Results Reliability Generalization and Variance Decomposition

Lukas Beinhauer

paths\_rel <- list.files(file.path("Data/Reliability Estimates"), full.names = TRUE)  
  
paths\_alpha <- paths\_rel[grep("\_Alpha.csv$", paths\_rel)]  
  
alpha\_estimates.list <- lapply(paths\_alpha, FUN = function(x){read.csv(x)})  
  
  
  
Alpha\_rma.list <- readRDS(file.path("Data/Shiny Data/Alpha\_rma.list.RData"))  
Bonett.Alpha\_rma.list <- readRDS(file.path("Data/Shiny Data/Bonett.Alpha\_rma.list.RData"))  
  
varT\_est.list <- readRDS(file.path("Notes/bootstrapped\_varT.RData"))  
varE\_est.list <- readRDS(file.path("Notes/bootstrapped\_varE.RData"))  
  
varT\_rma.list <- readRDS(file.path("Notes/bootstrapped\_varT\_rma.RData"))  
varE\_rma.list <- readRDS(file.path("Notes/bootstrapped\_varE\_rma.RData"))

s\_Caciop\_Alpha <- summary(alpha\_estimates.list[[1]]$Reliability)  
  
r.dec <- 3  
  
s\_Caciop\_Alpha\_rma <- Alpha\_rma.list[[1]]  
s\_Caciop\_B.Alpha\_rma <- Bonett.Alpha\_rma.list[[1]]  
  
s\_Caciop\_varT <- summary(varT\_est.list[[1]]$var.est)  
s\_Caciop\_varE <- summary(varE\_est.list[[1]]$var.est)  
  
s\_Caciop\_varT\_rma <- varT\_rma.list[[1]]  
s\_Caciop\_varE\_rma <- varE\_rma.list[[1]]  
  
  
  
  
  
s\_ShnabelRPP\_Alpha <- summary(alpha\_estimates.list[[9]]$Reliability)  
  
s\_ShnabelRPP\_Alpha\_rma <- Alpha\_rma.list[[9]]  
s\_ShnabelRPP\_B.Alpha\_rma <- Bonett.Alpha\_rma.list[[9]]  
  
s\_ShnabelRPP\_varT <- summary(varT\_est.list[[9]]$var.est)  
s\_ShnabelRPP\_varE <- summary(varE\_est.list[[9]]$var.est)  
  
s\_ShnabelRPP\_varT\_rma <- varT\_rma.list[[9]]  
s\_ShnabelRPP\_varE\_rma <- varE\_rma.list[[9]]  
  
  
  
  
s\_ShnabelRev\_Alpha <- summary(alpha\_estimates.list[[8]]$Reliability)  
  
s\_ShnabelRev\_Alpha\_rma <- Alpha\_rma.list[[8]]  
s\_ShnabelRev\_B.Alpha\_rma <- Bonett.Alpha\_rma.list[[8]]  
  
s\_ShnabelRev\_varT <- summary(varT\_est.list[[8]]$var.est)  
s\_ShnabelRev\_varE <- summary(varE\_est.list[[8]]$var.est)  
  
s\_ShnabelRev\_varT\_rma <- varT\_rma.list[[8]]  
s\_ShnabelRev\_varE\_rma <- varE\_rma.list[[8]]  
  
  
  
  
  
s\_HexHH\_Alpha <- summary(alpha\_estimates.list[[6]]$Reliability)  
  
s\_HexHH\_Alpha\_rma <- Alpha\_rma.list[[6]]  
s\_HexHH\_B.Alpha\_rma <- Bonett.Alpha\_rma.list[[6]]  
  
s\_HexHH\_varT <- summary(varT\_est.list[[6]]$var.est)  
s\_HexHH\_varE <- summary(varE\_est.list[[6]]$var.est)  
  
s\_HexHH\_varT\_rma <- varT\_rma.list[[6]]  
s\_HexHH\_varE\_rma <- varE\_rma.list[[6]]  
  
  
  
  
s\_HexEM\_Alpha <- summary(alpha\_estimates.list[[4]]$Reliability)  
  
s\_HexEM\_Alpha\_rma <- Alpha\_rma.list[[4]]  
s\_HexEM\_B.Alpha\_rma <- Bonett.Alpha\_rma.list[[4]]  
  
s\_HexEM\_varT <- summary(varT\_est.list[[4]]$var.est)  
s\_HexEM\_varE <- summary(varE\_est.list[[4]]$var.est)  
  
s\_HexEM\_varT\_rma <- varT\_rma.list[[4]]  
s\_HexEM\_varE\_rma <- varE\_rma.list[[4]]  
  
  
  
  
  
s\_HexEX\_Alpha <- summary(alpha\_estimates.list[[5]]$Reliability)  
  
s\_HexEX\_Alpha\_rma <- Alpha\_rma.list[[5]]  
s\_HexEX\_B.Alpha\_rma <- Bonett.Alpha\_rma.list[[5]]  
  
s\_HexEX\_varT <- summary(varT\_est.list[[5]]$var.est)  
s\_HexEX\_varE <- summary(varE\_est.list[[5]]$var.est)  
  
s\_HexEX\_varT\_rma <- varT\_rma.list[[5]]  
s\_HexEX\_varE\_rma <- varE\_rma.list[[5]]  
  
  
  
  
  
s\_HexAG\_Alpha <- summary(alpha\_estimates.list[[2]]$Reliability)  
  
s\_HexAG\_Alpha\_rma <- Alpha\_rma.list[[2]]  
s\_HexAG\_B.Alpha\_rma <- Bonett.Alpha\_rma.list[[2]]  
  
s\_HexAG\_varT <- summary(varT\_est.list[[2]]$var.est)  
s\_HexAG\_varE <- summary(varE\_est.list[[2]]$var.est)  
  
s\_HexAG\_varT\_rma <- varT\_rma.list[[2]]  
s\_HexAG\_varE\_rma <- varE\_rma.list[[2]]  
  
