variable.



It starts to become apparent that if $u_x > 1$ (i.e., $\sigma_x > \sigma_{x_S}$) the observed correlation over-estimates the true, unrestricted correlation and under-estimates the unrestricted correlation when $u_x < 1$ (i.e., $\sigma_x < \sigma_{x_S}$, Sackett and Yang 2000).

A bias correction formula for univariate direct range restriction was first developed by Pearson (1903) and provided more recently by J. E. Hunter and Schmidt (1990). To correct for the systematic bias in correlations, we can use the u-ratio of the independent variable such that,

$$r_c = \frac{r_{xy_S}}{u_x \sqrt{1 + r_{xy_S}^2 \left[\frac{1}{u_x^2} - 1\right]}}. \tag{9.1}$$

Where the sampling variance of the corrected correlation is,

$$\sigma_{\varepsilon_c}^2 = \sigma_{\varepsilon_o}^2 \left(\frac{r_c}{r_{xy_S}}\right)^2. \tag{9.2}$$

If we want to also correct for measurement error in both samples, then we can also incorporate measurement into these equations. Note that the following equations will incorporate the reliability within the selected sample ($r_{xx'_S}$) rather than the unrestricted population ($r_{xx'}$). If the reliability coefficient comes from the unrestricted population, then we can estimate the selected (restricted or enhanced) sample reliability with the corresponding u-ratio,

$$r_{xx'_S} = 1 - \frac{1 - r_{xx'}}{u_x^2}. \tag{9.3}$$

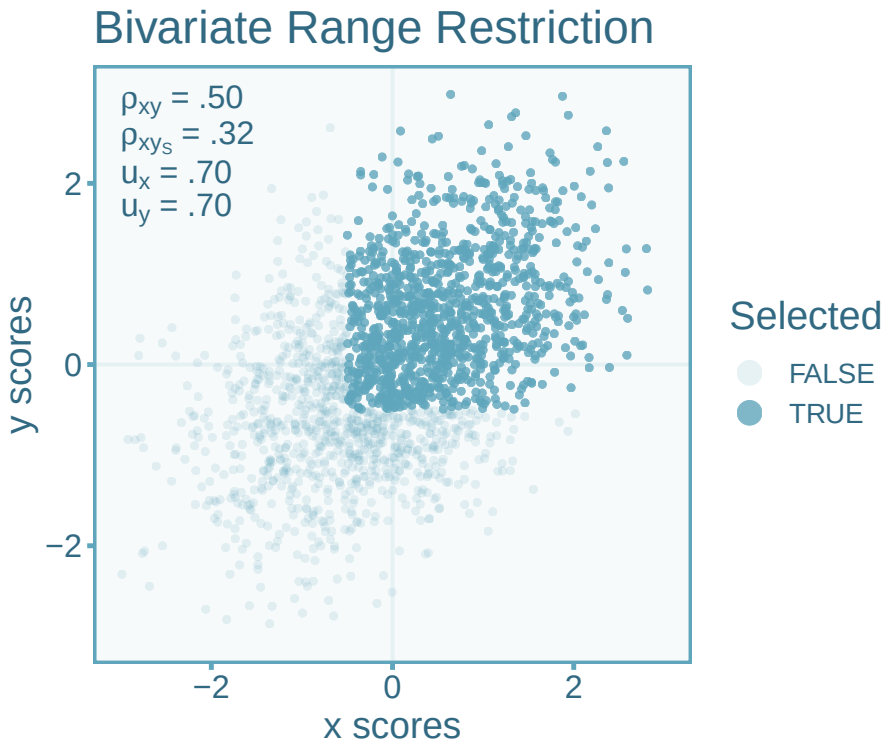
Then we can use the calculated reliability in the following equation to obtain an unbiased estimate of the true score unrestricted population correlation ($r_c = \hat{\rho}_{TU}$),

$$r_c = \frac{r_{xy_S}}{u_x\sqrt{1 - u_x^2(1 - r_{xx'_S})}\sqrt{r_{yy'_S} + r_{xy_S}^2\left[\frac{1}{u_x^2} - 1\right]}}.$$

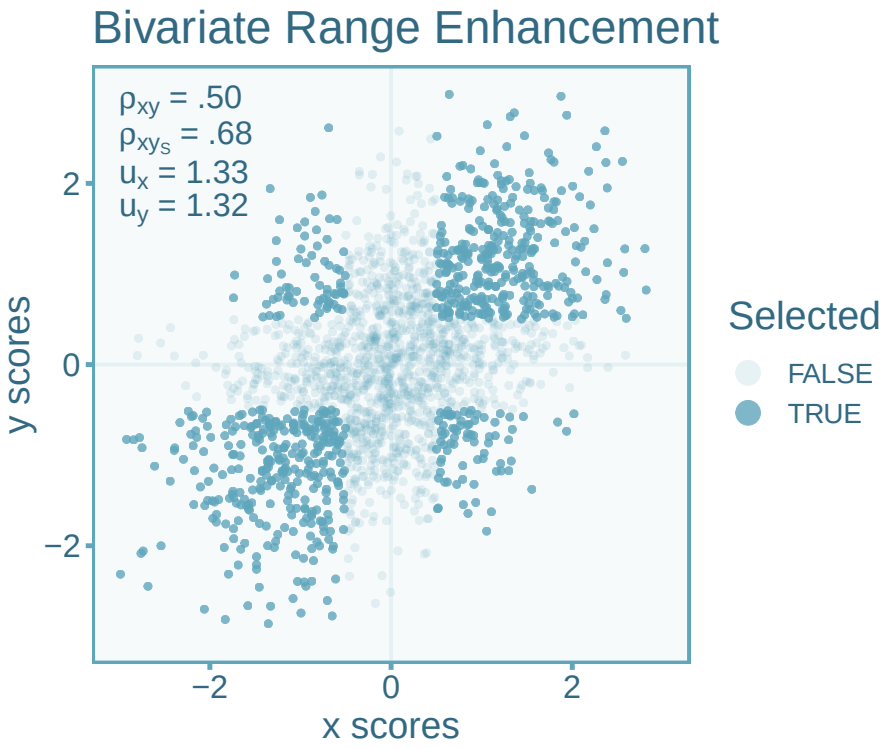
We can use the same equation as Equation 10.4 to calculate the corrected sampling variance.

The Bivariate Case

Bivariate direct range restriction/enhancement occurs when the variability in both independent and dependent variables within the selected sample is less than or greater than the variability in the unrestricted population. Let us consider a case where just the independent variable is restricted (or enhanced) such that $u_x \neq 1$ and $u_y \neq 1$. Like we showed for the univariate case, let's visualize the correlation between independent (x) and dependent (y) variables under range restriction by only selecting individuals above some cut off point for both x and y. The scores of individuals that have been selected will show less variance than the entire pool of individuals. Specifically, the scenario below shows a u-ratio of about 0.70 in the independent variable and dependent variables. We see in the figure that the correlation in the restricted sample (ρ_{xy_S}) is attenuated relative to the unrestricted (true) correlation (ρ_{xy}).



Likewise let's visualize what happens to the correlation when the range is enhanced. Enhancement in both variables can be accomplished by selecting individuals at the ends of the distribution of x and y. In the visualization below, we observe an over-estimation of observed correlation relative to the unrestricted correlation. The scenario below has a u-ratio of about 1.32 in both the independent variable and dependent variable.



A bias correction formula for bivariate range restriction is much more complicated than the univariate formulation. This is due to the fact that there is inter-dependence between the correlation, the u-ratio of x, and the u-ratio of y. For instance, if x and y are positively correlated and if there is direct range restriction in x this will also restrict the variability in y even if there is no range restriction in y. To break down the correction formula into simpler parts, let us first define a factor we will denote with ψ,

$$\psi = \frac{u_x u_y \left(r_{xy_S}^2 - 1\right)}{2 r_{xy_S}} \tag{9.4}$$

This factor contains all the parameters needed to correct the correlation coefficient under direct selection (r_{xy_S}). Then we can plug it into the formula

$$r_c = \psi + \text{sign}\left[r_{xy_S}\right] \sqrt{\psi^2 + 1} \tag{9.5}$$

Where the sampling variance of the corrected correlation is,

$$\sigma^2_{\varepsilon_c} = \sigma^2_{\varepsilon_o} \left(\frac{r_c}{r_{xy_S}} \right)^2. \tag{9.6}$$

Now let us also consider a case where measurement error is present in both independent and dependent variables. Note again that the following equations will incorporate the reliability within the selected sample ($r_{xx'_S}$) rather than the unrestricted population ($r_{xx'}$). Then we can use the restricted/enhanced (selected) sample reliability and the u-ratios in the following equation to obtain an unbiased estimate of the true score unrestricted population correlation ($r_c = \hat{\rho}_{TU}$),

$$r_c = \frac{\psi + \text{sign}\left[r_{xy_S}\right]\sqrt{\psi^2 + 1}}{\sqrt{1 - u_x^2(1 - r_{xx'_S})}\sqrt{1 - u_y^2(1 - r_{yy'_S})}}.$$

If the reliability coefficient comes from the unrestricted population, the formula simplifies to,

$$r_c = \frac{\psi + \text{sign}\left[r_{xy_S}\right]\sqrt{\psi^2 + 1}}{\sqrt{r_{xx'}}\sqrt{r_{yy'}}}.$$

We can use the same equation as Equation 9.6 to calculate the corrected sampling variance.

9.4.3 Correcting Correlations in R

To correct correlations for range restriction we can start by simulating data from the the `mvrnorm` function in the `MASS` package. Lets first simulate 200 data points.

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

# set seed
set.seed(343)

# define parameters
rho <- .50
n <- 200

# sample data from a bivariate normal distribution
data <- mvrnorm(n = n,
                mu = c(0,0),
```

```
                Sigma = data.frame(x = c(1,rho),
                                   y = c(rho,1)),
                empirical = TRUE)

# obtain unrestricted scores
x <- data[,1]
y <- data[,2]
```

Univariate Direct Range Restriction

We can start with univariate direct range restriction by selecting only on the independent variable. We will select only the values above the mean.

```
# obtain scores when x > Mean(x)
selected <- x > mean(x)
xS <- x[selected]
yS <- y[selected]

# calculate correlation between unrestricted and restricted scores
rxy <- cor(x,y) # unrestricted
rxyS <- cor(xS,yS) # restricted

# print results
rbind(paste0('unrestricted: rxy = ',round(rxy,2)),
      paste0('restricted: rxyS = ',round(rxyS,2))
      )
```

```
     [,1]
[1,] "unrestricted: rxy = 0.5"
[2,] "restricted: rxyS = 0.32"
```

As expected, we observe an attenuation of the correlation under range restriction. Now lets calculate the u-ratios for both variables. Remember that even though we only selected on x, we should expect the variability in y in the restricted sample to also be smaller than the unrestricted sample when x and y are correlated.

```
# calculate u-ratios
ux <- sd(xS)/sd(x)
uy <- sd(yS)/sd(y)

# print results
rbind(paste0('ux = ',round(ux,2)),
      paste0('uy = ',round(uy,2))
```

```
      )
```

```
     [,1]
[1,] "ux = 0.59"
[2,] "uy = 0.86"
```

As anticipated, not only is u_x below 1 indicating range restriction, but also u_y is slightly below 1 since x and y covary. Now we can apply the correction for univariate direct range restriction by hand from Equation 10.1 and Equation 10.1.

```
# correct the restricted correlation
rc <- rxyS / (ux * sqrt(1 + rxyS^2 * (1/ux^2-1)) )

# acquire sample size from
n <- length(xS)

# calculate the observed correlation sampling variance
var_e_o <- (1-rxyS^2)^2 / (n-1)

# correct sampling variance
var_e_c <- var_e_o * (rc/rxyS)^2

# print results
rbind(paste0('corrected cor: r = ',round(rc,2)),
      paste0('corrected var: var_e = ',round(var_e_c,3))
      )
```

```
     [,1]
[1,] "corrected cor: r = 0.49"
[2,] "corrected var: var_e = 0.02"
```

The correction formula produced a very close estimate of the true population correlation ($r_c = .49$ vs $\rho_{xy} = .50$). Lets also correct the correlation using the `correct_r` function in the psychmeta package, `psychmeta` (Dahlke and Wiernik 2019).

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

# correct the restricted correlation for univariate direct range restriction
correct_r(rxyi = rxyS,
          correction = 'uvdrr_x',  # uvdrr_x = univariate direct range restriction i
          ux = ux,
          n = n)
```

```
Correlations Corrected for Measurement Error and Univariate Direct Range Restriction:
--------------------------------------------------------------------------------------
  value CI_LL_95 CI_UL_95  n n_effective
1 0.492    0.209    0.685 97        40.8
```

We can see that the correction made by the `correct_r` function provides identical results to the one done by hand.

9.4.3.1 Bivariate Direct Range Restriction

For bivariate direct range restriction we can select values above the mean in both independent and dependent variables.

```
# obtain scores when x > Mean(x) and y > Mean(y)
selected <- x > mean(x) & y > mean(y)
xS <- x[selected]
yS <- y[selected]

# calculate correlation between unrestricted and restricted scores
rxy <- cor(x,y) # unrestricted
rxyS <- cor(xS,yS) # restricted

# print results
rbind(paste0('unresticted: rxy = ',round(rxy,2)),
      paste0('restricted: rxyS = ',round(rxyS,2))
      )
```

```
     [,1]
[1,] "unresticted: rxy = 0.5"
[2,] "restricted: rxyS = 0.29"
```

Notice that there is even more attenuation in the selected correlation coefficient than there was in the univariate case. Now we can correct for bivariate range restriction by hand using Equation 9.4, Equation 9.5, Equation 9.6.

```
# calculate the factor, psi
psi <- ux*uy*(rxyS^2-1) / (2*rxyS)

# calculate the corrected correlation using psi
rc <- psi + sign(rxyS)*sqrt(psi^2 + 1)

# acquire sample size from
n <- length(xS)

# calculate the observed correlation sampling variance
var_e_o <- (1-rxyS^2)^2 / n
```

```
# correct sampling variance
var_e_c <- var_e_o * (rc/rxyS)^2

# print results
rbind(paste0('corrected cor: r = ',round(rc,2)),
      paste0('corrected var: var_e = ',round(var_e_c,3))
      )
```

```
     [,1]
[1,] "corrected cor: r = 0.48"
[2,] "corrected var: var_e = 0.036"
```

Again, we see that the corrected correlation closely resembles the unrestricted correlation ($r_c = .48$ vs $\rho_{xy} = .50$). lets

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

# correct the restricted correlation for univariate direct range restriction
correct_r(rxyi = rxyS,
          correction = 'bvdrr',  # bvdrr = bivariate direct range restriction
          ux = ux,
          uy = uy,
          n = n)
```

