Reliability Estimation

Tate Paxton

2025-10-3

Set Working Directory, change to the folder where you have stored the three relevant excel documents.

```
# Set working directory
setwd("~/Documents/PhD Files/Externships/CC Assess")

# Optional: print to confirm
getwd()
```

```
## [1] "/Users/tatepaxton/Documents/PhD Files/Externships/CC Assess"
```

```
# Note: library designation chunk and data import chunk suppressed in knitted file
```

Reliability Estimates

The primary analysis estimated interrater reliability following, where possible, the methods outlined by the test publishers. The WISC-V Technical and Interpretive Manual specifies the use of intraclass correlations (ICC), while the GORT-5 reports product-moment correlations between two raters and multiple administrations. The Woodcock Johnson-IV Tests of Achievement authors provide no mention or estimate of interrater reliability. We adopted ICCs as our primary index because they are the most widely recommended for continuous or ordinal scoring of standardized test items and can be applied consistently across measures.

We sought additional information as needed within their cited sources (McGraw & Wong, 1996; Shrout & Fleiss, 1979) and more recent papers (Koo and Li, 2016). We used a "Case 3" design for total score reliability [ICC(3k)] and item-level reliability [ICC(3,1)], as "each target is rated by each of the same k judges, who are the only judges of interest" (Shrout & Fleiss, 1979, p. 421). This employs a two-way, random effects ANOVA to estimate intraclass correlations.

We also used Krippendorff's alpha to confirm item-level ICCs (Hayes and Krippendorff, 2007; Marzi et al., 2024) because it is robust to non-interval scaling and rater imbalance. These values were found to be nearly identical to ICC in each case, though we report both to maintain thoroughness.

WISC V Reliability Estimates

Total Score ICC across two administrations with 11-13 raters

```
# Read in the data
wisc_data <- read_excel(
  "/Users/tatepaxton/Documents/PhD Files/Externships/CC Assess/IRR All.xlsx",
  sheet = "WISC Raw"
)

# Convert to matrix
wisc_data <- as.matrix(wisc_data)

# Ensure it's numeric
wisc_data <- apply(wisc_data, c(1, 2), as.numeric)

# Transpose: now rows = subjects (subtests), columns = raters
wisc_data_t <- t(wisc_data)

# Run ICC for Similarities (only Sim2_17 and Sim7_25)
sim_icc <- psych::ICC(wisc_data_t[c("Sim2_17", "Sim7_25"), ])

# Run ICC for Vocabulary (Voc2_17 and Voc7_25)
voc_icc <- psych::ICC(wisc_data_t[c("Voc2_17", "Voc7_25"), ])

# Print results
sim_icc
```

```
## Call: psych::ICC(x = wisc_data_t[c("Sim2_17", "Sim7_25"), ])
##
## Intraclass correlation coefficients
##                          type  ICC  F df1 df2       p lower bound upper bound
## Single_raters_absolute   ICC1 0.77 45   1  24 6.4e-07        0.34           1
## Single_random_raters     ICC2 0.77 80   1  12 1.2e-06        0.35           1
## Single_fixed_raters      ICC3 0.86 80   1  12 1.2e-06        0.46           1
## Average_raters_absolute ICC1k 0.98 45   1  24 6.4e-07        0.87           1
## Average_random_raters   ICC2k 0.98 80   1  12 1.2e-06        0.88           1
## Average_fixed_raters    ICC3k 0.99 80   1  12 1.2e-06        0.92           1
##
##  Number of subjects = 2     Number of Judges =  13
## See the help file for a discussion of the other 4 McGraw and Wong estimates,
```

```
voc_icc
```

```
## Call: psych::ICC(x = wisc_data_t[c("Voc2_17", "Voc7_25"), ])
##
## Intraclass correlation coefficients
##                          type  ICC   F df1 df2       p lower bound upper bound
## Single_raters_absolute   ICC1 0.96 341   1  24 1.1e-15        0.82           1
## Single_random_raters     ICC2 0.96 373   1  12 2.1e-10        0.82           1
## Single_fixed_raters      ICC3 0.97 373   1  12 2.1e-10        0.81           1
## Average_raters_absolute ICC1k 1.00 341   1  24 1.1e-15        0.98           1
## Average_random_raters   ICC2k 1.00 373   1  12 2.1e-10        0.98           1
```

```
## Average_fixed_raters    ICC3k 1.00 373   1  12 2.1e-10        0.98           1
##
##  Number of subjects = 2     Number of Judges =  13
## See the help file for a discussion of the other 4 McGraw and Wong estimates,
```