  
  
  
  
s\_HexCO\_Alpha <- summary(alpha\_estimates.list[[3]]$Reliability)  
  
s\_HexCO\_Alpha\_rma <- Alpha\_rma.list[[3]]  
s\_HexCO\_B.Alpha\_rma <- Bonett.Alpha\_rma.list[[3]]  
  
s\_HexCO\_varT <- summary(varT\_est.list[[3]]$var.est)  
s\_HexCO\_varE <- summary(varE\_est.list[[3]]$var.est)  
  
s\_HexCO\_varT\_rma <- varT\_rma.list[[3]]  
s\_HexCO\_varE\_rma <- varE\_rma.list[[3]]  
  
  
  
  
  
s\_HexOX\_Alpha <- summary(alpha\_estimates.list[[7]]$Reliability)  
  
s\_HexOX\_Alpha\_rma <- Alpha\_rma.list[[7]]  
s\_HexOX\_B.Alpha\_rma <- Bonett.Alpha\_rma.list[[7]]  
  
s\_HexOX\_varT <- summary(varT\_est.list[[7]]$var.est)  
s\_HexOX\_varE <- summary(varE\_est.list[[7]]$var.est)  
  
s\_HexOX\_varT\_rma <- varT\_rma.list[[7]]  
s\_HexOX\_varE\_rma <- varE\_rma.list[[7]]

## Results Reliability Generalization and Variance Decomposition

### Cacioppo

Data set 1 concerned the scale of Need for Cognition (Cacioppo & Petty, 1982), consisting of 6 Likert-stye items (5-point scale). Measured across 21 labs in Ebersole et al. (2016), using Cronbach’s Alpha, score reliability varied between 0.406 and 0.819, with a mean value of 0.586 and median value 0.578. Performing a random-effects meta-analysis on the untransformed Cronbach’s Alpha estimates, we find a meta-analytic mean estimate of 0.601 (0.558 : 0.644). Additionally, we find a significant degree of heterogeneity in score reliability, at Q(20) = 192.728, p <.001. In terms of I^2, heterogeneity is 77.075%, which, according to XXX constitutes a large level of heterogeneity. In terms of tau, we find a standard deviation of 0.082 in true values of Cronbach’s Alpha. Similarly, when assessing Bonett-transformed Cronbach’s Alpha estimates, we find a meta-analytic mean of -0.906. Back-transformed this constitutes a Cronbach’s Alpha of 0.596 (CI: 0.549 : 0.638) (difference in MA-est. emerges at the 3rd decimal). Heterogeneity in Bonett-transformed estimates of Cronbach’s Alpha is also of statistically significant degree: Q(20) = 80.554, p <.001. Standardized heterogeneity estimates of I^2 = 77.298% and H^2 = 4.405 display highly similar amounts of heterogeneity. When using McDonald’s Omega to estimate score reliability, the results change only to a very small degree. Therefore, estimates using McDonald’s Omega can be found in the supplemental materials.

Also, we use Cronbach’s Alpha to decompose the observed variance into its true and error score variance components:

| {} | True Score Variance | Error Score Variance |
| --- | --- | --- |
| Min | 0.101 | 0.111 |
| Q1 | 0.153 | 0.126 |
| Median | 0.169 | 0.13 |
| Q3 | 0.203 | 0.131 |
| Max | 0.238 | 0.134 |
| Mean | 0.555 | 0.15 |

Performing a random-effects meta-analysis on the components directly leads to:

| {} | True Score Variance | Error Score Variance |
| --- | --- | --- |
| Meta-Analytic Mean Estimate | 0.201 | 0.127 |
| Confidence Interval | 0.159 : 0.243 | 0.124 : 0.131 |
| Heterogeneity tau | 0.089 | 0.003 |
| Heterogeneity I^2 | 83.888 | 9.444 |
| Heterogeneity H^2 | 6.206 | 1.104 |
| Q(20) | 129.574 | 22.735 |
| p-value QE | <.001 | 0.302 |

### Shnabel Sense of Power (RPP)

Data set 2 concerned the scale of Sense of Power (Shnabel & Nadler, 2008), consisting of 3 Likert-stye items (7-point scale). Measured across 8 labs in Baranski et al. (2020), using Cronbach’s Alpha, specifically for those labs following the RPP-protocol, score reliability varied between 0.723 and 0.828, with a mean value of 0.763 and median value 0.752. Performing a random-effects meta-analysis on the untransformed Cronbach’s Alpha estimates, we find a meta-analytic mean estimate of 0.769 (0.739 : 0.799). Additionally, we find a significant degree of heterogeneity in score reliability, at Q(7) = 15.851, p = 0.027. In terms of I^2, heterogeneity is 55.557%, which, according to XXX constitutes a medium level of heterogeneity. In terms of tau, we find a standard deviation of 0.031 in true values of Cronbach’s Alpha. Similarly, when assessing Bonett-transformed Cronbach’s Alpha estimates, we find a meta-analytic mean of -1.458. Back-transformed this constitutes a Cronbach’s Alpha of 0.767 (CI: 0.736 : 0.795) (difference in MA-est. emerges at the 3rd decimal). Heterogeneity in Bonett-transformed estimates of Cronbach’s Alpha is also of statistically significant degree: Q(7) = 16.013, p = 0.025. Standardized heterogeneity estimates of I^2 = 56.122% and H^2 = 2.279 display highly similar amounts of heterogeneity.