```
Correlations Corrected for Measurement Error and Bivariate Direct Range Restriction:
--------------------------------------------------------------------------------
  value CI_LL_95 CI_UL_95  n n_effective
1  0.48   0.0943    0.689 64        17.3
```

We can see that the correction exactly reflects the correction done by hand.

9.5 Correcting Standardized Mean Differences for Direct Range Restriction

9.5.1 Defining our Estimand

The quantity of interest is the unrestricted population standardized mean difference between groups A and B on variable, y. We can denote this standardized mean difference as δ. Within a study that suffers from direct selection, the observed standardized mean difference (d_S) will be biased relative to our estimand, δ. This bias can be denoted by a such that,

$$d_S = a\delta + \varepsilon.$$

Therefore an unbiased estimate of the unrestricted population standardized mean difference would be

$$d_c = \frac{d_S}{a}.$$

Artifact Correction for Standardized Mean Difference

9.5.1.1 Selection on the Continuous Variable

To correct for direct selection on the continuous variable, we can first convert the observed standardized mean difference (d_S) to a point-biserial correlation (r_o). Converting d_S to r_o can be done by using the observed proportion of individuals in group A (or B), p,

$$r_o = \frac{d_o}{\sqrt{\frac{1}{p(1-p)} - d_o^2}}.$$

We can then correct the point-biserial correlation for univariate direct range restriction using the formulas in **?@sec-cor-DDR**. Note that if you want to correct for measurement error as well, replace $r_{xx'}$ with $r_{gg'}$ (i.e., group classification reliability; see chapter on *group misclassification*) whenever you are working with standardized mean differences. Once we obtained the corrected correlation, r_c, we can convert back to a standardized mean difference, we need to use an adjusted group proportions, p^*:

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

Where p^* is

$$p^* = \frac{1 - \sqrt{1 - 4p(1-p)\left[1 + \frac{r_o^2}{u_x^2} - r_o^2\right]}}{2}$$

The adjusted proportion, p^*, can also be estimated from the proportion of individuals in the unrestricted population (e.g., the proportion of men vs women in the general population). This adjustment is necessary in order to account for indirect selection in the grouping variable when $d \neq 0$. This is similar to the situation described in Section 9.4.2, where one variable suffers from direct range restriction and any variable that is correlated with it, will suffer from indirect selection. The corresponding corrected sampling error $\sigma_{\varepsilon_c d}$ can also be computed with the observed and adjusted proportions such that,

$$\sigma^2_{\varepsilon_c} = \frac{\sigma^2_{\varepsilon_o}\left(\frac{r_c}{r_o}\right)^2}{\left(1 + d_o^2 p[1-p]\right)^2\left(d_o^2 + \frac{1}{p(1-p)}\right)p^*(1-p^*)(1-r_c^2)^3}.$$

9.5.2 Correcting Standardized Mean Differences in R

To correct for standardized mean differences in R, let us start by simulating data for groups, A and B with a true standardized mean difference of $\delta = 0.50$.

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

# define parameters
delta <- .50
nA <- nB <- 500
n <- nA+nB

# sample data from a normal distribution for each group
yA <- rnorm(nA,delta,1)
yB <- rnorm(nB,0,1)

# obtain unrestricted data
y = c(yA,yB)
group = c(rep('A',nA),rep('B',nB))
```

Now lets select individuals above the mean (y=0) and calculate the total sample u_y.

```
# select individuals above the mean
selected <- y > 0
yS <- y[selected]
groupS <- group[selected]

# calculate u-ratio
uy <- sd(yS)/sd(y)

# print results
paste0('uy = ',round(uy,3))
```

```
[1] "uy = 0.613"
```

We can then calculate the observed standardized mean difference and correct it for univariate range restriction using the equations from the previous section.

```
# calculate the observed standardized mean difference
MeanDiff <- mean(yS[groupS=='A']) - mean(yS[groupS=='B'])
SD_pooled <- sqrt((var(yS[groupS=='A']) + var(yS[groupS=='B']))/2)
dS <- MeanDiff / SD_pooled
p <- mean(groupS=='A')

## correct dS with three step procedure
# step 1: convert d to r
ro = dS / sqrt( 1/(p*(1-p)) + dS^2 )

# step 2: correct r
rc = ro / (uy * sqrt(1 + ro^2 * (1/uy^2-1)) )

# step 3: obtain unrestricted proportion and then convert r to d
p_true <- .5
dc <- rc / sqrt( p_true*(1-p_true)*(1-rc^2) )


# print results
rbind(paste0('observed: do = ',round(dS,2)),
      paste0('corrected: dc = ',round(dc,3)))
```

```
     [,1]
[1,] "observed: do = 0.27"
[2,] "corrected: dc = 0.426"
```

We can also use the `correct_d` function in the `psychmeta` package to correct for range restriction.

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

# correct using correct_d
correct_d(d=dS,
          correction = "uvdrr_y",
          pi = p,
          pa = p_true,
          uy = uy)
```

