S-061 Assignment 1

Thomas Kelley-Kemple, Joanna Moody, Vinh Nguyen

September 18, 2017

Part 1: Classical Test Theory & Validation

1) On the basis of this correlation, the researcher states that the reliability of the ratings is 0.7. What *score* is she assuming is relevant, that has a reliability of 0.7? Is it the score from e-Rater A? The score from e-Rater B? Some combination of both scores? A score from any single e-Rater? Or other?

2) If the researcher considers 0.7 to be the reliability, what is the replication that she assumes is relevant? What is random, and what is fixed?

3) Apply Spearman-Brown (actually perform a calculation) to estimate the reliability of the average these two e-Rater scores.

4) The company asks you what a valid use of the third score would be. Remember this score is available for only 10% of examinees. Should the company use this score alone? Should it use the unweighted average of the three scores (the two e-rater scores and the human score)? Should it ignore the score and use the average of the two e-Rater scores? Answer the following questions:

a) If you were an examinee with a high (well above average) true score, would you rather have the human score, the average of the two e-rater scores, or the average of all three scores?

b) If you were an examinee with a low (well below average) true score, would you rather have the human score, the average of the two e-rater scores, or the average of all three scores?

c) Weighing all considerations for the intended use of these scores for college admissions, what would be your recommendation to the company for how they should use this third score?

Part 2: Classical Test Theory and Exploratory Analysis

```
# Read in data
data_raw <- read_dta("./Assignment1.dta")
```