Also, we use Cronbach’s Alpha to decompose the observed variance into its true and error score variance components:

| {} | True Score Variance | Error Score Variance |
| --- | --- | --- |
| Min | 0.67 | 0.204 |
| Q1 | 0.755 | 0.256 |
| Median | 0.81 | 0.263 |
| Q3 | 0.864 | 0.262 |
| Max | 0.935 | 0.277 |
| Mean | 1.236 | 0.296 |

Performing a random-effects meta-analysis on the components directly leads to:

| {} | True Score Variance | Error Score Variance |
| --- | --- | --- |
| Meta-Analytic Mean Estimate | 0.829 | 0.26 |
| Confidence Interval | 0.726 : 0.931 | 0.238 : 0.281 |
| Heterogeneity tau | 0.063 | 0.025 |
| Heterogeneity I^2 | 18.662 | 66.311 |
| Heterogeneity H^2 | 1.229 | 2.968 |
| Q(7) | 9.723 | 24.052 |
| p-value QE | 0.205 | 0.001 |

### Shnabel Sense of Power (Rev)

Comparing those results to the labs employing the revised protocol, score reliability varied between 0.603 and 0.757, with a mean value of 0.696 and median value 0.701. Performing a random-effects meta-analysis on the untransformed Cronbach’s Alpha estimates, we find a meta-analytic mean estimate of 0.706 (0.672 : 0.74). Additionally, we find no significant degree of heterogeneity in score reliability, at Q(7) = 13.045, p = 0.071. In terms of I^2, heterogeneity is 47.571%, which, according to XXX constitutes a low to medium level of heterogeneity. In terms of tau, we find a standard deviation of 0.033 in true values of Cronbach’s Alpha. Similarly, when assessing Bonett-transformed Cronbach’s Alpha estimates, we find a meta-analytic mean of -1.209. Back-transformed this constitutes a Cronbach’s Alpha of 0.701 (CI: 0.663 : 0.736) (difference in MA-est. emerges at the 3rd decimal). Heterogeneity in Bonett-transformed estimates of Cronbach’s Alpha is of statistically significant degree: Q(7) = 14.839, p = 0.038. Standardized heterogeneity estimates of I^2 = 53.705% and H^2 = 2.16 display highly similar amounts of heterogeneity.

Also, we use Cronbach’s Alpha to decompose the observed variance into its true and error score variance components:

| {} | True Score Variance | Error Score Variance |
| --- | --- | --- |
| Min | 0.541 | 0.231 |
| Q1 | 0.604 | 0.289 |
| Median | 0.679 | 0.304 |
| Q3 | 0.705 | 0.301 |
| Max | 0.821 | 0.322 |
| Mean | 0.914 | 0.355 |

Performing a random-effects meta-analysis on the components directly leads to:

| {} | True Score Variance | Error Score Variance |
| --- | --- | --- |
| Meta-Analytic Mean Estimate | 0.689 | 0.299 |
| Confidence Interval | 0.584 : 0.794 | 0.27 : 0.328 |
| Heterogeneity tau | 0.101 | 0.036 |
| Heterogeneity I^2 | 45.728 | 75.265 |
| Heterogeneity H^2 | 1.843 | 4.043 |
| Q(7) | 12.528 | 36.801 |
| p-value QE | 0.084 | <.001 |

### HEXACO

#### Cronbach’s Alpha

| Scale | HH | EM | EX | AG | CO |
| --- | --- | --- | --- | --- | --- |
| Meta-analytic mean estimate | 0.706 | 0.767 | 0.798 | 0.735 | 0.756 |
| Confidence Interval | 0.68 : 0.731 | 0.746 : 0.788 | 0.773 : 0.822 | 0.712 : 0.758 | 0.727 : 0.785 |
| Heterogeneity tau | 0.039 | 0.033 | 0.046 | 0.036 | 0.054 |
| Heteroeneity I^2 | 50.048 | 52.721 | 75.37 | 50.528 | 75.125 |
| Heterogeneity H^2 | 2.002 | 2.115 | 4.06 | 2.021 | 4.02 |
| Q(18) | 39.346 | 44.673 | 64.587 | 44.159 | 57.724 |
| P-value QE | 0.003 | <.001 | <.001 | <.001 | <.001 |

| Scale | HH | EM | EX | AG | CO |
| --- | --- | --- | --- | --- | --- |
| Min | 0.433 | 0.461 | 0.65 | 0.377 | 0.521 |
| Q1 | 0.662 | 0.745 | 0.75 | 0.709 | 0.701 |
| Median | 0.715 | 0.763 | 0.803 | 0.739 | 0.784 |
| Q3 | 0.689 | 0.75 | 0.788 | 0.716 | 0.741 |
| Max | 0.745 | 0.803 | 0.836 | 0.763 | 0.796 |
| Mean | 0.773 | 0.826 | 0.879 | 0.801 | 0.826 |

##### Back-transformed Bonett-estimates (meta-analytic results)

| Scale | HH | EM | EX | AG | CO |
| --- | --- | --- | --- | --- | --- |
| Meta-analytic mean estimate | 0.696 | 0.759 | 0.796 | 0.726 | 0.751 |
| Confidence Interval | 0.663 : 0.726 | 0.729 : 0.786 | 0.769 : 0.82 | 0.692 : 0.757 | 0.717 : 0.782 |
| Heteroeneity I^2 | 66.793 | 74.745 | 77.198 | 73.843 | 78.945 |
| Heterogeneity H^2 | 3.011 | 3.96 | 4.386 | 3.823 | 4.75 |
| Q(18) | 55.265 | 74.094 | 78.648 | 69.962 | 90.827 |
| P-value QE | <.001 | <.001 | <.001 | <.001 | <.001 |