```
d Values Corrected for Measurement Error and Univariate Direct Range Restriction:
--------------------------------------------------------------------------------
   dgyi  dgpi  dGyi  dGpi  dgya  dgpa  dGya  dGpa
1 0.262 0.262 0.262 0.262 0.426 0.426 0.426 0.426
```

In this case, columns ending with the letter `a` (indicating unrestricted sample)

are the corrected values we are looking for.

Chapter 10

Indirect Selection

10.1 Introduction

Indirect range restriction/enhancement occurs when selection of sample participants is based on a variable that is correlated with the variables of interest. If the selector Whereas range enhancement indicates that there is *more* variation in a sample then there is in the population. Direct range restriction/enhancement (as opposed to indirect range restriction) is when selection into the sample is based on the variable(s) of interest (i.e., the independent and/or dependent variable). This selection into the sample will either restrict or enhance the variation in the variable, thus causing

10.2 An Applied Example of Indirect Range Restriction

Imagine a research team is conducting a study on academic motivation among college students using a survey that includes various questions related to academic engagement, goal orientation, and effort investment. The researchers administer the survey to a large sample of students across different universities. However, during the data cleaning process, the researchers identify a subset of respondents who exhibited signs of inattentiveness and carelessness in their responses. These signs include straight-lining questions (e.g., consistently selecting the same response option without reading the questions) or responding randomly without considering the content of the questions. Recognizing that inattentive or careless responding can distort the measurement of academic motivation, the researchers decide to exclude these individuals from the analysis. The rationale is to ensure that the data collected represents genuine responses and validly measures academic motivation. The unintended consequence of this decision is indirect range restriction. By removing inattentive and careless

responders, who likely also have lower academic motivation and engagement, from the dataset, the observed range of academic motivation scores is reduced. The excluded individuals, who may have had lower academic motivation scores, are not accounted for in the analysis, resulting in an underestimation of the variability of academic motivation relative to the population.

10.3 Indexing Range Restriction with the *u*-ratio

The distribution of scores in the unrestricted pool of individuals will exhibit a greater (or lesser) degree of variability compared to the sample that has been selected into the study. Therefore the standard deviation of scores in the unrestricted population (σ_x) will differ from that of the selected (restricted/enhanced) sample (σ_{x_S}). To index the difference between the two standard deviations, we can calculate the u-ratio Wiernik and Dahlke (2020). The u-ratio is the ratio between the standard deviations of the selected sample to the unrestricted sample such that,

$$u_x = \frac{\sigma_{x_S}}{\sigma_x}$$

The u-ratio in cases of range restriction will exist in the interval (0–1). Conversely, when the u-ratio is greater than 1 it is indicative of range enhancement. The unrestricted standard deviation is often quite difficult to acquire since we do not usually have access to the unrestricted group. However, the unrestricted standard deviation can be estimated from some reference study that has been conducted on the unrestricted group. This often comes in the form of standardization samples or norm samples (obtained from test manuals) if the unrestricted group is the general population. For example, the distribution full-scale IQ scores derived from the Wechsler Adult Intelligence Test has a standard deviation of 15 in the US population (Wechsler 2008). We can use this estimate as the standard deviation for the unrestricted population. Lets say we select a sample from members of Harvard students, who tend to have higher IQs than the general population (this is due to the fact that selection criterion, such as GPA and SAT scores are positively correlated with IQ). If the standard deviation of IQ in Harvard students is 10, then the u-ratio would be,

$$u = \frac{\sigma_{x_S}}{\sigma_x} = \frac{10}{15} = .67$$

However it is not always the case that an estimate of the unrestricted standard deviation is readily available. Therefore if the reliability coefficient from the unrestricted and selected sample can be used to estimate the u-ratio,

$$u_x = \sqrt{\frac{1 - r_{xx'}}{1 - r_{xx'_S}}}$$

Where $r_{xx'_S}$ and $r_{xx'}$ are the reliability estimates within the selected and unrestricted groups respectively. In the context of indirect range restriction, the selection does not occur directly on x (or y), instead it occurs on a selector variable, z. Therefore we can see how the u-ratio of the selector variable (u_z) relates to the u-ratio of x,

$$u_x = \sqrt{\rho_{xz}^2 u_z^2 - \rho_{xz}^2 + 1}$$

The formulation above is also dependent on the correlation between the selector, z, and x. If the correlation between z and x is $\rho_{xz} = 0$, then you will notice that $u_x = u_z$. Notice that a correlation of $\rho_{xz} = 1$ would also simplify to a direct range restriction problem. A correlation of $\rho_{xz} = 0$, would effectively have no selection effect (i.e., restriction nor enhancement) since the equation would simplify to $u_x = 1$.

10.4 Correcting Correlations for Direct Range Restriction

10.4.1 Defining our Estimand

For our study we want to estimate the population correlation of the unrestricted scores of the independent (x) and dependent variable (y). We can denote this correlation as ρ_{xy}. The population correlation within the selected sample can be denoted as ρ_{xy_S}. Within a study sample that suffers from indirect selection (and sampling error), the study correlation (r_{xy_S}) will be biased relative to our estimand, ρ_{xy}. This bias can be denoted by a such that,

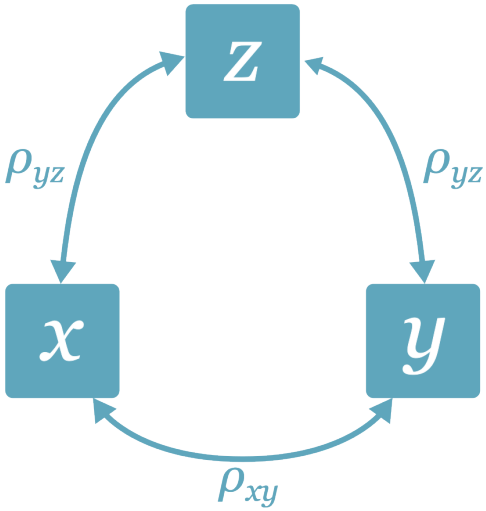
$$r_{xy} = a\rho_{xy_S} + \varepsilon$$

Therefore an unbiased estimate of the unrestricted population correlation would be

$$r_c = \frac{r_{xy_S}}{a}.$$

Note that we may also want to correct for measurement on top of range restriction. In this case, using the true score model for x ($x = T + e_x$) and y ($y = U + e_y$), we may want to estimate the unrestricted correlation between true scores (ρ_{TU}). The figure below shows the relationship between three variables: the selector (z), the independent variable (x), and the dependent variable

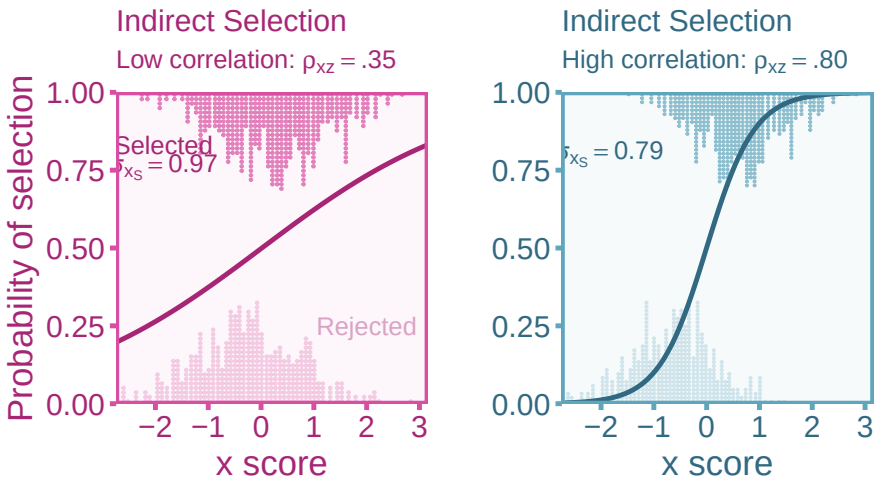
(y). The correlation between x and y (ρ_{xy}) is our estimand (i.e., the quantity we want to estimate), however as we will show, under selection on z, the correlations between the selector and the independent/dependent variable can severely bias the observed correlation.



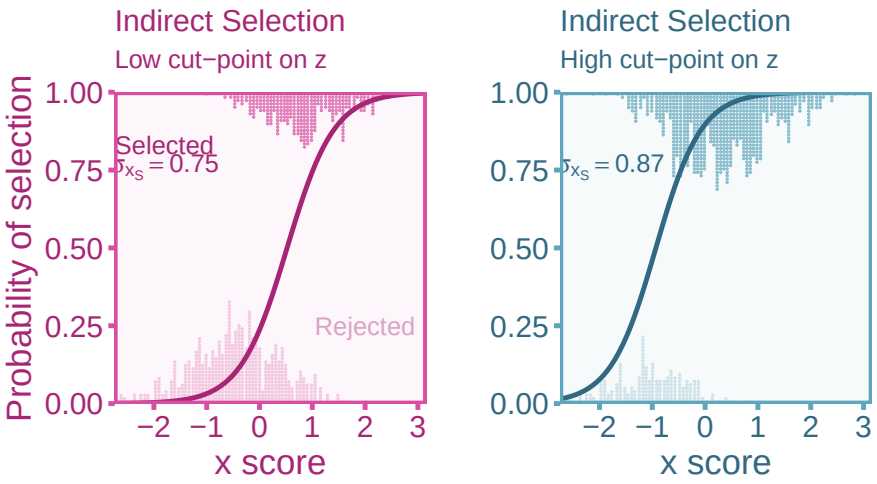
10.4.2 Artifact Correction for Correlations

The Univariate Case

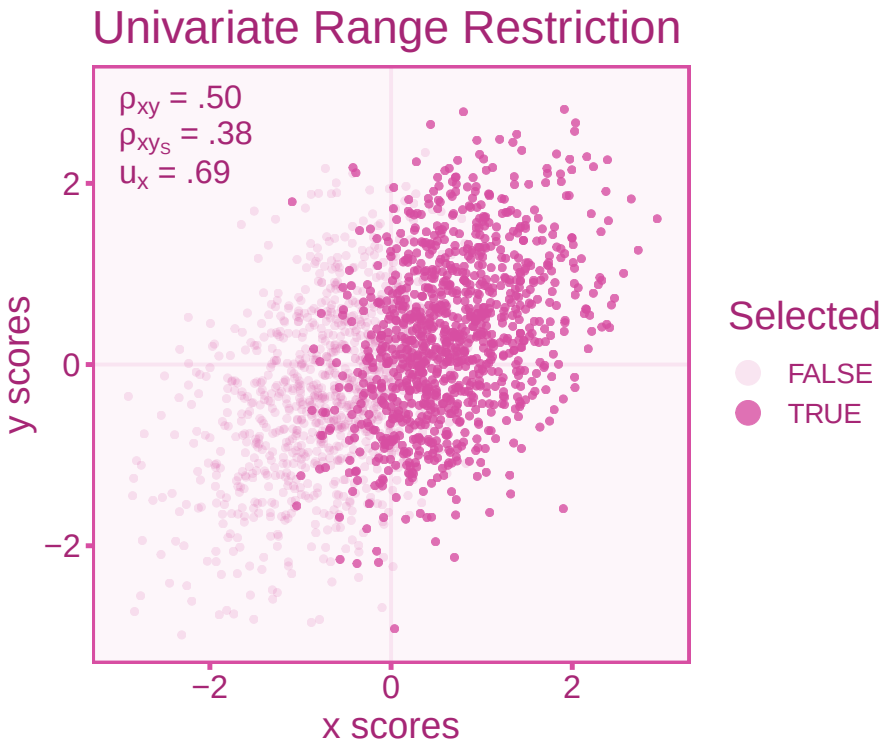
Range restriction (or enhancement) in either the independent or dependent variable will induce bias into the correlation coefficient. Let us consider a case where we select individuals based on meeting some criterion of some third variable, z. The extent to which selection on z induces restriction (or enhancement) on x (or y) depends on the correlation between z and x. We can simulate an example where individuals are selected into the sample if they are above the mean of z. Lets see how the selection process affects the distribution of x when we vary the correlation between x and z (ρ_{xz}).



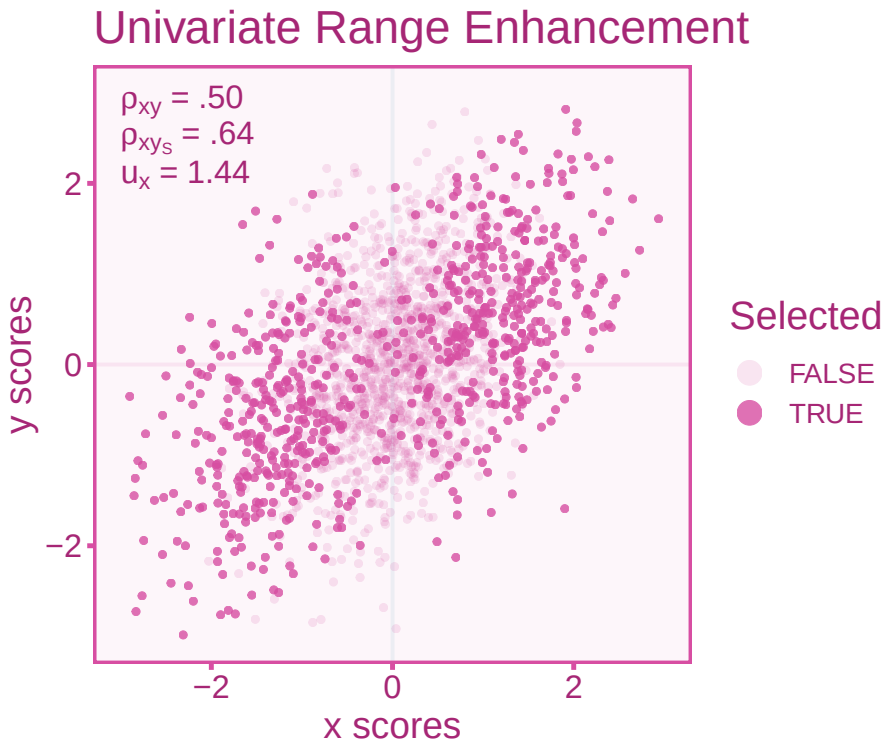
Notice that the distribution of rejected and accepted participants are more similar within the plot on the left where there is a relatively low correlation between the selector variable and the independent variable. More importantly, when the correlation is lower we see wider distributions (larger standard deviation, σ_{x_S}) than in the selected sample (and rejected), however, when the correlation is higher the standard deviation shrinks accordingly. Not only is the correlation between x and z important in the resulting variance of x, but so is where the cut-point, that is the threshold where individuals are selected above (or below). For example, the standard deviation of x will be smaller when individuals are selected above the 90th percentile of z than when individuals are selected above the median of z (i.e., the 50th percentile). We can visualize this phenomenon as well by setting a low cut-point for selection ($z > -1.0$) and a high cut-point for selection ($z > 0.5$). Notice in the figure below that the standard deviation is lower when the cut-point is higher on z.



Now lets consider a study where we want to calculate the unrestricted correlation between an independent variable, x, and a dependent variable, y. However, the sample is selected based on meeting some criterion on a selector, z. Lets look at a diagram looking at the relationship between variables. We can visualize the correlation between independent (x) and dependent (y) variables under range restriction by only selecting individuals above some cut off of our selector variable, z. The scores of individuals that have been selected will show less variance than the entire pool of individuals. Specifically, the scenario below shows a u-ratio of about 0.69 in the independent variable. We see in the figure that the correlation in the restricted scores (ρ_{xy_S}) is attenuated relative to the unrestricted (true) correlation (ρ_{xy}).



We can also visualize what happens to the correlation when the range is enhanced. Enhancement can be accomplished by selecting individuals at the ends of the distribution (Taylor and Griess 1976). In the visualization below, we see an opposite effect on the correlation, that is, an over-estimate of the unrestricted correlation rather than an attenuation like we see under range restriction. The scenario below has a u-ratio of about 1.44 in the independent variable.



It starts to become apparent that if $u_x > 1$ (i.e., $\sigma_x > \sigma_{x_S}$) the observed correlation over-estimates the true, unrestricted correlation and under-estimates the unrestricted correlation when $u_x < 1$ (i.e., $\sigma_x < \sigma_{x_S}$, Sackett and Yang 2000). A bias correction formula for univariate direct range restriction was first developed by Pearson (1903) and provided more recently by J. E. Hunter and Schmidt (1990). To correct for the systematic bias in correlations, we can use the u-ratio of the independent variable such that,

$$r_c = \frac{r_{xy_S}}{\sqrt{r_{xy_S}^2 + u_x^2(1 - r_{xy_S}^2)}} \tag{10.1}$$

This correction formula is only meant for observed scores only. If one wants to correct for range restriction and measurement error, we need to adjust the u-ratio for measurement error, as well as adjust the reliability coefficients for range restriction/enhancement. We can incorporate these adjustments into a single correction formula using the equation in table 3 of Wiernik and Dahlke

(2020),

$$r_c = \frac{r_{xy_S}}{\sqrt{r_{xy_S}^2 + \frac{u_x^2 r_{xx'_S}\left(r_{xx'_S} r_{yy'_S} - r_{xy_S}^2\right)}{1 - u_x^2\left(1 - r_{xx'_S}\right)}}}$$

Where $r_{xx'_S}$ and $r_{yy'_S}$ are the reliabilities within the selected sample. If the reliability coefficients come from the unrestricted population (e.g., from a norm/standardization sample), we can estimate the restricted reliability using the following formulas:

$$r_{xx'_S} = 1 - \frac{1 - r_{xx'}}{u_x^2} \tag{10.2}$$

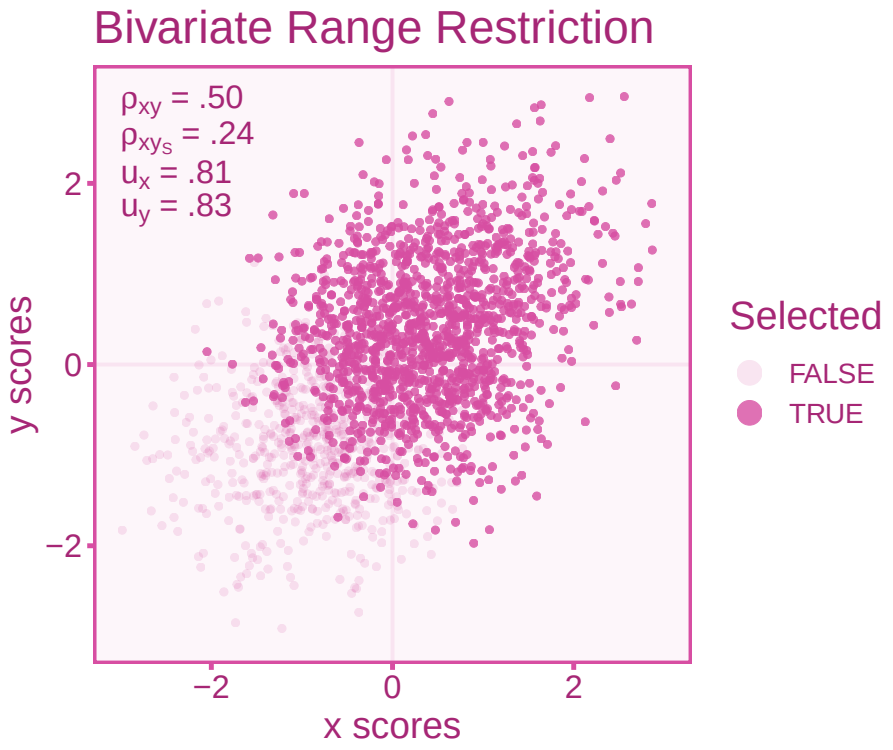
$$r_{yy'_S} = 1 - \frac{1 - r_{yy'}}{u_y^2} \tag{10.3}$$

Now once the the correlation is corrected the observed sampling variance (σ_{ε_o}) must also be adjusted as well. To do this, we can simply use the corrected correlation and the observed correlation to adjust the sampling variance:

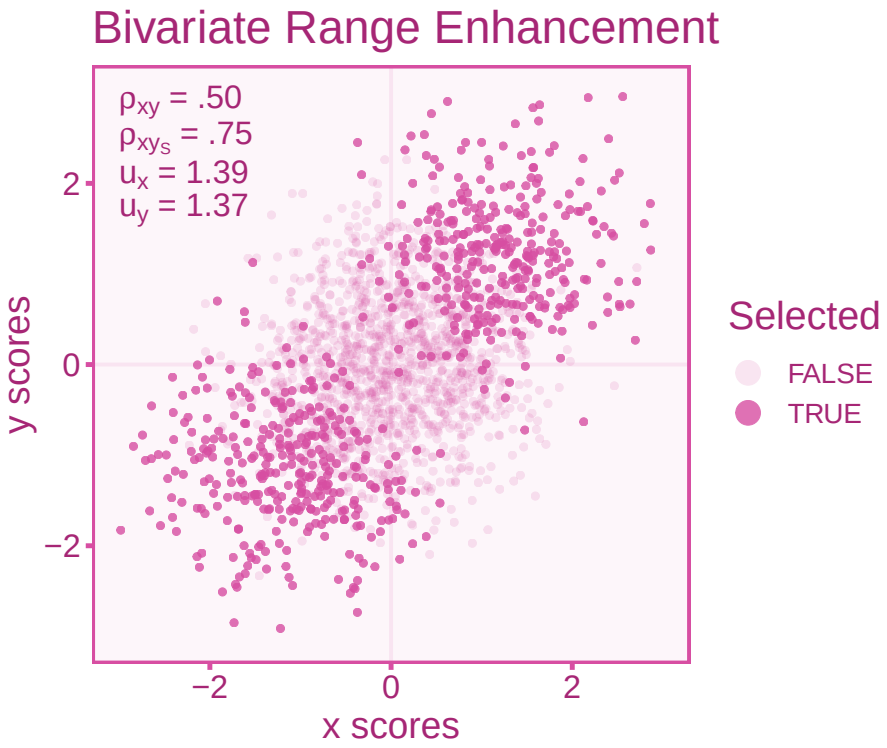
$$\sigma_{\varepsilon_c}^2 = \sigma_{\varepsilon_o}^2 \left(\frac{r_c}{r_{xy_S}}\right)^2. \tag{10.4}$$

The Bivariate Case

Bivariate direct range restriction/enhancement occurs when the selection variable has independent relationships with both the independent and dependent variable. Like we did for the univariate case, let's visualize the correlation between independent (x) and dependent (y) variables under range restriction by only selecting individuals above some cut off point in our selector variable, z. We will fix the correlations between the selector and independent variable (ρ_{xz}), as well as the selector and dependent variable (ρ_{yz}) to be .80. The x and y scores of individuals that have been selected above some threshold of z will show less variance than the entire pool of individuals. Specifically, the scenario below shows a u-ratio of about ~0.82 in the independent variable and dependent variables. We see in the figure below that the correlation in the restricted sample ($\rho_{xy_S} = .24$) is attenuated relative to the unrestricted (true) correlation ($\rho_{xy} = .50$).



Likewise let's visualize what happens to the correlation when the range is enhanced. Enhancement in both variables can be accomplished by selecting individuals at the ends of the distribution of z. In the visualization below, we observe an over-estimation of observed correlation ($\rho_{xy_S} = .75$) relative to the unrestricted correlation ($\rho_{xy} = .50$). The scenario below has a u-ratio of about ~1.38 in both the independent variable and dependent variable.



Note that bivariate restriction or bivariate enhancement will increase the amount of bias in the correlation coefficients similar to that of measurement error in both variables rather than just one (see chapter 5). A bias correction formula for bivariate range restriction is much more complicated than the univariate formulation. Particularly we need to have a basic idea of the selection mechanism at play (Dahlke and Wiernik 2019). The correction formula requires the direction of the correlation between the selector variable, z, and the independent (ρ_{xz}) and dependent variable (ρ_{yz}). For example, if we run a study on a college admissions test and college academic performance (indexed by grade-point average), we can be reasonably certain that the selection (i.e., admissions decisions) are positively correlated with both academic performance and SAT scores. Other situations this may not be as straight-forward however laying out the correlational structure of the selection mechanism is vital step of the correction procedure. To break down the correction formula into simpler parts, let us first define a factor we will denote with λ. This factor is what takes into account the direction of the correlation of ρ_{xz} (if positive, we can set $\rho_{yz} = 1$, if negative, $\rho_{yz} = -1$, if zero, $\rho_{yz} = 0$) and ρ_{yz} (repeat the same procedure as ρ_{xz}).

$$\lambda = \text{sign}\left(\rho_{xz}\rho_{yz}[1-u_x][1-u_y]\right) \times \tag{10.5}$$

$$\frac{\text{sign}\,(1-u_x)\min\left(u_x, \frac{1}{u_x}\right) + (1-u_x)\min\left(u_y, \frac{1}{u_y}\right)}{\min\left(u_x, \frac{1}{u_x}\right) + \min\left(u_y, \frac{1}{u_y}\right)} \tag{10.6}$$

The output of λ will be either -1, 0, or 1. We can then plug this factor into the full correction equation that provides us with an unbiased estimate of the correlation in the unrestricted population,

$$r_c = r_{xy_S}u_x u_y + \lambda\sqrt{|1-u_x^2||1-u_y^2|}$$

Similar to the univariate formula, we can also incorporate measurement error into the correction. Measurement error will bias the observed correlation on top of the bias induced by range restriction/enhancement. Therefore we can incorporate the reliabilities estimated within the restricted sample ($_{xx'_S}$ and $r_{yy'_S}$), into our correction formula:

$$r_c = \frac{r_{xy_S}u_x u_y + \lambda\sqrt{|1-u_x^2||1-u_y^2|}}{\sqrt{1-u_x^2(1-r_{xx'_S})}\sqrt{1-u_y^2(1-r_{yy'_S})}}$$

If the reliability estimates come from an unrestricted sample, we can get estimates of the reliability coefficients in the selected sample using Equation 10.2 and Equation 10.3. We then can correct the observed sampling variance ($\sigma^2_{\varepsilon_o}$),

$$\sigma^2_{\varepsilon_c} = \sigma^2_{\varepsilon_o}\left(\frac{r_c}{r_{xy_S}}\right)^2.$$

10.4.3 Correcting Correlations in R

Univariate Indirect Range Restriction

To correct correlations for range restriction we can start by simulating data from the the `mvrnorm` function in the `MASS` package. Lets first simulate 1000 data points. Then we will select values above the mean of the selector variable, z.

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

# set seed
```

```
set.seed(1)

# define parameters
rho_xy <- .5
n <- 1000

# simulate data
data <- mvrnorm(n=n,
                mu=c(0,0),
                Sigma = reshape_vec2mat(c(rho_xy)),
                empirical=TRUE)
x <- data[,1]
y <- data[,2]
z <- x + rnorm(n,0,.5)
selected <- z > 0
```