Item level ICC

```
## --- WISC item-level ICC(3,1) ---
# Two-way mixed, absolute agreement, SINGLE rater (ICC(3,1))

icc_sim_217 <- irr::icc(sim2.17_t, model = "twoway", type = "agreement", unit = "single")  # ICC(3,1)
icc_sim_725 <- irr::icc(sim7.25_t, model = "twoway", type = "agreement", unit = "single")  # ICC(3,1)
icc_voc_217 <- irr::icc(voc2.17_t, model = "twoway", type = "agreement", unit = "single")  # ICC(3,1)
icc_voc_725 <- irr::icc(voc7.25_t, model = "twoway", type = "agreement", unit = "single")  # ICC(3,1)

cat(
  "\nItem-level ICC(3,1):",
  "\n  Similarities (2/17):", round(icc_sim_217$value, 2),
  "\n  Similarities (7/25):", round(icc_sim_725$value, 2),
  "\n  Vocabulary   (2/17):", round(icc_voc_217$value, 2),
  "\n  Vocabulary   (7/25):", round(icc_voc_725$value, 2), "\n"
)
```

```
##
## Item-level ICC(3,1):
##   Similarities (2/17): 0.86
##   Similarities (7/25): 0.82
##   Vocabulary   (2/17): 0.97
##   Vocabulary   (7/25): 0.9
```

Item-level Krippendorff's alpha

```
# WISC item-level Krippendorff's alpha

# Krippendorff's alpha (interval) with correct orientation: raters x items ---
ka_sim_217 <- irr::kripp.alpha(t(sim2.17_t), method = "interval")
ka_sim_725 <- irr::kripp.alpha(t(sim7.25_t), method = "interval")
ka_voc_217 <- irr::kripp.alpha(t(voc2.17_t), method = "interval")
ka_voc_725 <- irr::kripp.alpha(t(voc7.25_t), method = "interval")

cat(
  "\nKrippendorff's alpha (interval):",
  "\n  Similarities (2/17):", round(ka_sim_217$value, 2),
  "\n  Similarities (7/25):", round(ka_sim_725$value, 2),
  "\n  Vocabulary   (2/17):", round(ka_voc_217$value, 2),
  "\n  Vocabulary   (7/25):", round(ka_voc_725$value, 2), "\n"
)
```

```
##
## Krippendorff's alpha (interval):
##   Similarities (2/17): 0.85
##   Similarities (7/25): 0.81
##   Vocabulary   (2/17): 0.97
##   Vocabulary   (7/25): 0.89
```

WJ-IV Reliability Estimates

The WJ-IV does not have multiple adiminstrations rated. Nor were we able to convert the raw total scores into standard scores to estimate the total score reliability on a shared scale. Therefore, we will focus on item-level reliability using ICC and K's alpha.

ICC

```
## --- WJ-IV item-level ICC(3,1) ---
# Two-way mixed, absolute agreement, SINGLE rater (ICC(3,1))
# Inputs are items x raters matrices you already made: wjpc_t, wjwa_t, wjor_t

icc_wj_pc <- irr::icc(wjpc_t, model = "twoway", type = "agreement", unit = "single")  # Passage Compreh
icc_wj_wa <- irr::icc(wjwa_t, model = "twoway", type = "agreement", unit = "single")  # Word Attack
icc_wj_or <- irr::icc(wjor_t, model = "twoway", type = "agreement", unit = "single")  # Oral Reading

cat(
  "\nWJ-IV Item-level ICC(3,1):",
  "\n  Passage Comprehension:", round(icc_wj_pc$value, 2),
  "\n  Word Attack:          ", round(icc_wj_wa$value, 2),
  "\n  Oral Reading:         ", round(icc_wj_or$value, 2), "\n"
)
```

```
##
## WJ-IV Item-level ICC(3,1):
##   Passage Comprehension: 1
##   Word Attack:           0.99
##   Oral Reading:          0.84
```