#### True Score Variance

| Scale | HH | EM | EX | AG | CO |
| --- | --- | --- | --- | --- | --- |
| Min | 0.077 | 0.086 | 0.144 | 0.052 | 0.102 |
| Q1 | 0.204 | 0.287 | 0.272 | 0.239 | 0.219 |
| Median | 0.258 | 0.319 | 0.307 | 0.256 | 0.268 |
| Q3 | 0.241 | 0.318 | 0.337 | 0.253 | 0.268 |
| Max | 0.279 | 0.391 | 0.409 | 0.292 | 0.326 |
| Mean | 0.336 | 0.44 | 0.525 | 0.383 | 0.43 |

| Scale | HH | EM | EX | AG | CO |
| --- | --- | --- | --- | --- | --- |
| Meta-analytic mean estimate | 0.235 | 0.308 | 0.327 | 0.245 | 0.26 |
| Confidence Interval | 0.204 : 0.265 | 0.265 : 0.351 | 0.283 : 0.372 | 0.208 : 0.282 | 0.223 : 0.296 |
| Heterogeneity tau | 0.049 | 0.08 | 0.083 | 0.069 | 0.068 |
| Heteroeneity I^2 | 52.902 | 72.703 | 72.339 | 76.582 | 74.663 |
| Heterogeneity H^2 | 2.123 | 3.663 | 3.615 | 4.27 | 3.947 |
| Q(18) | 39.551 | 90.652 | 64.645 | 120.464 | 80.405 |
| P-value QE | 0.002 | <.001 | <.001 | <.001 | <.001 |

#### Error Score Variance

| Scale | HH | EM | EX | AG | CO |
| --- | --- | --- | --- | --- | --- |
| Min | 0.085 | 0.086 | 0.071 | 0.079 | 0.074 |
| Q1 | 0.095 | 0.092 | 0.075 | 0.085 | 0.078 |
| Median | 0.104 | 0.095 | 0.081 | 0.09 | 0.084 |
| Q3 | 0.102 | 0.097 | 0.084 | 0.092 | 0.086 |
| Max | 0.108 | 0.102 | 0.094 | 0.096 | 0.089 |
| Mean | 0.12 | 0.114 | 0.104 | 0.112 | 0.117 |

| Scale | HH | EM | EX | AG | CO |
| --- | --- | --- | --- | --- | --- |
| Meta-analytic mean estimate | 0.102 | 0.097 | 0.083 | 0.092 | 0.085 |
| Confidence Interval | 0.097 : 0.106 | 0.093 : 0.1 | 0.078 : 0.088 | 0.088 : 0.095 | 0.081 : 0.09 |
| Heterogeneity tau | 0.008 | 0.005 | 0.009 | 0.007 | 0.009 |
| Heteroeneity I^2 | 69.422 | 49.982 | 80.355 | 65.765 | 77.466 |
| Heterogeneity H^2 | 3.27 | 1.999 | 5.09 | 2.921 | 4.438 |
| Q(18) | 59.251 | 35.907 | 80.68 | 50.173 | 68.688 |
| P-value QE | <.001 | 0.007 | <.001 | <.001 | <.001 |

library(tidyverse)

Warning: Paket 'tidyverse' wurde unter R Version 4.2.2 erstellt

── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
✔ ggplot2 3.4.2 ✔ purrr 1.0.1  
✔ tibble 3.2.1 ✔ dplyr 1.1.1  
✔ tidyr 1.3.0 ✔ stringr 1.5.0  
✔ readr 2.1.4 ✔ forcats 1.0.0

Warning: Paket 'ggplot2' wurde unter R Version 4.2.3 erstellt

Warning: Paket 'tibble' wurde unter R Version 4.2.3 erstellt

Warning: Paket 'tidyr' wurde unter R Version 4.2.2 erstellt

Warning: Paket 'readr' wurde unter R Version 4.2.3 erstellt

Warning: Paket 'purrr' wurde unter R Version 4.2.2 erstellt

Warning: Paket 'dplyr' wurde unter R Version 4.2.3 erstellt

Warning: Paket 'stringr' wurde unter R Version 4.2.2 erstellt

Warning: Paket 'forcats' wurde unter R Version 4.2.2 erstellt

── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()

### Simulation Results  
  
vis.df\_empF <- read.csv(file.path("Notes/vis\_df.csv"), sep = " ")  
vis.df\_empT <- read.csv(file.path("Notes/vis\_df\_empT.csv"), sep = " ")  
  