We can start with univariate indirect range restriction by selecting only on the independent variable. We will select only the values above the mean.

```
# calculate correlation between unrestricted and restricted scores
rxy <- cor(x,y) # unrestricted
rxyS <- cor(x[selected],y[selected]) # restricted

# print results
rbind(paste0('unrestricted: rxy = ',round(rxy,2)),
      paste0('restricted: rxyS = ',round(rxyS,2))
      )
```

```
     [,1]
[1,] "unrestricted: rxy = 0.5"
[2,] "restricted: rxyS = 0.39"
```

As expected, we observe an attenuation of the correlation under range restriction. Now lets calculate the u-ratios for both variables. We should expect the variability not only in x, but also y in the restricted sample to be smaller than the unrestricted sample. Since x and y are positively correlated, restriction x will restrict

```
# calculate u-ratios
ux <- sd(x[selected])/sd(x)
uy <- sd(y[selected])/sd(y)

# print results
rbind(paste0('ux = ',round(ux,2)),
      paste0('uy = ',round(uy,2))
```

```
      )
```

```
     [,1]
[1,] "ux = 0.69"
[2,] "uy = 0.93"
```

Now we can apply the correction for univariate direct range restriction by hand from Equation 10.1 and Equation 10.1.

```
# correct the restricted correlation
rc <- rxyS / ( ux*sqrt((1/ux^2-1)*rxyS^2+1 ))

# acquire sample size from
n <- length(x[selected])

# calculate the observed correlation sampling variance
var_e_o <- (1-rxyS^2)^2 / (n-1)

# correct sampling variance
var_e_c <- var_e_o * (rc/rxyS)^2

# print results
rbind(paste0('corrected cor: r = ',round(rc,3)),
      paste0('corrected var: var_e = ',round(var_e_c,3))
      )
```

```
     [,1]
[1,] "corrected cor: r = 0.52"
[2,] "corrected var: var_e = 0.003"
```

The correction formula produced a very close estimate of the true population correlation ($r_c = .50$ vs $\rho_{xy} = .50$). Lets also correct the correlation using the `correct_r` function in the psychmeta package, `psychmeta` (Dahlke and Wiernik 2019).

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

# correct the restricted correlation for univariate direct range restriction
correct_r(rxyi = rxyS,
          correction = 'bvirr',  # uvdrr_x = univariate direct range restriction in
          ux = ux,
          uy = uy,
          n = n)
```

```
Correlations Corrected for Measurement Error and Bivariate Indirect Range Restriction:
-------------------------------------------------------------------------------------
  value CI_LL_95 CI_UL_95   n n_effective
1 0.523    0.474    0.568 517         220
```

We can see that the correction made by the `correct_r` function provides identical results to the one done by hand.

Bivariate Indirect Range Restriction

To correct correlations for range restriction we can start by simulating data from the the `mvrnorm` function in the `MASS` package. Lets first simulate 1000 data points. Then we will select values above the mean of the selector variable, z.

```r
# load packages
# install.packages('MASS')
library(MASS)

# set seed
set.seed(1)

# define parameters
rho_xz <- .8
rho_yz <- .8
rho_xy <- .5
n <- 1000

# simulate data
data <- mvrnorm(n=n,
                mu=c(0,0,0),
                Sigma = reshape_vec2mat(c(.5,.8,.8)),
                empirical=TRUE)
x <- data[,1]
y <- data[,2]
z <- data[,3]
selected <- z > 0
```

We can see how the correlations are attenuated under bivariate indirect range restriction. We will select only the values above the mean.

```r
# calculate correlation between unrestricted and restricted scores
rxy <- cor(x,y) # unrestricted
rxyS <- cor(x[selected],y[selected]) # restricted

# print results
rbind(paste0('unrestricted: rxy = ',round(rxy,2)),
```

```
      paste0('restricted: rxyS = ',round(rxyS,2))
      )
```

```
     [,1]
[1,] "unrestricted: rxy = 0.5"
[2,] "restricted: rxyS = 0.18"
```

As expected, we observe an attenuation of the correlation under range restriction. Now lets calculate the u-ratios for both variables. We should expect the variability in x and y in the restricted sample to be smaller than the unrestricted sample.

```
# calculate u-ratios
ux <- sd(x[selected])/sd(x)
uy <- sd(y[selected])/sd(y)

# print results
rbind(paste0('ux = ',round(ux,2)),
      paste0('uy = ',round(uy,2))
      )
```

```
     [,1]
[1,] "ux = 0.8"
[2,] "uy = 0.78"
```

Now we can apply the correction for univariate direct range restriction by hand from Equation 10.1 and Equation 10.1.

```
# calculate lambda
rho_xz <- 1 # assume a positive correlation for rho_xz
rho_yz <- 1 # assume a positive correlation for rho_yz
lambda <- sign(rho_xz*rho_yz*(1-ux)*(1-uy)) * (sign(1-ux)*min(c(ux,1/ux))+sign(1-uy)

# correct the restricted correlation
rc <- rxyS*ux*uy + lambda*sqrt(abs(1-ux^2)*abs(1-uy^2))

# acquire sample size from
n <- length(x[selected])

# calculate the observed correlation sampling variance
var_e_o <- (1-rxyS^2)^2 / (n-1)

# correct sampling variance
var_e_c <- var_e_o * (rc/rxyS)^2
```

```
# print results
rbind(paste0('corrected cor: r = ',round(rc,3)),
      paste0('corrected var: var_e = ',round(var_e_c,3))
      )
```

```
     [,1]
[1,] "corrected cor: r = 0.493"
[2,] "corrected var: var_e = 0.014"
```

The correction formula produced a very close estimate of the true population correlation ($r_c = .50$ vs $\rho_{xy} = .50$). Lets also correct the correlation using the `correct_r` function in the psychmeta package, `psychmeta` (Dahlke and Wiernik 2019).

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

# correct the restricted correlation for univariate direct range restriction
correct_r(rxyi = rxyS,
          correction = 'bvirr',  # uvdrr_x = univariate direct range restriction in x
          ux = ux,
          uy = uy,
          n = n)
```

```
Correlations Corrected for Measurement Error and Bivariate Indirect Range Restriction:
--------------------------------------------------------------------------------------
  value CI_LL_95 CI_UL_95   n n_effective
1 0.493     0.44    0.545 500         465
```

We can see that the correction made by the `correct_r` function provides identical results to the one done by hand.

Part II

Application to Meta-Analysis

Chapter 11

Introduction to Meta-Analysis Methods

11.1 Introduction

Meta-analysis is an analytic tool to synthesize quantitative evidence from multiple studies. By systematically combining and analyzing the results of multiple studies, meta-analysis provides a comprehensive overview, unveiling patterns, trends, and insights that individual studies might not be able to capture. Combining research findings also has the added benefit of increasing the precision of our results (i.e., greater statistical power). In this section we will cover the method described by (J. E. Hunter and Schmidt 1990) since it is readily compatible with artifact corrections (see next chapter). For the random-effects model however, we use an integrated approach that incorporates methods from J. E. Hunter and Schmidt (1990) and Hedges and Vevea (1998) that was first introduced by Morris et al. (2014). However it is important to note that there are other common methods to conduct meta-analyses that have their strengths and weaknesses (Hedges and Olkin 2014; Callender and Osburn 1980; Johnson, Mullen, and Salas 1995).

11.2 Common-Effect Model

A common effect model is the simplest form of meta-analysis. It assumes that all the variation in observed effect sizes is attributable to sampling error. In other words, all the observed effect sizes are estimates of the same population effect size. Note that there is a distinction between *fixed-effects* models and a *common effect* model (Viechtbauer, n.d.; Laird and Mosteller 1990). The common effect model assumes that the true effect size is identical for each study while the fixed effects model does not assume this. Instead, the fixed effects model can

be interpreted as the weighted average of true effects. Computationally, they are the same and provide the same parameter estimates, yet the interpretation differs.

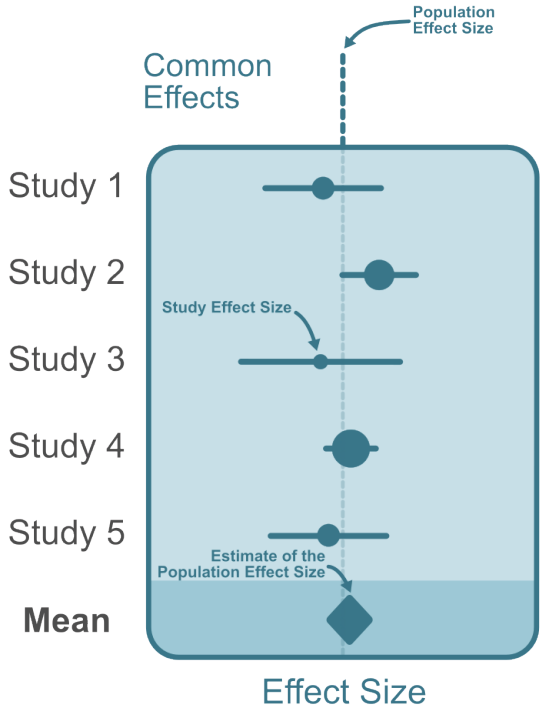


Figure 11.1: The diagram above depicts a common effect meta-analysis of five studies. The study effect sizes are homogenous and all estimate a single true population effect size.

11.2.1 The General Case

The common effect model can be modeled such that population effect size ϑ is held constant each sample (study) effect sizes (θ_i), such that,

$$\theta_i = \vartheta + \varepsilon_i \tag{11.1}$$

Where ε_i indicates sampling error and the subscript i denotes each study. Similar to the true score theory model that we discussed in chapter 4, the variance components of each term can similarly be written out as,

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

However in our fixed effects model, the population effect size is constant across studies and will not vary, simplifying the formula to,

$$\sigma_\theta^2 = \sigma_\varepsilon^2 \tag{11.2}$$

Therefore the only source of variation in the observed effect sizes, is sampling error. Since sampling error varies from study to study, we can take the average sampling variance across studies to estimate σ_ε^2:

$$\sigma_\varepsilon^2 = \frac{1}{k}\sum_{i=1}^{k}\sigma_{\varepsilon_i}^2$$

Ultimately, our goal is to obtain a precise estimate of the population effect size. To obtain an estimate of the population effect size, ϑ, we can calculate the average observed effect size, $\bar{\theta}_i$ from k studies. However, in practice, effect sizes from different studies have varying levels of precision (i.e., varying sample size). A simple average will not account for the differences between studies in their precision. Instead, we can calculate a weighted average where the weights each study can be calculated by the inverse variance (i.e., precision) of each study such that,

$$w_i = \frac{1}{\sigma_{\varepsilon_i}^2}$$

Then we can calculate a weighted average,

$$\hat{\vartheta} = \frac{\sum_{i=1}^{k} w_i\theta_i}{\sum_{i=1}^{k} w_i}$$

This weighted average will be an unbiased estimate of the population effect size. However, even though this mean effect size is more precise compared to single-study estimates, it is not exempt from error itself. In the fixed-effects model, we can obtain the standard error of our estimate of the population effect size using,

$$SE_{\hat{\vartheta}} = \sqrt{\frac{\sigma_\varepsilon^2}{k}}$$

The standard error can be used to comput the 95% confidence intervals of the meta-analytic point estimate:

$$\vartheta_{\text{Lower}} = \hat{\vartheta} - 1.96 \cdot SE_{\hat{\vartheta}}$$
$$\vartheta_{\text{Upper}} = \hat{\vartheta} + 1.96 \cdot SE_{\hat{\vartheta}}$$

11.2.2 Fixed Effects Meta-Analysis of Correlations

To apply the general case in the previous section to correlation coefficients, lets define our model similarly to Equation 11.1,

$$r_i = \rho + \varepsilon_i.$$

Where r_i is our sample (study) correlation and ρ is the population correlation. We can breakdown the variance components the same way as we did in the general case,

$$\sigma_r^2 = \sigma_\rho^2 + \sigma_\varepsilon^2$$

For each sample correlation, the large sample formulation for sampling variance is,

$$\sigma_{\varepsilon_i}^2 = \frac{\left(1-\rho^2\right)^2}{n_i} \tag{11.3}$$

Note that formulation includes the population correlation, which is unknown. Also notice that, since the population correlation is fixed, the inverse sampling variance would be proportional to the sample size ($1/\sigma_{\varepsilon_i}^2 \propto n_i$). For this reason, we can use the sample size as our weights. We can estimate the population correlation, ρ, by taking the n-weighted average,

$$\hat{\rho} = \frac{\sum_{i=1}^k n_i r_i}{\sum_{i=1}^k n_i}$$

We can use this estimate of the population correlation in the equation in Equation 11.3 to estimate the sampling variance for each study.

$$\sigma_{\varepsilon_i}^2 = \frac{\left(1-\hat{\rho}^2\right)^2}{n_i}$$

We also can acquire the standard error of our population correlation estimate ($\hat{\rho}$). To do so, we must first calculate the weighted average of the sampling variance from each study,

$$\sigma_{\bar{\varepsilon}}^2 = \frac{\sum_{i=1}^k n_i \sigma_{\varepsilon_i}^2}{\sum_{i=1}^k n_i}$$

then we can calculate the standard error of the population correlation from this by dividing by the number of studies, k,

$$SE_{\hat{\rho}} = \sqrt{\frac{\sigma_{\varepsilon}^2}{k}}$$

The standard error can be used to comput the 95% confidence intervals of the meta-analytic point estimate:

$$\rho_{\text{Lower}} = \hat{\rho} - 1.96 \cdot SE_{\hat{\rho}}$$
$$\rho_{\text{Upper}} = \hat{\rho} + 1.96 \cdot SE_{\hat{\rho}}$$

11.2.3 Fixed Effects Meta-Analysis of Standardized Mean Differences

Similar to Equation 11.1, we can model sample standardized mean differences similarly,

$$d_i = \delta + \varepsilon_i$$

The most staightforward method for meta-analyzing standardized mean differences (i.e., d values) is to first convert all the sample d values to point-biserial correlations by using,

$$r_i = \frac{d_i}{\sqrt{\frac{1}{p_i(1-p_i)} + d_i^2}}$$

Where p_i is the observed proportion of group membership in either group A or group B. The sampling variance of the study standardized mean difference is defined as

$$\sigma_{\varepsilon_d i}^2 = \frac{n_A + n_B}{n_A n_B} + \frac{d_i^2}{2(n_A + n_B)}$$