K's Alpha

```
# WJ-IV item-level Krippendorff's alpha (interval)

ka_wj_pc <- irr::kripp.alpha(t(wjpc_t), method = "interval")
ka_wj_wa <- irr::kripp.alpha(t(wjwa_t), method = "interval")
ka_wj_or <- irr::kripp.alpha(t(wjor_t), method = "interval")

cat(
  "\nWJ-IV Krippendorff's alpha (interval):",
  "\n  Passage Comprehension:", round(ka_wj_pc$value, 2),
  "\n  Word Attack:          ", round(ka_wj_wa$value, 2),
  "\n  Oral Reading:         ", round(ka_wj_or$value, 2), "\n"
)
```

```
##
## WJ-IV Krippendorff's alpha (interval):
##   Passage Comprehension: 1
##   Word Attack:           0.99
##   Oral Reading:          0.83
```

GORT-5 Reliability Estimates

For the total score reliability, we used the scaled scores across the four domains, which gave a shared metric on which we can fairly estimate reliability across different domains. Of note, this means that the total score reliability is a general reliability across all four domains. We calculated item-level reliability for each domain using the raw scores of each story. We did this for both ICC(3,1) and K's alpha estimates.

ICC Total Score reliability

```
## --- GORT total-score ICC(3,k) using scaled scores (GORT SS tab) ---
# Rows = subjects, cols = raters
gort_icc_tbl <- psych::ICC(gortss_t)

gort_icc_tbl
```

```
## Call: psych::ICC(x = gortss_t)
##
## Intraclass correlation coefficients
##                             type  ICC  F df1 df2       p lower bound upper bound
## Single_raters_absolute     ICC1 0.87 35   3  16 2.7e-07        0.61        0.99
## Single_random_raters       ICC2 0.87 62   3  12 1.4e-07        0.58        0.99
## Single_fixed_raters        ICC3 0.92 62   3  12 1.4e-07        0.72        0.99
## Average_raters_absolute ICC1k 0.97 35   3  16 2.7e-07        0.88        1.00
## Average_random_raters   ICC2k 0.97 62   3  12 1.4e-07        0.87        1.00
## Average_fixed_raters    ICC3k 0.98 62   3  12 1.4e-07        0.93        1.00
##
##  Number of subjects = 4     Number of Judges =  5
## See the help file for a discussion of the other 4 McGraw and Wong estimates,
```

Item-level

```
## --- GORT item-level ICC(3,1) by domain ---
icc_gort_rate <- irr::icc(gort_rate_t, model = "twoway", type = "agreement", unit = "single")
icc_gort_acc  <- irr::icc(gort_acc_t,  model = "twoway", type = "agreement", unit = "single")
icc_gort_flu  <- irr::icc(gort_flu_t,  model = "twoway", type = "agreement", unit = "single")
icc_gort_comp <- irr::icc(gort_comp_t, model = "twoway", type = "agreement", unit = "single")

cat(
  "\nGORT Item-level ICC(3,1):",
  "\n  Rate:         ", round(icc_gort_rate$value, 2),
  "\n  Accuracy:     ", round(icc_gort_acc$value, 2),
  "\n  Fluency:      ", round(icc_gort_flu$value, 2),
  "\n  Comprehension:", round(icc_gort_comp$value, 2), "\n"
)
```

```
##
## GORT Item-level ICC(3,1):
##   Rate:         0.99
##   Accuracy:     0.8
##   Fluency:      0.94
##   Comprehension: 0.8
```

K's alpha

```r
## --- GORT item-level Krippendorff's alpha (interval) by domain ---
ka_gort_rate <- irr::kripp.alpha(t(gort_rate_t), method = "interval")
ka_gort_acc  <- irr::kripp.alpha(t(gort_acc_t),  method = "interval")
ka_gort_flu  <- irr::kripp.alpha(t(gort_flu_t),  method = "interval")
ka_gort_comp <- irr::kripp.alpha(t(gort_comp_t), method = "interval")

cat(
  "\nGORT Krippendorff's alpha (interval):",
  "\n  Rate:          ", round(ka_gort_rate$value, 2),
  "\n  Accuracy:      ", round(ka_gort_acc$value, 2),
  "\n  Fluency:       ", round(ka_gort_flu$value, 2),
  "\n  Comprehension:", round(ka_gort_comp$value, 2), "\n"
)
```

```
##
## GORT Krippendorff's alpha (interval):
##   Rate:          0.99
##   Accuracy:      0.78
##   Fluency:       0.93
##   Comprehension: 0.78
```