vis.df\_summarised\_empF <- vis.df\_empF %>%   
 group\_by(CVT, CVE, rel) %>%  
 summarise(T\_m = mean(tau\_T, na.rm = T),  
 E\_m = mean(tau\_E, na.rm = T),  
 T\_sd = sd(tau\_T, na.rm = T),  
 E\_sd = sd(tau\_E, na.rm = T),  
 T\_ll = quantile(tau\_T, .025, na.rm = T),  
 T\_ul = quantile(tau\_T, .975, na.rm = T),  
 E\_ll = quantile(tau\_E, .025, na.rm = T),  
 E\_ul = quantile(tau\_E, .975, na.rm = T),  
 VT.MAE\_m = mean(varT.MAE, na.rm = T),  
 VE.MAE\_m = mean(varE.MAE, na.rm = T),  
 VT.MAE\_ll = quantile(varT.MAE, .025, na.rm = T),  
 VT.MAE\_ul = quantile(varT.MAE, .975, na.rm = T),  
 VE.MAE\_ll = quantile(varE.MAE, .025, na.rm = T),  
 VE.MAE\_ul = quantile(varE.MAE, .975, na.rm = T),  
 VT.M\_m = mean(varT.M, na.rm = T),  
 VE.M\_m = mean(varE.M, na.rm = T),  
 VT.M\_ll = quantile(varT.M, .025, na.rm = T),  
 VT.M\_ul = quantile(varT.M, .975, na.rm = T),  
 VE.M\_ll = quantile(varE.M, .025, na.rm = T),  
 VE.M\_ul = quantile(varE.M, .975, na.rm = T),  
 rel.M\_m = mean(rel.M, na.rm = T),  
 rel.M\_ll = quantile(rel.M, .025, na.rm = T),  
 rel.M\_ul = quantile(rel.M, .975, na.rm = T),  
 rel.MAE\_m = mean(rel.MAE, na.rm = T),  
 rel.MAE\_ll = quantile(rel.MAE, .025, na.rm = T),  
 rel.MAE\_ul = quantile(rel.MAE, .975, na.rm = T),  
 tau\_rel\_m = mean(tau\_rel, na.rm = T),  
 tau\_rel\_ll = quantile(tau\_rel, .025, na.rm = T),  
 tau\_rel\_ul = quantile(tau\_rel, .975, na.rm = T),  
 tau\_rel\_cT\_m = mean(tau\_rel\_cT, na.rm = T),  
 tau\_rel\_cT\_ll = quantile(tau\_rel\_cT, .025, na.rm = T),  
 tau\_rel\_cT\_ul = quantile(tau\_rel\_cT, .975, na.rm = T),  
 tau\_rel\_cE\_m = mean(tau\_rel\_cE, na.rm = T),  
 tau\_rel\_cE\_ll = quantile(tau\_rel\_cE, .025, na.rm = T),  
 tau\_rel\_cE\_ul = quantile(tau\_rel\_cE, .975, na.rm = T),  
 tau\_rel\_cTE\_m = mean(tau\_rel\_cTE, na.rm = T),  
 tau\_rel\_cTE\_ll = quantile(tau\_rel\_cTE, .025, na.rm = T),  
 tau\_rel\_cTE\_ul = quantile(tau\_rel\_cTE, .975, na.rm = T))

`summarise()` has grouped output by 'CVT', 'CVE'. You can override using the  
`.groups` argument.

vis.df\_summarised\_empT <- vis.df\_empT %>%   
 group\_by(CVT, CVE, rel) %>%  
 summarise(T\_m = mean(tau\_T, na.rm = T),  
 E\_m = mean(tau\_E, na.rm = T),  
 T\_sd = sd(tau\_T, na.rm = T),  
 E\_sd = sd(tau\_E, na.rm = T),  
 T\_ll = quantile(tau\_T, .025, na.rm = T),  
 T\_ul = quantile(tau\_T, .975, na.rm = T),  
 E\_ll = quantile(tau\_E, .025, na.rm = T),  
 E\_ul = quantile(tau\_E, .975, na.rm = T),  
 VT.MAE\_m = mean(varT.MAE, na.rm = T),  
 VE.MAE\_m = mean(varE.MAE, na.rm = T),  
 VT.MAE\_ll = quantile(varT.MAE, .025, na.rm = T),  
 VT.MAE\_ul = quantile(varT.MAE, .975, na.rm = T),  
 VE.MAE\_ll = quantile(varE.MAE, .025, na.rm = T),  
 VE.MAE\_ul = quantile(varE.MAE, .975, na.rm = T),  
 VT.M\_m = mean(varT.M, na.rm = T),  
 VE.M\_m = mean(varE.M, na.rm = T),  
 VT.M\_ll = quantile(varT.M, .025, na.rm = T),  
 VT.M\_ul = quantile(varT.M, .975, na.rm = T),  
 VE.M\_ll = quantile(varE.M, .025, na.rm = T),  
 VE.M\_ul = quantile(varE.M, .975, na.rm = T),  
 rel.M\_m = mean(rel.M, na.rm = T),  
 rel.M\_ll = quantile(rel.M, .025, na.rm = T),  
 rel.M\_ul = quantile(rel.M, .975, na.rm = T),  
 rel.MAE\_m = mean(rel.MAE, na.rm = T),  
 rel.MAE\_ll = quantile(rel.MAE, .025, na.rm = T),  
 rel.MAE\_ul = quantile(rel.MAE, .975, na.rm = T),  
 tau\_rel\_m = mean(tau\_rel, na.rm = T),  
 tau\_rel\_ll = quantile(tau\_rel, .025, na.rm = T),  
 tau\_rel\_ul = quantile(tau\_rel, .975, na.rm = T),  
 tau\_rel\_cT\_m = mean(tau\_rel\_cT, na.rm = T),  
 tau\_rel\_cT\_ll = quantile(tau\_rel\_cT, .025, na.rm = T),  
 tau\_rel\_cT\_ul = quantile(tau\_rel\_cT, .975, na.rm = T),  
 tau\_rel\_cE\_m = mean(tau\_rel\_cE, na.rm = T),  
 tau\_rel\_cE\_ll = quantile(tau\_rel\_cE, .025, na.rm = T),  
 tau\_rel\_cE\_ul = quantile(tau\_rel\_cE, .975, na.rm = T),  
 tau\_rel\_cTE\_m = mean(tau\_rel\_cTE, na.rm = T),  
 tau\_rel\_cTE\_ll = quantile(tau\_rel\_cTE, .025, na.rm = T),  
 tau\_rel\_cTE\_ul = quantile(tau\_rel\_cTE, .975, na.rm = T))

`summarise()` has grouped output by 'CVT', 'CVE'. You can override using the  
`.groups` argument.