Which can then be converted to the standard error of the point-biserial correlation,

$$\sigma_{\varepsilon_r i}^2 = \frac{\sigma_{\varepsilon_d i}^2}{\left(d_i^2 p_i[1-p_i] + 1\right)\left(\frac{1}{p_i(1-p_i)} + d_i^2\right)}$$

Note that the subscripts, r and d denote the sampling variances for correlations and d values respectively. Once the d values and sampling variances are converted to point-biserial correlations, the meta-analysis can then be conducted by using the methods from Section 11.2.2. Once the meta-analysis is completed,

the estimate of the population correlation and it's standard error can be converted back to a d value using the weighted average proportion of individuals in group A or B ($\bar{p}$),

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

$$SE_{\hat{\delta}} = \sqrt{\frac{\sigma_{\varepsilon}^2}{k\,\bar{p}(1-\bar{p})(1-\hat{\rho}^2)^3}}$$

Where k is the number of studies. The standard error can be used to comput the 95% confidence intervals of the meta-analytic point estimate:

$$\delta_{\text{Lower}} = \hat{\delta} - 1.96 \cdot SE_{\hat{\delta}}$$
$$\delta_{\text{Upper}} = \hat{\delta} + 1.96 \cdot SE_{\hat{\delta}}$$

11.2.4 Example from Roth et al. (2015)

Lets use a meta-analytic data set investigating the correlation of school grades and intelligence test scores from Roth (2015). This data set has correlation coefficients from $k = 240$ studies (total sample size: $n = 105,151$) and is available within the developmental version of the `psychmeta` package (Dahlke and Wiernik 2019). Note that there is substantial heterogeneity in effect sizes here, far more than what could be accounted for by sampling error alone, but for the sake of this example we will assume that there is no heterogeneity. Lets conduct a common effect meta-analysis using the equations from the previous section:

```
# Load in packages (we need the development version of psychmeta)
# install.packages("devtools")
# devtools::install_github("psychmeta/psychmeta")
library(psychmeta)

# obtain data for correlations (r) amd sample size (n)
r <- data_r_roth_2015$rxyi
n <- data_r_roth_2015$n
k <- length(r)

# calculate the sample size weighted average of r
r_bar = sum(r*n) / sum(n)

# calculate the sampling variance for each study
var_ei <- (1-r_bar^2)^2 / n
```

```r
# estimate the mean population correlation
rho_hat <- sum(r*n) / sum(n)

# calculate the variance in study correlations (r)
var_r <- sum(n*(r-rho_hat)^2) / sum(n)

# calculate average sampling variance
var_e <- sum(n*var_ei) / sum(n)

# calculate the variance in true population correlations
var_rho <- var_r - var_e

# calculate standard error of rho estimate
SE_rho = sqrt(var_r/k)

# compute confidence interval
CI_lower = rho_hat - qnorm(.975)*SE_rho
CI_upper = rho_hat + qnorm(.975)*SE_rho

# print results
data.frame(rho_hat,
           SE = SE_rho,
           CI_lower,
           CI_upper)
```

```
   rho_hat         SE  CI_lower  CI_upper
1 0.4413862 0.01188354 0.4180949 0.4646775
```

Due to the massive sample size and the assumption that there is no variation in population correlations (i.e., fixed effects), the standard error is quite small. We can also use the `metafor` package (Viechtbauer 2010) to conduct a fixed effects meta-analysis without having to write each equation by hand.

```r
# install.packages("metafor")
library(metafor)

# fixed effects model
mdl <- rma(data = data_r_roth_2015,
           yi = rxyi,
           vi = var_ei,
           method = 'EE')

# print results
data.frame(rho_hat = mdl$b[1],
           SE = mdl$se[1],
```

```
            CI_lower = mdl$ci.lb[1],
            CI_upper = mdl$ci.ub[1])
```

```
    rho_hat          SE  CI_lower  CI_upper
1 0.4413862 0.002483049 0.4365195 0.4462529
```

11.3 Random Effects Model

The random-effects model refers to a model that allows for the population effect size to vary from study to study. Random-effects differs from the fixed effects model in an important way: it does not assume that all observed effect sizes come from a single (fixed) population effect size (Borenstein et al. 2010). This variation in population effect sizes is called heterogeneity. In the traditional J. E. Hunter and Schmidt (1990) the weights utilized in the random effects meta-analysis are identical to the common effect version (sample size weights). In other conventional meta-analysis methods (Hedges and Vevea 1998), random-effect weights include a random effect component containing the variation in population effect sizes (this has the effect of making study weights more similar to each other with more variation in population effects). A modern approach introduced by Morris et al. (2014) and later tested by Brannick et al. (2019), combined these two approaches. The simulation study by Brannick et al. (2019), concluded that weights incorporating random effect components improved the J. E. Hunter and Schmidt (1990) estimates. This section will thus use Morris's method.

11.3.1 The General Case

The model from Equation 11.1 can be changed slightly to encompass variation of the population effect size from study to study:

$$\theta_i = \vartheta_i + \varepsilon_i.$$

In the common effect model, we assumed that all the variation in study effect sizes is accounted for by variation in sampling error ($\sigma^2_\theta = \sigma^2_\varepsilon$; see Equation 11.2). However in the random-effects model the variance in population effect sizes (σ^2_ϑ) is allowed to be greater than zero. The variance components can be written out as

$$\sigma^2_\theta = \sigma^2_\vartheta + \sigma^2_\varepsilon. \tag{11.4}$$

The variance of population effects, σ^2_ϑ, can be calculated by first calculating σ^2_θ and σ^2_ε. Since the variation in study effect sizes is no longer solely accounted for by sampling error, this would suggest that $\sigma^2_\theta \neq \sigma^2_\varepsilon$, therefore we must calculate them separately. First we need to calculate study weights using the inverse of

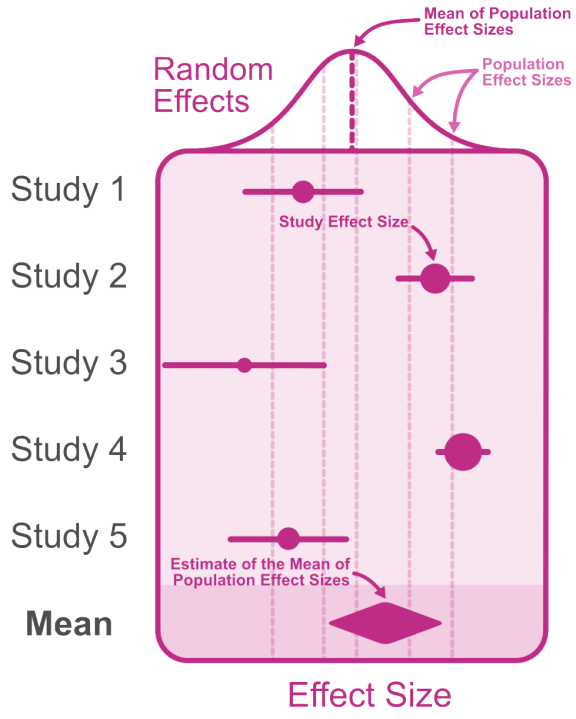


Figure 11.2: The diagram above depicts a random-effects meta-analysis of five studies. The study effect sizes are heterogeneous as population effect sizes vary.

the sampling variance and a the variance in population effect sizes (i.e., the random effect component) from each study,

$$w_i = \frac{1}{\sigma^2_{\varepsilon_i} + \sigma^2_{\vartheta}}$$

In order to estimate random effects component, σ^2_{ϑ} (i.e., the variance in population effect sizes), we can calculate it by subtracting the average sampling variance (σ^2_{ε}) from the the observed variance in effect sizes (σ^2_{θ}). The problem however is that in order to calculate the variance components, we need estimates of the population effect size and the weights, and in order to calculate the population effect size and the weights, we need the variance components. So instead, we will use sample size weights and the sample size weighted mean effect size ($\bar{\theta}$) as an estimate of the population correlation to estimate the weights:

$$w_i = \frac{1}{\sigma^2_{\varepsilon_i} + \sigma^2_{\vartheta}} = \frac{1}{\sigma^2_{\varepsilon_i} + (\sigma^2_{\theta} - \sigma^2_{\varepsilon})} \approx \frac{1}{\sigma^2_{\varepsilon_i} + \left(\frac{\sum_{i=1}^{k} n_i (\theta_i - \bar{\theta})^2}{\sum_{i=1}^{k} n_i} - \frac{\sum_{i=1}^{k} n_i \sigma^2_{\varepsilon_i}}{\sum_{i=1}^{k} n_i} \right)}$$

Now with these weights, we can calculate a more precise estimate of the mean population effect size,

$$\hat{\bar{\vartheta}} = \frac{\sum_{i=1}^{k} w_i \theta_i}{\sum_{i=1}^{k} w_i}$$

With these weights and the estimate of the population effect size, we can now estimate each of the three variance components from Equation 11.4:

1) variance in study effect sizes

$$\sigma^2_{\theta} = \frac{\sum_{i=1}^{k} w_i (\theta_i - \hat{\bar{\vartheta}})}{\sum_{i=1}^{k} w_i}$$

2) sampling error variance (mean)

$$\sigma^2_{\varepsilon} = \frac{\sum_{i=1}^{k} w_i \sigma^2_{\varepsilon_i}}{\sum_{i=1}^{k} w_i}$$

3) variance in true effect sizes

$$\sigma^2_{\vartheta} = \sigma^2_{\theta} - \sigma^2_{\varepsilon}$$

In other conventions, σ^2_ϑ is denoted as τ^2 (Borenstein et al. 2010; DerSimonian and Kacker 2007; Hedges and Vevea 1998), but conceptually these are identical. Taking the root of σ^2_ϑ, σ_ϑ is the standard deviation of population effect sizes which can be a useful measure of heterogeneity. Furthermore, we can use σ_ϑ to calculate credibility (prediction) intervals which allows us to draw inferences about the range of plausible population effect sizes. For example, the 90% credibility interval can be calculated with the following equations:

$$\vartheta_{\text{Upper}} = \hat{\bar{\vartheta}} + 1.645\sigma_\vartheta$$

$$\vartheta_{\text{Lower}} = \hat{\bar{\vartheta}} - 1.645\sigma_\vartheta$$

We can also calculate the standard error of the mean of population effect sizes ($SE_{\hat{\bar{\vartheta}}}$) by dividing the sampling error variance component by the number of studies, k,

$$SE_{\hat{\bar{\vartheta}}} = \sqrt{\frac{\sigma^2_\theta}{k}}$$

Which can then be used to calculate 95% confidence intervals:

$$\bar{\vartheta}_{\text{Upper}} = \hat{\bar{\vartheta}} + 1.96 \cdot SE_{\hat{\bar{\vartheta}}}$$

$$\bar{\vartheta}_{\text{Lower}} = \hat{\bar{\vartheta}} - 1.96 \cdot SE_{\hat{\bar{\vartheta}}}$$

The confidence interval and credibility interval have fundamentally different interpretations that are often misinterpreted in published work (Whitener 1990). When we are interpreting a single realized interval (i.e., our estimate-in-hand), the 90% credibility interval can be interpreted as the region in which 90% of population effect sizes exist, however, a 95% confidence interval describes the interval in which there is a 95% probability of containing the true mean of population effect sizes. It is important to note that the confidence interval interpretation here is only valid in the case of a single realized interval (Vos and Holbert 2022), if there is more than one computed intervals the same population of studies, then the interpretation does not hold (this would be an exceedingly rare scenario in a meta-analysis).

11.3.2 Random Effects Meta-Analysis of Correlations

Lets now specifically apply the random effects model to pearson correlation coefficients. Let us again start by defining the meta-analytic model allowing the population correlation to vary for each study,

$$r_i = \rho_i + \varepsilon_i$$

Where it's corresponding variance components are defined similarly as,

$$\sigma_r^2 = \sigma_\rho^2 + \sigma_\varepsilon^2$$

Like in the general case, we must calculate the study weights using the method by Morris et al. (2014) and further described in Brannick et al. (2019). The weights are a function of the study-level sampling variance ($\sigma_{\varepsilon_i}^2$) and the variance in population correlations (σ_ρ^2).

$$w_i = \frac{1}{\sigma_{\varepsilon_i}^2 + \sigma_\rho^2}$$

However as described in the last section, to estimate the variance in population effect sizes (σ_ρ^2), we need estimates of the mean of true population effect sizes and the weights, but to get both those parameters, we need the weights. In order to get around this dilemma we can instead replace the weights with n_i and the mean of population correlations with the n-weighted average correlation ($\bar{r}$). Lets first define the sampling variance for a pearson correlation:

$$\sigma_{\varepsilon_i}^2 = \frac{\left(1-\rho^2\right)^2}{n_i}$$

Therefore we can approximate the weights with,

$$w_i = \frac{1}{\sigma_{\varepsilon_i}^2 + \sigma_\vartheta^2} = \frac{1}{\sigma_{\varepsilon_i}^2 + (\sigma_\theta^2 - \sigma_\varepsilon^2)} \approx \frac{1}{\sigma_{\varepsilon_i}^2 + \left(\frac{\sum_{i=1}^k n_i (r_i - \bar{r})^2}{\sum_{i=1}^k n_i} - \frac{\sum_{i=1}^k n_i \sigma_{\varepsilon_i}^2}{\sum_{i=1}^k n_i}\right)}$$

With the weights we can estimate a precise estimate of the mean of population correlations ($\bar{\rho}$)

$$\hat{\bar{\rho}} = \frac{\sum_{i=1}^k w_i r_i}{\sum_{i=1}^k w_i}$$

Where the variance components can be calculated as:

1) Variance in study correlations:

$$\sigma_r^2 = \frac{\sum_{i=1}^k w_i (r_i - \hat{\bar{\rho}})}{\sum_{i=1}^k w_i}.$$

2) Sampling error variance (mean):

$$\sigma_{\varepsilon}^2 = \frac{\sum_{i=1}^{k} w_i \sigma_{\varepsilon_i}^2}{\sum_{i=1}^{k} w_i}.$$

3) Variance in population correlations:

$$\sigma_{\rho}^2 = \sigma_r^2 - \sigma_{\varepsilon}^2.$$

Now lets use these variance components to calculate the 90% credibility (prediction) interval and the 95% confidence interval. The 90% credibility interval can be calculated with the following equations:

$$\rho_{\text{Upper}} = \hat{\bar{\rho}} + 1.645\sigma_{\rho}$$

$$\rho_{\text{Lower}} = \hat{\bar{\rho}} - 1.645\sigma_{\rho}$$

We can also calculate the standard error of the mean of population effect sizes ($SE_{\hat{\bar{\rho}}}$) by dividing the sampling error variance component by the number of studies, k,

$$SE_{\hat{\bar{\rho}}} = \sqrt{\frac{\sigma_r^2}{k}}$$

Which can then be used to calculate 95% confidence intervals:

$$\bar{\rho}_{\text{Upper}} = \hat{\bar{\rho}} + 1.96 \cdot SE_{\hat{\bar{\rho}}}$$

$$\bar{\rho}_{\text{Lower}} = \hat{\bar{\rho}} - 1.96 \cdot SE_{\hat{\bar{\rho}}}$$