vis.df\_summarized2\_empF <- vis.df\_empF %>%  
 group\_by(CVT, CVE, rel) %>%  
 summarise(bias\_varT.M = mean(varT.M - rel\*10, na.rm = T),  
 bias\_varT.MAE = mean(varT.MAE - rel\*10, na.rm = T),  
 bias\_varE.M = mean(varE.M - (1-rel)\*10, na.rm = T),  
 bias\_varE.MAE = mean(varE.MAE - (1-rel)\*10, na.rm = T),  
 bias\_rel.M = mean(rel.M - rel, na.rm = T),  
 bias\_rel.MAE = mean(rel.MAE-rel, na.rm = T),  
 bias\_tauT = mean(tau\_T - CVT\*rel\*10, na.rm = T),  
 bias\_tauE = mean(tau\_E - CVE\*(1-rel)\*10, na.rm = T),  
 MSE\_varT.M = mean((varT.M - rel\*10)^2, na.rm = T),  
 MSE\_varT.MAE = mean((varT.MAE - rel\*10)^2, na.rm = T),  
 MSE\_varE.M = mean((varE.M - (1-rel)\*10)^2, na.rm = T),  
 MSE\_varE.MAE = mean((varE.MAE - (1-rel)\*10)^2, na.rm = T),  
 MSE\_rel.M = mean((rel.M - rel)^2, na.rm = T),  
 MSE\_rel.MAE = mean((rel.MAE-rel)^2, na.rm = T),  
 MSE\_tauT = mean((tau\_T - CVT\*rel\*10)^2, na.rm = T),  
 MSE\_tauE = mean((tau\_E - CVE\*(1-rel)\*10)^2, na.rm = T),  
 RMSE\_varT.M = sqrt(mean((varT.M - rel\*10)^2, na.rm = T)),  
 RMSE\_varT.MAE = sqrt(mean((varT.MAE - rel\*10)^2, na.rm = T)),  
 RMSE\_varE.M = sqrt(mean((varE.M - (1-rel)\*10)^2, na.rm = T)),  
 RMSE\_varE.MAE = sqrt(mean((varE.MAE - (1-rel)\*10)^2, na.rm = T)),  
 RMSE\_rel.M = sqrt(mean((rel.M - rel)^2, na.rm = T)),  
 RMSE\_rel.MAE = sqrt(mean((rel.MAE-rel)^2, na.rm = T)),  
 RMSE\_tauT = sqrt(mean((tau\_T - CVT\*rel\*10)^2, na.rm = T)),  
 RMSE\_tauE = sqrt(mean((tau\_E - CVE\*(1-rel)\*10)^2, na.rm = T)))

`summarise()` has grouped output by 'CVT', 'CVE'. You can override using the  
`.groups` argument.

vis.df\_summarized2\_empT <- vis.df\_empT %>%  
 group\_by(CVT, CVE, rel) %>%  
 summarise(bias\_varT.M = mean(varT.M - rel\*10, na.rm = T),  
 bias\_varT.MAE = mean(varT.MAE - rel\*10, na.rm = T),  
 bias\_varE.M = mean(varE.M - (1-rel)\*10, na.rm = T),  
 bias\_varE.MAE = mean(varE.MAE - (1-rel)\*10, na.rm = T),  
 bias\_rel.M = mean(rel.M - rel, na.rm = T),  
 bias\_rel.MAE = mean(rel.MAE-rel, na.rm = T),  
 bias\_tauT = mean(tau\_T - CVT\*rel\*10, na.rm = T),  
 bias\_tauE = mean(tau\_E - CVE\*(1-rel)\*10, na.rm = T),  
 MSE\_varT.M = mean((varT.M - rel\*10)^2, na.rm = T),  
 MSE\_varT.MAE = mean((varT.MAE - rel\*10)^2, na.rm = T),  
 MSE\_varE.M = mean((varE.M - (1-rel)\*10)^2, na.rm = T),  
 MSE\_varE.MAE = mean((varE.MAE - (1-rel)\*10)^2, na.rm = T),  
 MSE\_rel.M = mean((rel.M - rel)^2, na.rm = T),  
 MSE\_rel.MAE = mean((rel.MAE-rel)^2, na.rm = T),  
 MSE\_tauT = mean((tau\_T - CVT\*rel\*10)^2, na.rm = T),  
 MSE\_tauE = mean((tau\_E - CVE\*(1-rel)\*10), na.rm = T)^2,  
 RMSE\_varT.M = sqrt(mean((varT.M - rel\*10)^2, na.rm = T)),  
 RMSE\_varT.MAE = sqrt(mean((varT.MAE - rel\*10)^2, na.rm = T)),  
 RMSE\_varE.M = sqrt(mean((varE.M - (1-rel)\*10)^2, na.rm = T)),  
 RMSE\_varE.MAE = sqrt(mean((varE.MAE - (1-rel)\*10)^2, na.rm = T)),  
 RMSE\_rel.M = sqrt(mean((rel.M - rel)^2, na.rm = T)),  
 RMSE\_rel.MAE = sqrt(mean((rel.MAE-rel)^2, na.rm = T)),  
 RMSE\_tauT = sqrt(mean((tau\_T - CVT\*rel\*10)^2, na.rm = T)),  
 RMSE\_tauE = sqrt(mean((tau\_E - CVE\*(1-rel)\*10), na.rm = T)^2)  
 )

`summarise()` has grouped output by 'CVT', 'CVE'. You can override using the  
`.groups` argument.

vis.df2\_empF <- vis.df\_empF %>%  
 mutate(bias\_varT.M = (varT.M - rel\*10),  
 bias\_varT.MAE = (varT.MAE - rel\*10),  
 bias\_varE.M = (varE.M - (1-rel)\*10),  
 bias\_varE.MAE = (varE.MAE - (1-rel)\*10),  
 bias\_rel.M = (rel.M - rel),  
 bias\_rel.MAE = (rel.MAE-rel),  
 bias\_tauT = (tau\_T - CVT\*rel\*10),  
 bias\_tauE = (tau\_E - CVE\*(1-rel)\*10)  
 )  
  
vis.df2\_empT <- vis.df\_empT %>%  
 mutate(bias\_varT.M = (varT.M - rel\*10),  
 bias\_varT.MAE = (varT.MAE - rel\*10),  
 bias\_varE.M = (varE.M - (1-rel)\*10),  
 bias\_varE.MAE = (varE.MAE - (1-rel)\*10),  
 bias\_rel.M = (rel.M - rel),  
 bias\_rel.MAE = (rel.MAE-rel),  
 bias\_tauT = (tau\_T - CVT\*rel\*10),  
 bias\_tauE = (tau\_E - CVE\*(1-rel)\*10)  
 )  
  
  
  
colnames(vis.df\_empT) <- paste0(colnames(vis.df\_empF), "\_empT")  
colnames(vis.df\_summarised\_empT) <- paste0(colnames(vis.df\_summarised\_empF), "\_empT")  
colnames(vis.df\_summarized2\_empT) <- paste0(colnames(vis.df\_summarized2\_empF), "\_empT")  
colnames(vis.df2\_empT) <- paste0(colnames(vis.df2\_empF), "\_empT")  
  
vis.df <- cbind(vis.df\_empF, vis.df\_empT)  
vis.df\_summarised <- cbind(vis.df\_summarised\_empF, vis.df\_summarised\_empT)  
vis.df\_summarized2 <- cbind(vis.df\_summarized2\_empF, vis.df\_summarized2\_empT)  
vis.df2 <- cbind(vis.df2\_empF, vis.df2\_empT)

vis.df\_summarized2

# A tibble: 80 × 54  
# Groups: CVT, CVE [16]  
 CVT CVE rel bias\_varT.M bias\_varT.MAE bias\_varE.M bias\_varE.MAE  
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
 1 0 0 0.1 0.00215 -0.297 -0.00207 -0.0376   
 2 0 0 0.3 0.000652 -0.310 -0.00273 -0.0303   
 3 0 0 0.5 -0.000509 -0.319 0.000297 -0.0198   
 4 0 0 0.7 0.00310 -0.320 -0.000659 -0.0127   
 5 0 0 0.9 0.00682 -0.324 0.000102 -0.00388  
 6 0 0.1 0.1 0.00808 -0.295 0.00388 -0.0347   
 7 0 0.1 0.3 0.00338 -0.307 -0.00429 -0.0343   
 8 0 0.1 0.5 0.0104 -0.309 -0.00410 -0.0256   
 9 0 0.1 0.7 -0.00419 -0.329 0.00393 -0.00902  
10 0 0.1 0.9 0.00516 -0.321 -0.000591 -0.00484  
# ℹ 70 more rows  
# ℹ 47 more variables: bias\_rel.M <dbl>, bias\_rel.MAE <dbl>, bias\_tauT <dbl>,  
# bias\_tauE <dbl>, MSE\_varT.M <dbl>, MSE\_varT.MAE <dbl>, MSE\_varE.M <dbl>,  
# MSE\_varE.MAE <dbl>, MSE\_rel.M <dbl>, MSE\_rel.MAE <dbl>, MSE\_tauT <dbl>,  
# MSE\_tauE <dbl>, RMSE\_varT.M <dbl>, RMSE\_varT.MAE <dbl>, RMSE\_varE.M <dbl>,  
# RMSE\_varE.MAE <dbl>, RMSE\_rel.M <dbl>, RMSE\_rel.MAE <dbl>, RMSE\_tauT <dbl>,  
# RMSE\_tauE <dbl>, CVT\_empT <dbl>, CVE\_empT <dbl>, rel\_empT <dbl>, …

vis.df\_summarised$T\_ul - vis.df\_summarised$T\_ll

97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5%   
0.9980080 0.9481348 0.9496716 0.9469212 0.9712860 0.9618731 0.9528763 0.9408710   
 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5%   
0.9307054 0.9715433 0.8815830 0.9289250 0.9236190 0.9636163 0.9582408 0.8610613   
 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5%   
0.8787312 0.9381103 0.9585010 0.9998592 0.9869175 0.9833532 1.0263499 0.7860384   
 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5%   
0.7152636 0.9431323 0.9779654 1.1092559 0.7585112 0.7422973 0.9010322 0.9817085   
 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5%   
1.0573597 0.7943605 0.7775251 0.8949700 0.9555611 1.0536290 0.8316046 0.7892287   
 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5%   
0.9889657 1.1077520 0.7923271 0.7063435 0.7363034 0.9838759 0.9989508 0.7463248   
 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5%   
0.7129990 0.6933129 0.9283158 1.0862069 0.7878726 0.7488063 0.7027118 0.8529257   
 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5%   
0.8535446 0.7353154 0.7555984 0.7592874 1.0226183 0.8562861 0.6868790 0.7970994   
 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5%   
0.8772766 0.9955845 0.8128083 0.6934829 0.7643218 0.8783175 0.9743909 0.8176535   
 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5% 97.5%   
0.7294808 0.7782159 0.9569009 0.9011559 0.6824830 0.7433174 0.7051011 0.8995885

Simulating data under seemingly ideal circumstances – normal distribution of random true and error score variances, tau-equivalence of scores , …? – has shown that an analysis of score variance components is not without its flaws.