11.3.3 Random Effects Meta-Analysis of Standardized Mean Differences

We can model sample standardized mean differences similarly to that of correlations,

$$d_i = \delta_i + \varepsilon_i$$

Like we did in the common effect model, instead of meta-analyzing the d values, we can instead convert all the sample d values to point-biserial correlations by using,

$$r_i = \frac{d_i}{\sqrt{\frac{1}{p_i(1-p_i)} + d_i^2}}$$

Where p_i is the observed proportion of group membership in either group A or group B. The sampling variance of the study standardized mean difference can be defined as

$$\sigma^2_{\varepsilon_i d} = \frac{n_A + n_B}{n_A n_B} + \frac{\delta_i^2}{2(n_A + n_B)}$$

Where the population standardized mean difference, δ_i can be approximated with the sample size weighted mean d value ($\bar{d}$).

$$\sigma^2_{\varepsilon_i d} = \frac{n_A + n_B}{n_A n_B} + \frac{\bar{d}^2}{2(n_A + n_B)}$$

Which can then be converted to the standard error of the point-biserial correlation,

$$\sigma^2_{\varepsilon_i r} = \frac{\sigma^2_{\varepsilon_i d}}{(d_i^2 p_i[1-p_i]+1)\left(\frac{1}{p_i(1-p_i)} + d_i^2\right)}$$

The subscripts, r and d denote the sampling variances for correlations and d values respectively. Once the d values and sampling variances are converted to point-biserial correlations, the meta-analysis can then be conducted by using the methods from Section 11.2.2. Once the meta-analysis is completed, the estimate of the population correlation and it's standard error can be converted back to a d value,

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

$$SE_{\hat{\bar{\delta}}} = \sqrt{\frac{\sigma_r^2}{k\,\bar{p}\,(1-\bar{p})\left(1-\hat{\bar{\rho}}^2\right)^3}}$$

Where k is the number of studies. Likewise we can also convert the variance of the population correlations to the standard deviation of population standardized mean differences.

$$\sigma_\delta = \sqrt{\frac{\sigma_\rho^2}{\bar{p}(1-\bar{p})(1-\hat{\rho}^2)^3}}$$

11.3.4 Random Effects Meta-Analysis in R

Lets conduct a random effects meta-analysis using the equations from the previous section and the data set we used earlier (Roth 2015). For this dataset, it is more appropriate to use a random-effects model due to the large amount of heterogeneity we observe in the correlations.

```
# Load in packages (we need the development version of psychmeta)
# install.packages("devtools")
# devtools::install_github("psychmeta/psychmeta")
library(psychmeta)

# obtain data for correlations (r) amd sample size (n)
r <- data_r_roth_2015$rxyi
n <- data_r_roth_2015$n
k <- length(r)

# calculate the sample size weighted average of r
r_bar = sum(r*n) / sum(n)

# calculate the sampling variance for each study
var_ei <- (1-r_bar^2)^2 / n

# calculate weights
w <- 1 / (var_ei + ( (sum(n*(r-r_bar)^2)/sum(n)) - (sum(n*var_ei)/sum(n)) ) )

# estimate the mean population correlation
mean_rho_hat <- sum(r*w) / sum(w)

# calculate the variance in study correlations (r)
var_r <- sum(w*(r-mean_rho_hat)^2) / sum(w)

# calculate average sampling variance
var_e <- sum(w*var_ei) / sum(w)

# calculate the variance in population correlations
var_rho <- var_r - var_e

# calculate standard error of rho estimate
SE_rho = sqrt(var_r/k)

# compute 95% confidence interval
CI_lower = mean_rho_hat - qnorm(.975)*SE_rho
CI_upper = mean_rho_hat + qnorm(.975)*SE_rho
```

```
# compute 90% credibility interval
CR_lower = mdl$b[1] - qnorm(.95)*sqrt(var_rho)
CR_upper = mdl$b[1] + qnorm(.95)*sqrt(var_rho)

# print results
data.frame(mean_rho_hat,
           SE = SE_rho,
           CI_lower,
           CI_upper,
           SD_rho = sqrt(var_rho),
           CR_lower,
           CR_upper)
```

```
  mean_rho_hat         SE  CI_lower  CI_upper    SD_rho  CR_lower  CR_upper
1    0.4265249 0.01131785 0.4043424 0.4487075 0.1601417 0.1779766 0.7047958
```

Notice that the standard error of the mean correlation is larger than the common effects model. The reason for this disparity, is that the random effects model has two sources of variance, sampling error and variance in true correlations. We can also use the `metafor` package (Viechtbauer 2010) to conduct a random effects meta-analysis. The method used in the previous sections is not available in metafor so slight deviations with the resulting statistics may occur.

```
# install.packages("metafor")
library(metafor)

# fixed effects model
mdl <- rma(data = data_r_roth_2015,
           yi = rxyi,
           vi = var_ei,
           method = 'HS')

# print results
data.frame(mean_rho_hat = mdl$b[1],
           SE = mdl$se[1],
           CI_LO = mdl$ci.lb[1],
           CI_HI = mdl$ci.ub[1],
           SD_rho = sqrt(mdl$tau2),
           CR_LO = mdl$b[1] - qnorm(.95)*sqrt(mdl$tau2),
           CR_HI = mdl$b[1] + qnorm(.95)*sqrt(mdl$tau2))
```

```
  mean_rho_hat         SE     CI_LO     CI_HI    SD_rho     CR_LO     CR_HI
1    0.4265249 0.01250165 0.4020222 0.4510277 0.1800354 0.1303931 0.7226568
```

Chapter 12

Artifact Correction Meta-Analysis

12.1 Introduction

Artifact correction meta-analysis, also referred to as psychometric meta-analysis, is a form of meta-analysis where effect sizes are systematically corrected for sources of bias. These sources of bias have been discussed in previous chapters 4-10. Methodology for conducting artifact correction style meta-analyses were originally pioneered by Frank Schmidt and John Hunter (J. E. Hunter and Schmidt 1990; Schmidt and Hunter 1977) and then reviewed more recently by Brenton Wiernik and Jeffrey Dahlke (Wiernik and Dahlke 2020). There has also been powerful R packages developed to aide in the application of artifact correction meta-analyses that we have used in previous chapters (Dahlke and Wiernik 2019). You will notice that in this section, we do not discuss standardized mean differences. This is due to the fact that the artifact correction model is designed for pearson correlations, in order to use this method for standardized mean differences, convert to pearson correlations using the methods described in chapter 11, and then use the correction methods used below. Once you apply the corrections to the converted correlations they can then be converted back to a standardized mean difference.

12.2 Bare Bones vs Artifact Correction Meta-Analysis

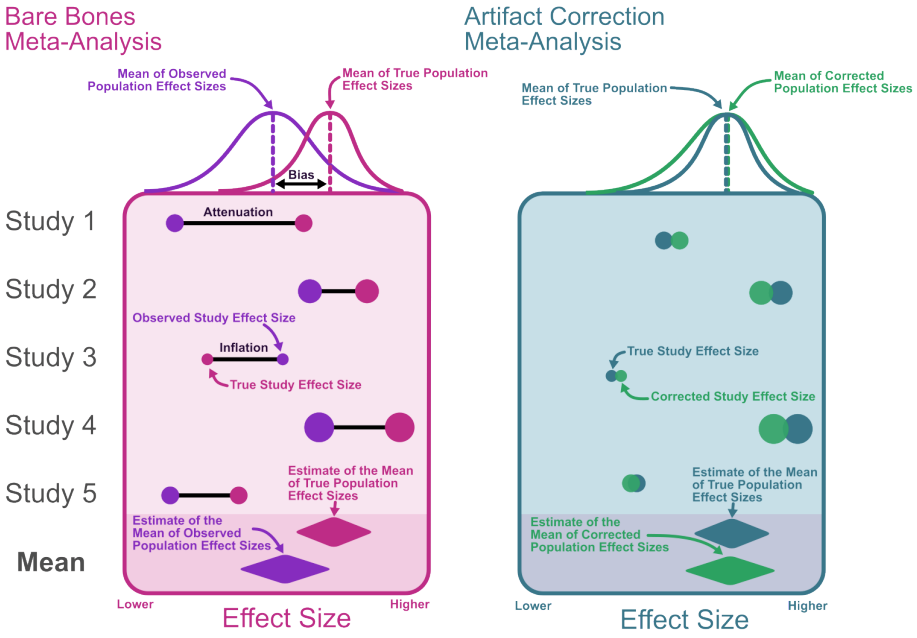
This is because even if the estimates are biased relative to our estimand (i.e., the thing we are trying to estimate), the observed value still has its own population value. Chapter 11 focused on bare-bones meta-analysis, that is, meta-analyses

that do not correct for biases in effect size estimates. This section will be dedicated to the artifact correction style of meta-analysis that does aim to correct for such artifactual biases. The choice between these two types of meta-analyses depends on the research question, the available data, and the assumptions researchers are willing to make. If the goal is to investigate the state of the quantitative evidence while avoiding additional assumptions about the data, then a bare-bones meta-analysis might be the way to go. On the other hand, if the goal is to obtain a more accurate estimate of the true effect size by accounting for biases induced by statistical artifacts, an artifact correction meta-analysis is preferable.

1. Bare-Bones Meta-Analysis: In a bare-bones meta-analysis, the focus is on aggregating effect sizes from various studies without explicitly correcting for potential biases in these effect size estimates.

2. Artifact Correction Meta-Analysis: In contrast, an artifact correction meta-analysis takes into account and attempts to correct for biases that may be present in the effect size estimates from individual studies. This involves addressing potential sources of bias, such as measurement error or selection effects, through statistical techniques or adjustments. By doing so, the meta-analysis aims to provide a more accurate and unbiased estimate of the true effect size. Although it is important to note that this method will require additional assumptions about the nature of the data.

Note that the bare-bones model does not assume that there is no bias, rather, the bare-bones model is estimating something else entirely, that is, the observed population effect size.



12.3 Individual Artifact Correction Model

The individual artifact correction model corrects each effect size individually prior to conducting the meta-analysis. This method is ideal if we have high-quality artifact estimates for most/all studies in the meta-analysis. If there is a substantial amount of missingness in the artifact values, then the artifact distribution model may be a better choice.

12.3.1 The General Case

Let us recall the random effects model in chapter 11, where $\theta_i = \vartheta_i + \varepsilon_i$. This model would be considered a bare-bones meta-analytic model; we can re-write it slightly to denote that these are observed values: $\theta_{o_i} = \vartheta_{o_i} + \varepsilon_{o_i}$. Ultimately, observed values tend to be biased relative to true values due to many artifactual factors, some that we can account for and some we can not. If we decide that corrections to observed effect sizes are necessary to answer our research question, then we can construct an artifact correction model. In the artifact correction framework, we can incorporate a compound artifact biasing factor, A, to the bare-bones formula such that,

$$\theta_{o_i} = A_i \vartheta_i + \varepsilon_{o_i} \tag{12.1}$$

So now instead of the model being in terms of the observed population value

(ϑ_{o_i}), it is now in terms of the *true* population value (ϑ_i). The compound biasing factor, A_i is a product of multiple independent artifact values (e.g., unreliability and range restriction),

$$A_i = a_{1i}a_{2i}a_{3i}...$$

This compound artifact formula assumes that the values are independent of one another, which is not always the case, see chapter 11 on independence of artifacts. Therefore adjustments must be made to ensure independence. eq-art-mdl can be re-arranged to obtain unbiased estimates of the true population effect size:

$$\frac{\theta_{o_i}}{A_i} = \vartheta_i + \frac{\varepsilon_{o_i}}{A_i} \tag{12.2}$$

This division of A_i will provide us with our corrected effect size estimates that we can denote with the subscript, c,

$$\theta_{c_i} = \frac{\theta_{o_i}}{A_i}$$

and the corresponding error term must also be corrected

$$\varepsilon_{c_i} = \frac{\varepsilon_{o_i}}{A_i}. \tag{12.3}$$

Therefore Equation 12.2 can be reformulated as,

$$\theta_{c_i} = \vartheta_i + \varepsilon_{c_i} \tag{12.4}$$

These corrections cause changes in the point estimate and the error variance of the study effect sizes. Like we saw in chapter 11, we can breakdown the variance components of the model,

$$\sigma^2_{\theta_c} = \sigma^2_{\vartheta} + \sigma^2_{\varepsilon_c} \tag{12.5}$$

To obtain these variance components, we can start by correcting the observed sampling variances from each study. We can calculate the corrected sampling variance ($\sigma^2_{\varepsilon_c}$) by first correcting each study-level sampling variance estimate,

$$\sigma^2_{\varepsilon_c i} = \frac{\sigma^2_{\varepsilon_o i}}{A^2_i}$$

This may also be done by using the corrected effect size and the observed effect size to correct the sampling variance:

$$\sigma^2_{\varepsilon_c i} = \sigma^2_{\varepsilon_o i} \left(\frac{\theta_{c_i}}{\theta_{o_i}} \right)^2$$

The next step is to obtain the random effects weights of the study, we can do this with the inverse corrected variance for each study, $w_i = 1/(\sigma^2_{\varepsilon_c i} + \sigma^2_\rho)$. From here we can calculate our estimate of the mean of true population correlations,

$$\hat{\bar{\vartheta}} = \frac{\sum_{i=1}^k n_i \theta_{c_i}}{\sum_{i=1}^k n_i}$$

Remember that because this is a random effects model, $\hat{\bar{\vartheta}}$ is not an estimate of the true population effect size, instead it is an estimate of the mean of a distribution of true population effect sizes. Now that we have an estimate of the mean and the corrected sampling variances, the variance components from Equation 12.5 can be easily calculated as follows:

$$\sigma^2_{\theta_c} = \frac{\sum_{i=1}^k n_i (\theta_{c_i} - \hat{\bar{\vartheta}})^2}{\sum_{i=1}^k n_i}$$

$$\sigma^2_{\varepsilon_c} = \frac{\sum_{i=1}^k n_i \sigma^2_{\varepsilon_c i}}{\sum_{i=1}^k n_i}$$

$$\sigma^2_\vartheta = \sigma^2_{\theta_c} - \sigma^2_{\varepsilon_c}$$

The standard deviation of true effects is a useful measure of heterogeneity and is simply the square root of the variance of true population effect sizes (σ_ϑ). From the standard deviation in true effects, we can also calculate a credibility (prediction) interval that shows the range of plausible values for which a true effect size is likely to fall,

$$\vartheta_{\text{Upper}} = \hat{\bar{\vartheta}} + 1.645\sigma_\vartheta$$

$$\vartheta_{\text{Lower}} = \hat{\bar{\vartheta}} - 1.645\sigma_\vartheta \,.$$

Note that this is not to be confused with *confidence* intervals which denotes the range of plausible values that the *mean* of true effects can take on. This differentiation is akin to understanding the disparity between the standard error

of the mean and the standard deviation in the context of a normal distribution. We can also see how the corrections reduced the heterogeneity in the effect size estimates by comparing variance in true effect sizes ($\sigma^2_{\vartheta_o}$) to the variance in observed effect sizes ($\sigma^2_{\vartheta_o}$; this can be calculated by conducting a bare-bones random effects meta-analysis described in chapter 11). The percent reduction in heterogeneity can be computed by taking the ratio of the two, $\sigma^2_{\vartheta}/\sigma^2_{\vartheta_o}$. J. E. Hunter and Schmidt (1990) suggest that if 75% of the heterogeneity is accounted for by artifact corrections, then we can assume that the remaining heterogeneity is attributable to remaining artifacts that have not been addressed in the current meta-analysis. Although it is important to point out that this is simply a rule of thumb rather than a mathematical property.