Concerning the heterogeneity in true score variance, the simulation demonstrates that degrees of uncertainty remain, surrounding the estimates of heterogeneity. In total, the size of confidence intervals varied between 0.682 and 1.109. Across conditions, RMSE varied between 0.182 and 0.604. As demonstrated in figure X, these seem positively related to the degree of score reliability. However, even at the highest level of reliability, uncertainty in estimated true score variance heterogeneity remained between 0.693 and 1. In terms of RMSE, uncertainty in estimated true score variance hetergeneity remained between 0.186 and 0.604 at score reliability of .9. Uncertainty is also related to the degree of true score variance heterogeneity, at higher degrees of heterogeneity the uncertainty seems to fall. For example, at score reliability of .7, average confidence interval size falls from 0.95 at CV\_T of 0, to 0.761 at CV\_T of .3. Similarly, at score reliability of .7, mean RMSE falls from 0.585 at CV\_T of 0, to 0.216 at CV\_T of .3.

More importantly, heterogeneity in true score variance estimates is highly overestimated at low degrees of reliability and low degrees of simulated heterogeneity. Average bias across all conditions was 0.188. In figure X, we can see that this overestimation falls with increasing heterogeneity in true score variance, as well as larger degree of score reliability. For example, at a reliability of .7, we see that average bias in heterogeneity in true score variance drops from 0.519 at CV\_T = 0, to -0.082 at CV\_T = .3. If anything, we seem to observe some degree of underestimation of heterogeneity in true score variance, if reliability is larger than .5 and heterogeneity is large. On the other hand, the simulated heterogeneity in error score variances seems to have no obvious effect on the estimates of true score variance heterogeneity.

Compared to the estimation of heterogeneity in true score variance, estimation of error score variance heterogeneity is more precise and accurate. The size of confidence intervals varied more strongly between 0.028 and 0.727, see figure X. Similarly, across all conditions, RMSE varied between 0.008 and 0.194. Overall, it appears estimation was more precise at all levels of score reliability and degree of heterogeneity. Similarly to true score variance, for error score variance heterogeneity, we observed that its uncertainty in estimation fell an increase in score reliability. For example, average confidence interval size drops from 0.451 at score reliability of .1, to 0.05 at score reliability of .9. Similary, mean RMSE drops from 0.135 to 0.015, as score reliability rises from .1 to .9. At the same time, uncertainty in the esimtation of error score variance heterogeneity rises, as heterogeneity grows, across all levels of score reliability. For example, at score reliability of .7, average confidence interval size rises from 0.089 at CV\_E of 0, to 0.242 at CV\_E of .3. Similarly, at score reliability of .7, mean RMSE rises from 0.046 at CV\_E of 0, to 0.065 at CV\_E of .3. Concerning the relationship between score variance heterogeneity in uncertainty in its estimation, it appears that relationship is inverse for error score variance, compared to true score variance. While estimation of true score variance heterogeneity gets more precise as heterogeneity rises, estimation of error score variance gets more imprecise.

While we observed some overestimation in true score variance heterogeneity, we seem to find less so for error score variance heterogeneity. On average, across all conditions, bias was 0.188. When score reliability and error score variance heterogeneity is low, from figure X it become clear that we still find some overestimation. For example, at a reliability of .7, we see that average bias in heterogeneity in error score variance drops from 0.035 at CV\_E = 0, to -0.018 at CV\_T = .3. While over- and underestimation of error score variance heterogeneity follows the same pattern as that of true score variance heterogeneity, from a comparison of figures X and X it becomes clear that this pattern is less strongly pronounced for heterogeneity in error score variance.

The simulation has also shown some implications concerning the Estimation of heterogeneity in score reliability estimates Uncertainty varies with size of score reliability and score variance component heterogeneity, in total between 0.006 and 0.06 in terms of confidence interval width. First of all, as score reliability grows, uncertainty in estimating its heterogeneity falls. For example, average confidence interval width in score reliability heterogeneity is 0.055 at targeted score reliability of .1 and 0.013 at targeted score reliability of .9. The relationship between variance component heterogeneity and uncertainty in estimation of score reliability heterogeneity is less obvious. Concerning error score variance heterogeneity, at CV\_E of 0, average confidence interval width is 0.034 and at CV\_E of .3 it is 0.042. Concerning true score variance heterogeneity, at CV\_T of 0, average confidence interval width is 0.034 and at CV\_E of .3 it is 0.044. It would appear that uncertainty in estimation of score reliability heterogeneity rises with larger heterogeneity in both variance components, but this increase is less strongly pronounced than the increase due to degree of targeted score reliability heterogeneity.

In cases of substantial heterogeneity in true and/or error score variance, a non-linear relationship between estimated heterogeneity in score reliability and degree of score reliability can be observed. As the heterogeneity in either (or both) of the score variance components grows, we find that heterogeneity in score reliability estimates grows as targeted score reliability approaches .5, where it plateaus. As targeted score reliability approaches .1 or .9, we observed a decrease in heterogeneity in score reliability. For example, at CV\_T of 0, CV\_E of .3 we find average score reliability heterogeneity to be 0.04 at targeted score reliability of .1, 0.084 at targeted score reliability of .5 and 0.026 at targeted score reliability of .9. Similarly, at CV\_T of .3 and CV\_E of 0, we find the average heterogeneity in score reliability to be 0.032 at targeted score reliability of .1, 0.067 at targeted score reliability of .5 and 0.026 at targeted score reliability of .9. It appears that there is an interaction between heterogeneity in the score variance components and the score reliability, when estimating heterogeneity in score reliability.

While it is not clear what degree of heterogeneity in score reliability is to be expected at specific degrees of heterogeneity in score variance components, we know that we should not find any heterogeneity if none of the two score variance components vary, no matter the degree of score reliability. However, if there is no heterogeneity in either of the score variance components, we instead find score reliability heterogeneity of 0.025 at targeted score reliability of .1, 0.013 at targeted score reliability of .5 and 0.003 at targeted score reliability of .9. Similarly to the score variance components, we end up overestimating the heterogeneity in score reliability if there is little or no heterogeneity in the components present. Additionally, we do so more strongly if targeted score reliability is low.