12.3.2 Individual Corrections in Correlations

For correlation coefficients we can define the model similarly to Equation 12.1, with the only difference being that we will use the notation for pearson correlations,

$$r_{o_i} = A_i \rho_i + \varepsilon_{o_i}$$

The artifact correction formulation of this, corresponding to Equation 12.4, would be

$$r_{c_i} = \rho_i + \varepsilon_{c_i}$$

The corresponding variance components would then be,

$$\sigma^2_{r_c} = \sigma^2_{\rho} + \sigma^2_{\varepsilon_c}$$

In order to compute the variance components as well as the mean true population correlation, we first need to calculate the study weights. We will follow a similar procedure for calculating random effects weights in chapter 11. Lets define the corrected random effects weights as,

$$w_i = \frac{1}{\sigma^2_{\varepsilon_c i} + \sigma^2_{\rho}}.$$

However the variance components, $\sigma^2_{\varepsilon_c i}$ and σ^2_{ρ}, require the weights themselves to actually estimate them, so instead we can approximate the variance components using the sample size as the weights such that,

$$w_i = \frac{1}{\sigma^2_{\varepsilon_c i} + \sigma^2_{\rho}} = \frac{1}{\sigma^2_{\varepsilon_c i} + (\sigma^2_{r_c} - \sigma^2_{\bar{\varepsilon}})} \approx \frac{1}{\sigma^2_{\varepsilon_c i} + \left(\frac{\sum_{i=1}^k n_i (r_{c_i} - \bar{r}_c)^2}{\sum_{i=1}^k n_i} - \frac{\sum_{i=1}^k n_i \sigma^2_{\varepsilon_c i}}{\sum_{i=1}^k n_i} \right)}$$

Where $\bar{r}_c$ is the sample size weighted average corrected correlation. These weights can then be used to obtain a more precise estimate of the true population correlation,

$$\hat{\bar{\rho}} = \frac{\sum_{i=1}^{k} w_i r_{c_i}}{\sum_{i=1}^{k} w_i}$$

Now we can compute each of the three variance components:

1) Variance in corrected correlations:

$$\sigma^2_{r_c} = \frac{\sum_{i=1}^{k} w_i (r_{c_i} - \hat{\bar{\rho}})}{\sum_{i=1}^{k} w_i}.$$

2) Sampling error variance:

$$\sigma^2_{\varepsilon_c} = \frac{\sum_{i=1}^{k} w_i \sigma^2_{\varepsilon_c i}}{\sum_{i=1}^{k} w_i}.$$

3) Variance in population correlations:

$$\sigma^2_{\rho} = \sigma^2_{r} - \sigma^2_{\varepsilon}.$$

Now lets use these variance components to calculate the 90% credibility (prediction) interval and the 95% confidence interval. The 90% credibility interval can be calculated with the following equations:

$$\rho_{\text{Upper}} = \hat{\bar{\rho}} + 1.645\sigma_{\rho}$$

$$\rho_{\text{Lower}} = \hat{\bar{\rho}} - 1.645\sigma_{\rho}$$

We can also calculate the standard error of the mean of true population effect sizes ($SE_{\hat{\bar{\rho}}}$) by dividing the sampling error variance component by the number of studies, k,

$$SE_{\hat{\bar{\rho}}} = \sqrt{\frac{\sigma^2_{r_c}}{k}}$$

Which can then be used to calculate 95% confidence intervals:

$$\bar{\rho}_{\text{Upper}} = \hat{\bar{\rho}} + 1.96 \cdot SE_{\hat{\bar{\rho}}}$$

$$\bar{\rho}_{\text{Lower}} = \hat{\bar{\rho}} - 1.96 \cdot SE_{\hat{\bar{\rho}}}$$

12.3.3 Applied Example in R

Lets conduct an individual correction meta-analysis in r using the data set by Roth (2015). This data set consists of correlations between school grades and intelligence test scores. It also contains information on the reliability of the intelligence test scores and the extent of range restriction in test scores. We can conduct a meta-analysis correcting for univariate indirect range restriction and measurement error in test scores. The compound artifact biasing factor for the correlation would be:

$$A_i = \sqrt{r_{o_i}^2 + \frac{u_{x_i}^2 r_{xx'_i}(r_{xx'_i} - r_{o_i}^2)}{1 - u_{x_i}^2(1 - r_{xx'_i})}}$$

Sticking with our theme of doing everything in base R first, lets use the equations from the previous section to conduct the meta-analysis.

```
# Load in packages (we need the development version of psychmeta)
# install.packages("devtools")
# devtools::install_github("psychmeta/psychmeta")
library(psychmeta)

# obtain artifact values
rxx <- data_r_roth_2015$rxxi
ux <- data_r_roth_2015$ux
ro <- data_r_roth_2015$rxyi
n <- data_r_roth_2015$n
k <- length(ro)

# fill in missing artifact values with mean
rxx[is.na(rxx)] <- mean(rxx,na.rm=TRUE)
ux[is.na(ux)] <- mean(ux,na.rm=TRUE)

# calculate compound artifact biasing factor for univariate direct range restrictior
A <- sqrt(ro^2 + (ux^2*rxx*(rxx - ro^2)) / (1 - ux^2*(1-rxx)))

# calculate the sample size weighted average of r
ro_bar <- sum(ro*n) / sum(n)

# calculate the observed sampling variance for each study
var_eoi <- (1-ro_bar^2)^2 / (n-1)

# correct sampling variance
var_eci <- var_eoi / A^2

# calculate corrected correlations
rc <- ro / A
```

```
# calculate weights
w <- 1/var_eci

# calculate population effect size estimate
mean_rho_hat <- sum(rc*w) / sum(w)

# calculate the variance in corrected correlations (rc)
var_rc <- sum(w*(rc - mean_rho_hat)^2) / sum(w)

# calculate average corrected sampling variance
var_ec <- sum(var_eci*w) / sum(w)

# calculate the variance in true population correlations
var_rho <- var_rc - var_ec

# calculate standard error of rho estimate
SE_rho = sqrt(var_rc/k)

# print results
data.frame(k = k,
           n = sum(n),
           mean_rho_hat,
           SE = SE_rho,
           SD_rho = sqrt(var_rho))
```

```
    k      n mean_rho_hat         SE    SD_rho
1 240 105151    0.5398838 0.01339916 0.2022865
```

The estimated mean correlation of .540 is precisely what is precisely what the original paper reported (Roth 2015). Lets conduct the meta-analysis using the the `psychmeta` package (Dahlke and Wiernik 2019). The function `ma_r_ic` is designed to conduct an individual correction meta-analysis on correlation coefficients.

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

# conduct individual correction meta-analysis
mdl_ic <- ma_r_ic(rxyi = ro, n = n,
            correction_method = "uvirr",
            rxx = rxx,
            ux = ux,
            ux_observed = TRUE,
            rxx_restricted = TRUE)
```

```
summary_stats <- data.frame(k = mdl_ic$meta_tables$`analysis_id: 1`$individual_corre
                            n = mdl_ic$meta_tables$`analysis_id: 1`$individual_corre
                            mean_rho = mdl_ic$meta_tables$`analysis_id: 1`$individua
                            SE = mdl_ic$meta_tables$`analysis_id: 1`$individual_corr
                            SD_rho = mdl_ic$meta_tables$`analysis_id: 1`$individual_
summary_stats
```

```
    k      n  mean_rho        SE    SD_rho
1 240 105151 0.5404134 0.0134356 0.2036946
```

We can also obtain credibility intervals by using the `credibility` function in the `psychmeta` package. The interval defaults to 80% intervals, however we can change that to 90% by inputting .90 into the `cred_level` argument.

```
credibility(mean = summary_stats$mean_rho_hat,
            sd = summary_stats$SD_rho,
            cred_method = "norm",
            cred_level = .90)
```

```
  CR_LL_90 CR_UL_90
```

Lets compare these results to the bare-bones model. In `psychmeta` the bare-bones model can be conduced using `ma_r_bb`. However, the `ma_r_ic` function also reports the bare-bones results as well. Therefore we can just extract the necessary statistics from the model.

```
data.frame(
  k = mdl_ic$meta_tables$`analysis_id: 1`$barebones$k,
  n = mdl_ic$meta_tables$`analysis_id: 1`$barebones$N,
  mean_rho_obs = mdl_ic$meta_tables$`analysis_id: 1`$barebones$mean_r,
  SE = mdl_ic$meta_tables$`analysis_id: 1`$barebones$se_r,
  SD_rho_obs = mdl_ic$meta_tables$`analysis_id: 1`$barebones$sd_r)
```

```
    k      n mean_rho_obs         SE SD_rho_obs
1 240 105151    0.4418789 0.01191933  0.1846534
```

We can see that the estimate of the population correlation is largely attenuated in the observed values. This is due to the fact tests of intelligence are not perfectly reliable and the scores were restricted in their range.

12.4 Artifact Distribution Model

When we observe a lot of missingness in artifact values (e.g., studies not reporting reliability), we may choose to use an artifact distribution model. The artifact distribution model conducts a meta-analysis on the observed effect sizes

and artifact values separately, and then uses the aggregate artifact values to correct for the observed mean effect size. Since the artifact distribution method uses Taylor series approximations (Dahlke and Wiernik 2020) that are custom-tailored to estimate the sampling variance of corrected correlations, we will skip the general case to focus on its application to correlations.

12.4.1 The Correlational Case

The model here can be broken down into two parts, the first part aggregates the observed effect sizes and the second part aggregates the artifact values. The artifact values we will focus on here are the reliability coefficients (see chapter 5 and 6), however other artifact values like u-ratios will follow similar procedures. We can start with the bare-bones meta-analysis model: $r_{o_i} = \rho_{o_i} + \varepsilon_{o_i}$. We can estimate the observed population correlation (ϑ_{o_i}) by first calculating the weights (using the n-weighted mean correlation in the formula for sampling variance, $\bar{r}$):

$$\sigma^2_{\varepsilon_o i} \approx \frac{(1-\bar{r}^2)^2}{n_i - 1}$$

$$w_i = \frac{1}{\sigma^2_{\varepsilon_i} + \sigma^2_{\vartheta}} = \frac{1}{\sigma^2_{\varepsilon_i} + (\sigma^2_{\theta} - \sigma^2_{\varepsilon})} \approx \frac{1}{\sigma^2_{\varepsilon_i} + \left(\frac{\sum_{i=1}^{k} n_i(\theta_i - \bar{\theta})^2}{\sum_{i=1}^{k} n_i} - \frac{\sum_{i=1}^{k} n_i \sigma^2_{\varepsilon_o i}}{\sum_{i=1}^{k} n_i} \right)}$$

Taking the mean of the observed study correlations weighted by the inverse sampling variance,

$$\hat{\bar{\rho}}_o = \frac{\sum_{i=1}^{k} w_i r_{o_i}}{\sum_{i=1}^{k} w_i}$$

Then lets get the variance in observed population correlations, in order to do this we need the v

$$\sigma^2_{\rho_o} = \sigma^2_{r_o} - \sigma^2_{\varepsilon_o} = \frac{\sum_{i=1}^{k} w_i (r_{o_i} - \hat{\bar{\rho}}_o)^2}{\sum_{i=1}^{k} w_i} - \frac{\sum_{i=1}^{k} w_i \sigma^2_{\varepsilon_o i}}{\sum_{i=1}^{k} w_i}$$

With the weights we can also take the weighted average of the artifact values (such as u-ratios or reliabilities) that are available. For our example here, we will correct only for measurement error, therefore the weighted means for reliability in x and y will be:

$$\bar{r}_{xx'} = \frac{\sum_{i=1}^{k} w_i r_{xx'_i}}{\sum_{i=1}^{k} w_i}$$

$$\bar{r}_{yy'} = \frac{\sum_{i=1}^{k} w_i r_{yy'_i}}{\sum_{i=1}^{k} w_i}$$

Now recall from chapter 5 that the square root of the reliability is equal to the correlation between observed scores and true scores. We can denote the mean correlation as follows: $\bar{r}_{xT} = \sqrt{\bar{r}_{xx'}}$ and $\bar{r}_{yU} = \sqrt{\bar{r}_{xx'}}$. We then must also compute the average sampling variances of r_{xT_i} and r_{yU_i} between studies. These sampling variance of these correlations can be computed the same way as a pearson correlation:

$$\sigma^2_{r_{xT}i} \approx \frac{(1 - \bar{r}^2_{xT})^2}{n_i - 1}$$

$$\sigma^2_{r_{yU}i} \approx \frac{(1 - \bar{r}^2_{yU})^2}{n_i - 1}$$

Then weighted average of these sampling variances is

$$\sigma^2_{r_{xT}} = \frac{\sum_{i=1}^{k} w_i r_{xT_i}}{\sum_{i=1}^{k} w_i}$$

$$\sigma^2_{r_{yU}} = \frac{\sum_{i=1}^{k} w_i r_{yU_i}}{\sum_{i=1}^{k} w_i}$$

Now that we have the point-estimate of the population observed correlation, the variance of observed population correlations, the sampling variance of observed correlations, and the sampling variance of the square root of the reliability for x and y, we can now attempt to correct the point-estimate and the variance of population correlations.

12.4.1.0.1 * Correcting Using Summary Values

First, we can start by correcting the overall point-estimate for the observed population correlation in order to remove bias due to measurement error. Recall from chapter 5 the correction formula:

$$\hat{\bar{\rho}} = \frac{\hat{\bar{\rho}}_o}{\bar{A}} = \frac{\hat{\bar{\rho}}_o}{\bar{r}_{xT}\bar{r}_{yU}} = \frac{\hat{\bar{\rho}}_o}{\sqrt{\bar{r}_{xx'}}\sqrt{\bar{r}_{yy'}}}$$

Note that the artifact biasing factor, A, is the product of the two sources of attenuation. Correcting the variance in observed population correlations ($\sigma^2_{\rho_o}$), so that it is accurately estimating the variance of true population effect sizes

(σ^2_ρ), we must use a Taylor series approximation. This formula can become fairly complex with more types of artifacts involved. The taylor series approximation is for estimating specifically the amount of sampling variance within the correction factor we apply to the observed correlation. The first step is lay out our attenuation formula (the equation where observed effect size is on the left side of the equation and the artifact values and true effect size is on the right hand side of the equation). In the case of correcting only for measurement error, the attenuation formula is relatively simple

$$\hat{\bar{\rho}}_o = \hat{\bar{\rho}} \cdot \bar{r}_{xT} \cdot \bar{r}_{yU}$$

For the taylor series approximation, we want to first find the partial derivitive with respect to each artifact component:

$$B_{r_{xT}} = \frac{\partial}{\partial r_{xT}}(\hat{\bar{\rho}} \cdot \bar{r}_{xT} \cdot \bar{r}_{yU}) = \hat{\bar{\rho}} \cdot \bar{r}_{yU}$$
$$B_{r_{yU}} = \frac{\partial}{\partial r_{yU}}(\hat{\bar{\rho}} \cdot \bar{r}_{xT} \cdot \bar{r}_{yU}) = \hat{\bar{\rho}} \cdot \bar{r}_{xT}$$

The variance due to artifacts is then approximately,

$$\sigma^2_A \approx B^2_{r_{xT}} \sigma^2_{r_{xT}} + B^2_{r_{yU}} \sigma^2_{r_{yU}}$$

Now we can approximate the variance in true population correlations,

$$\sigma^2_\rho = \frac{\sigma^2_{\rho_o} - \sigma^2_A}{\bar{A}^2}$$

Where the artifact biasing factor is: $\bar{A} = \bar{r}_{xT} \cdot \bar{r}_{yU}$. See the supplementary materials of Dahlke and Wiernik (2020) for detailed Taylor series approximation derivations for the immensely more complicated bivariate indirect range restriction plus measurement error correction.

12.4.2 Applied Example in R

Lets conduct an artifact distribution correction meta-analysis in R, instead using data from the meta-analysis by McDaniel et al. (1994). This dataset contains correlations between employment interviews and job performance. This data set has a lot of missing values for reliability coefficients and u-ratios which might suggest that the artifact distribution approach is a better choice compared to the individual correction approach. We can conduct a meta-analysis correcting for univariate indirect range restriction and measurement error in both job performance and employment interviews. The attenuation formula will be important for calculating the Taylro series approximation can be defined as

$$\bar{\rho}_o = \bar{\rho}\sqrt{\bar{r}^2_{o_i} + \frac{\bar{u}^2_{x_i}\bar{r}_{xx'_i}(\bar{r}_{xx'_i}\bar{r}_{yy'_i} - \bar{r}^2_{o_i})}{1 - \bar{u}^2_{x_i}(1 - \bar{r}_{xx'_i})}}$$

Instead of conducting a taylor series approximation by hand, we will simply use the `psychmeta` package to perform the artifact distribution meta-analysis. The function `ma_r_ad` is designed to conduct an artifact distribution meta-analysis on correlation coefficients. The function also reports the bare-bones model allowing us to compare the corrected estimates to the uncorrected.

```
# Load in packages (we need the development version of psychmeta)
# install.packages("devtools")
# devtools::install_github("psychmeta/psychmeta")
library(psychmeta)

# obtain artifact values
rxx <- data_r_roth_2015$rxxi
ux <- data_r_roth_2015$ux
ro <- data_r_roth_2015$rxyi
n <- data_r_roth_2015$n
k <- length(ro)

# compute barebones meta-analysis
ma_obj <- ma_r_bb(r = rxyi,
                  n = n,
                  correct_bias = FALSE,
                  wt_type = "REML",
                  data = data_r_mcdaniel_1994)

# construct artifact distribution for x
ad_obj_x <- create_ad(ad_type = "tsa",
                      mean_rxxi = data_r_mcdaniel_1994$Mrxxi[1],
                      var_rxxi = data_r_mcdaniel_1994$SDrxxi[1]^2,
                      ux = data_r_mcdaniel_1994$ux,
                      wt_ux = data_r_mcdaniel_1994$`ux frequency`)

# construct artifact distribution for y
ad_obj_y <- create_ad(ad_type = "tsa",
                      rxxi = data_r_mcdaniel_1994$ryyi,
                      wt_rxxi = data_r_mcdaniel_1994$`ryyi frequency`)

# compute artifact-distribution meta-analysis, correcting for measurement error only
mdl_ad <- ma_r_ad(ma_obj = ma_obj,
                  ad_obj_x = ad_obj_x,
                  ad_obj_y = ad_obj_y,
```

```
                  correction_method = "meas")


# summary table of meta-analysis
summary_stats <- data.frame(
  type = c('Artifact Distribution', 'Bare-Bones'),
  k = c(mdl_ad$meta_tables$`analysis_id: 1`$artifact_distribution$true_score$k,mdl_ad$meta_tab
  n = c(mdl_ad$meta_tables$`analysis_id: 1`$artifact_distribution$true_score$N,mdl_ad$meta_tab
  mean_rho = c(mdl_ad$meta_tables$`analysis_id: 1`$artifact_distribution$true_score$mean_rho,m
  SE = c(mdl_ad$meta_tables$`analysis_id: 1`$artifact_distribution$true_score$se_r_c,mdl_ad$me
  SD_rho = c(mdl_ad$meta_tables$`analysis_id: 1`$artifact_distribution$true_score$sd_rho,0))

summary_stats
```

```
                  type   k     n  mean_rho         SE    SD_rho
1 Artifact Distribution 160 25244 0.3201828 0.02108407 0.1985571
2            Bare-Bones 160 25244 0.2205043 0.01452023 0.0000000
```

We can also obtain credibility intervals by using the `credibility` function in the `psychmeta` package. The interval defaults to 80% intervals, however we can change that to 90% by inputting .90 into the `cred_level` argument.

```
credibility(mean = summary_stats$mean_rho[1],
            sd = summary_stats$SD_rho[1],
            cred_method = "norm",
            cred_level = .90)
```

```
        CR_LL_90  CR_UL_90
[1,] -0.006414571 0.6467802
```

Lets compare these results to the bare-bones model. In `psychmeta` the bare-bones model can be conduced using `ma_r_bb`. However, the `ma_r_ic` function also reports the bare-bones results as well. Therefore we can just extract the necessary statistics from the model.

Part III

References

References

Aguinis, Herman, Charles A Pierce, and Steven A Culpepper. 2009. "Scale Coarseness as a Methodological Artifact," September.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2010. "A Basic Introduction to Fixed-Effect and Random-Effects Models for Meta-Analysis." *Research Synthesis Methods* 1 (2): 97–111. https://doi.org/10.1002/jrsm.12.

Borsboom, Denny, and Gideon J Mellenbergh. 2002. "True Scores, Latent Variables, and Constructs: A Comment on Schmidt and Hunter."

Borsboom, Denny, Gideon J. Mellenbergh, and Jaap Van Heerden. 2004. "The Concept of Validity." *Psychological Review* 111 (4): 1061–71. https://doi.org/10.1037/0033-295X.111.4.1061.

Brannick, Michael T., Sean M. Potter, Bryan Benitez, and Scott B. Morris. 2019. "Bias and Precision of Alternate Estimators in Meta-Analysis: Benefits of Blending Schmidt-Hunter and Hedges Approaches." *Organizational Research Methods* 22 (2): 490–514. https://doi.org/10.1177/1094428117741966.

Bravais, A. 1844. *Analyse mathématique sur les probabilités des erreurs de situation d'un point*. Impr. Royale.

Callender, John C., and H. G. Osburn. 1980. "Development and Test of a New Model for Validity Generalization." *Journal of Applied Psychology* 65 (5): 543–58. https://doi.org/10.1037/0021-9010.65.5.543.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

———. 2013. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52 (4): 281–302. https://doi.org/10.1037/h0040957.

Dahlke, Jeffrey A., and Brenton M. Wiernik. 2019. "Psychmeta: An R Package for Psychometric Meta-Analysis." *Applied Psychological Measurement* 43 (5): 415–16. https://doi.org/10.1177/0146621618795933.

———. 2020. "Not Restricted to Selection Research: Accounting for Indirect Range Restriction in Organizational Research." *Organizational Research Methods* 23 (4): 717–49. https://doi.org/10.1177/1094428119859398.

DerSimonian, Rebecca, and Raghu N. Kacker. 2007. "Random-Effects Model

for Meta-Analysis of Clinical Trials: An Update." *NIST* 28 (January): 105–14. https://www.nist.gov/publications/random-effects-model-meta-analysis-clinical-trials-update.

Galton, Francis. 1907. "Vox Populi." *Nature* 75 (1949): 450–51. https://doi.org/10.1038/075450a0.

Haertel, Edward H. 2006. "3. Reliability." In, 4th ed.

Hedges, Larry V. 1981. "Distribution Theory for Glass's Estimator of Effect Size and Related Estimators." *Journal of Educational Statistics* 6 (2): 107–28. https://doi.org/10.3102/10769986006002107.

———. 1989. "An Unbiased Correction for Sampling Error in Validity Generalization Studies." *Journal of Applied Psychology* 74 (3): 469–77. https://doi.org/10.1037/0021-9010.74.3.469.

Hedges, Larry V., and Ingram Olkin. 2014. *Statistical Methods for Meta-Analysis*. Academic press. https://books.google.com/books?hl=en&lr=&id=7GviBQAAQBAJ&oi=fnd&pg=PP1&dq=info:e6P1zfh2T6QJ:scholar.google.com&ots=Dx-YqN6_9B&sig=-39HgbYdWPp_BwSTzA9cRODs2Q0.

Hedges, Larry V., and Jack L. Vevea. 1998. "Fixed- and Random-Effects Models in Meta-Analysis." *Psychological Methods* 3 (4): 486–504. https://doi.org/10.1037/1082-989X.3.4.486.

Hunter, John E., and Frank L. Schmidt. 1990. *Methods of meta-analysis: correcting error and bias in research findings*. Newbury Park: Sage Publications.

Hunter, John, and Frank Schmidt. 1990. "Dichotomization of Continuous Variables: The Implications for Meta-Analysis." *Journal of Applied Psychology* 75 (June): 334–49. https://doi.org/10.1037/0021-9010.75.3.334.

Johnson, Blair T., Brian Mullen, and Eduardo Salas. 1995. "Comparison of Three Major Meta-Analytic Approaches." *Journal of Applied Psychology* 80 (1): 94–106. https://doi.org/10.1037/0021-9010.80.1.94.

Kelley, Truman Lee. 1927. *Interpretation of Educational Measurements*. World Book Company.

Laird, Nan M., and Frederick Mosteller. 1990. "Some Statistical Methods for Combining Experimental Results." *International Journal of Technology Assessment in Health Care* 6 (1): 5–30. https://doi.org/10.1017/S0266462300008916.

Maxwell, Scott, and Harold Delaney. 1993. "Bivariate Median Splits and Spurious Statistical Significance." *Psychological Bulletin* 113 (January): 181–90. https://doi.org/10.1037/0033-2909.113.1.181.

McDaniel, Michael A., Deborah L. Whetzel, Frank L. Schmidt, and Steven D. Maurer. 1994. "The Validity of Employment Interviews: A Comprehensive Review and Meta-Analysis." *Journal of Applied Psychology* 79 (4): 599–616. https://doi.org/10.1037/0021-9010.79.4.599.

Mendoza, Jorge L., and Michael Mumford. 1987. "Corrections for Attenuation and Range Restriction on the Predictor." *Journal of Educational Statistics* 12 (3): 282–93. https://doi.org/10.3102/10769986012003282.

Morris, Scott, Rebecca Daisley, Megan Wheeler, and Peggy Boyer. 2014. "A Meta-Analysis of the Relationship Between Individual Assessments and Job Performance." *The Journal of Applied Psychology* 100 (May). https://doi.

org/10.1037/a0036938.

Olkin, Ingram, and John W. Pratt. 1958. "Unbiased Estimation of Certain Correlation Coefficients." *The Annals of Mathematical Statistics* 29 (1): 201–11. https://www.jstor.org/stable/2237306.

Pearson, Karl. 1903. "I. Mathematical Contributions to the Theory of Evolution. —XI. On the Influence of Natural Selection on the Variability and Correlation of Organs." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 200 (321-330): 1–66. https://doi.org/10.1098/rsta.1903.0001.

———. 1913. "On the Probable Error of a Coefficient of Correlation as Found from a Fourfold Table." *Biometrika* 9 (1/2): 22–33. https://doi.org/10.2307/2331798.

Peters, Charles C., and Walter R. Van Voorhis. 1940. "Further Methods of Correlation." In, 362–403. New York, NY, US: McGraw-Hill Book Company. https://doi.org/10.1037/13596-013.

"Psych: Procedures for Personality and Psychological Research." 2017. https://CRAN.R-project.org/package=psych.

Roth, Bettina. 2015. "Intelligence and School Grades: A Meta-Analysis."

Sackett, Paul R., and Hyuckseung Yang. 2000. "Correction for Range Restriction: An Expanded Typology." *Journal of Applied Psychology* 85 (1): 112–18. https://doi.org/10.1037/0021-9010.85.1.112.

Schmidt, Frank, and John Hunter. 1977. "Development of a General Solution to the Problem of Validity Generalization." *Journal of Applied Psychology* 62 (October): 529–40. https://doi.org/10.1037/0021-9010.62.5.529.

Spearman, C. 1904. "The Proof and Measurement of Association Between Two Things." *International Journal of Epidemiology* 39 (5): 1137–50. https://doi.org/10.1093/ije/dyq191.

Taylor, Erwin K., and Thomas Griess. 1976. "The Missing Middle in Validation Research." *Personnel Psychology* 29 (1): 5–11. https://doi.org/10.1111/j.1744-6570.1976.tb00397.x.

Viechtbauer, Wolfgang. 2010. "Conducting meta-analyses in R with the metafor package." *Journal of Statistical Software* 36 (3): 1–48. https://doi.org/10.18637/jss.v036.i03.

———. n.d. "Fixed-Effects and Random-Effects Models in Meta-Analysis." https://wviechtb.github.io/metafor/index.html.

Viswesvaran, Chockalingam, Deniz S. Ones, Frank L. Schmidt, Huy Le, and In-Sue Oh. 2014. "Measurement Error Obfuscates Scientific Knowledge: Path to Cumulative Knowledge Requires Corrections for Unreliability and Psychometric Meta-Analyses." *Industrial and Organizational Psychology* 7 (4): 507–18. https://doi.org/10.1017/S1754942600006799.

Vos, Paul, and Don Holbert. 2022. "Frequentist Statistical Inference Without Repeated Sampling." *Synthese* 200 (2): 89. https://doi.org/10.1007/s11229-022-03560-x.

Wechsler, David. 2008. *Wechsler Adult Intelligence Scale–Fourth Edition*. 4th ed. https://doi.org/10.1037/t15169-000.

Whitener, Ellen M. 1990. "Confusion of Confidence Intervals and Credibility

Intervals in Meta-Analysis." *Journal of Applied Psychology* 75 (3): 315–21. https://doi.org/10.1037/0021-9010.75.3.315.

Wiernik, Brenton M., and Jeffrey A. Dahlke. 2020. "Obtaining Unbiased Results in Meta-Analysis: The Importance of Correcting for Statistical Artifacts." *Advances in Methods and Practices in Psychological Science* 3 (1): 94–123. https://doi.org/10.1177/2515245919885611.

Wylie, Peter B. 1976. "Effects of Coarse Grouping and Skewed Marginal Distributions on the Pearson Product Moment Correlation Coefficient." *Educational and Psychological Measurement* 36 (1): 1–7. https://doi.org/10.1177/001316447603600101.