5) Using Stata, calculate coefficient alpha for the first occasion and the second occasion separately. In a sentence or two, interpret coefficient alpha for the first occasion (see also Question 16).

```
# create variable lists
o1 <- paste0("x_o1_i", 1:12)
o2 <- paste0("x_o2_i", 1:12)
# subset data
data_o1 <- subset(data_raw, select = o1)
data_o2 <- subset(data_raw, select = o2)
# create alpha output
alpha1 <- alpha(data_o1, keys = NULL, title = NULL, cumulative = FALSE, max = 10,
    na.rm = TRUE, check.keys = TRUE, n.iter = 1, delete = TRUE)
alpha2 <- alpha(data_o2, keys = NULL, title = NULL, cumulative = FALSE, max = 10,
    na.rm = TRUE, check.keys = TRUE, n.iter = 1, delete = TRUE)
# get just the alpha numbers
alpha_time1 <- alpha1$total$std.alpha
alpha_time2 <- alpha2$total$std.alpha
# output alpha data
alpha1
```

```
##
## Reliability analysis
## Call: alpha(x = data_o1, keys = NULL, cumulative = FALSE, title = NULL,
##     max = 10, na.rm = TRUE, check.keys = TRUE, n.iter = 1, delete = TRUE)
##
##   raw_alpha std.alpha G6(smc) average_r S/N   ase mean   sd
##       0.88      0.87    0.94      0.36 6.9 0.033  2.6 0.91
##
##  lower alpha upper     95% confidence boundaries
## 0.82 0.88 0.95
##
##  Reliability if an item is dropped:
##          raw_alpha std.alpha G6(smc) average_r S/N alpha se
## x_o1_i1       0.86      0.85    0.92      0.34 5.6    0.040
## x_o1_i2       0.87      0.86    0.92      0.35 6.0    0.037
## x_o1_i3       0.88      0.87    0.93      0.38 6.8    0.034
## x_o1_i4       0.86      0.85    0.92      0.34 5.7    0.040
## x_o1_i5       0.86      0.86    0.93      0.35 5.9    0.038
## x_o1_i6       0.87      0.86    0.93      0.36 6.1    0.037
## x_o1_i7       0.87      0.86    0.93      0.37 6.4    0.035
## x_o1_i8       0.87      0.87    0.93      0.37 6.5    0.035
## x_o1_i9       0.88      0.87    0.93      0.38 6.7    0.033
## x_o1_i10      0.86      0.85    0.92      0.35 5.9    0.039
## x_o1_i11      0.89      0.89    0.94      0.42 7.9    0.033
## x_o1_i12      0.88      0.87    0.92      0.37 6.6    0.035
##
```

```
##  Item statistics
##            n raw.r std.r r.cor r.drop mean   sd
## x_o1_i1  25  0.84  0.83  0.84   0.78  3.5 1.58
## x_o1_i2  25  0.72  0.73  0.72   0.65  2.8 1.45
## x_o1_i3  25  0.52  0.52  0.50   0.43  2.1 1.29
## x_o1_i4  25  0.82  0.82  0.81   0.76  2.9 1.55
## x_o1_i5  25  0.76  0.75  0.73   0.70  3.0 1.35
## x_o1_i6  25  0.72  0.69  0.68   0.64  2.6 1.50
## x_o1_i7  25  0.61  0.63  0.60   0.54  1.6 1.08
## x_o1_i8  25  0.59  0.61  0.58   0.52  2.1 1.15
## x_o1_i9  25  0.53  0.54  0.52   0.43  2.1 1.41
## x_o1_i10 25  0.78  0.76  0.76   0.71  3.7 1.57
## x_o1_i11 25  0.23  0.28  0.22   0.16  1.6 0.82
## x_o1_i12 25  0.62  0.58  0.57   0.52  2.8 1.64
##
## Non missing response frequency for each item
##             1    2    3    4    5 miss
## x_o1_i1  0.20 0.08 0.12 0.20 0.40    0
## x_o1_i2  0.24 0.28 0.12 0.20 0.16    0
## x_o1_i3  0.52 0.08 0.24 0.12 0.04    0
## x_o1_i4  0.24 0.24 0.12 0.16 0.24    0
## x_o1_i5  0.12 0.32 0.20 0.16 0.20    0
## x_o1_i6  0.32 0.20 0.16 0.16 0.16    0
## x_o1_i7  0.68 0.16 0.08 0.04 0.04    0
## x_o1_i8  0.40 0.28 0.20 0.08 0.04    0
## x_o1_i9  0.48 0.28 0.04 0.08 0.12    0
## x_o1_i10 0.16 0.12 0.08 0.16 0.48    0
## x_o1_i11 0.60 0.28 0.08 0.04 0.00    0
## x_o1_i12 0.40 0.04 0.16 0.20 0.20    0
```

```
alpha2
```

```
##
## Reliability analysis
## Call: alpha(x = data_o2, keys = NULL, cumulative = FALSE, title = NULL,
##     max = 10, na.rm = TRUE, check.keys = TRUE, n.iter = 1, delete = TRUE)
##
##   raw_alpha std.alpha G6(smc) average_r S/N   ase mean   sd
##       0.85      0.85    0.92      0.33 5.8 0.042  2.8 0.85
##
##  lower alpha upper     95% confidence boundaries
## 0.77 0.85 0.94
##
##  Reliability if an item is dropped:
##          raw_alpha std.alpha G6(smc) average_r S/N alpha se
## x_o2_i1       0.84      0.84    0.89      0.32 5.2    0.046
## x_o2_i2       0.84      0.84    0.91      0.32 5.2    0.047
## x_o2_i3       0.83      0.83    0.90      0.31 4.8    0.050
## x_o2_i4       0.85      0.85    0.91      0.35 5.8    0.042
## x_o2_i5       0.83      0.83    0.90      0.31 5.0    0.048
## x_o2_i6       0.83      0.83    0.90      0.30 4.8    0.049
## x_o2_i7       0.84      0.84    0.90      0.32 5.2    0.046
## x_o2_i8       0.85      0.85    0.91      0.34 5.7    0.044
## x_o2_i9       0.83      0.83    0.91      0.31 4.9    0.050
## x_o2_i10      0.85      0.85    0.91      0.34 5.8    0.042
```

```
## x_o2_i11      0.86      0.86    0.91      0.36 6.3    0.040
## x_o2_i12      0.85      0.85    0.91      0.35 5.8    0.042
##
##  Item statistics
##           n raw.r std.r r.cor r.drop mean  sd
## x_o2_i1  25  0.65  0.66  0.65   0.56  3.6 1.4
## x_o2_i2  25  0.69  0.68  0.66   0.60  3.0 1.6
## x_o2_i3  25  0.79  0.78  0.78   0.72  2.5 1.5
## x_o2_i4  25  0.47  0.48  0.44   0.36  2.4 1.3
## x_o2_i5  25  0.75  0.74  0.73   0.68  3.1 1.3
## x_o2_i6  25  0.79  0.79  0.79   0.74  3.6 1.3
## x_o2_i7  25  0.66  0.68  0.66   0.59  2.1 1.2
## x_o2_i8  25  0.53  0.53  0.48   0.43  2.5 1.3
## x_o2_i9  25  0.78  0.77  0.75   0.70  2.9 1.5
## x_o2_i10 25  0.51  0.50  0.45   0.39  2.8 1.5
## x_o2_i11 25  0.34  0.35  0.30   0.22  2.1 1.3
## x_o2_i12 25  0.47  0.47  0.43   0.36  2.5 1.3
##
## Non missing response frequency for each item
##             1    2    3    4    5 miss
## x_o2_i1  0.16 0.04 0.16 0.28 0.36    0
## x_o2_i2  0.24 0.24 0.08 0.16 0.28    0
## x_o2_i3  0.40 0.12 0.12 0.28 0.08    0
## x_o2_i4  0.28 0.36 0.16 0.12 0.08    0
## x_o2_i5  0.12 0.20 0.36 0.12 0.20    0
## x_o2_i6  0.08 0.16 0.12 0.40 0.24    0
## x_o2_i7  0.44 0.20 0.16 0.20 0.00    0
## x_o2_i8  0.28 0.24 0.28 0.12 0.08    0
## x_o2_i9  0.24 0.20 0.20 0.12 0.24    0
## x_o2_i10 0.24 0.20 0.24 0.12 0.20    0
## x_o2_i11 0.48 0.16 0.20 0.08 0.08    0
## x_o2_i12 0.28 0.28 0.20 0.16 0.08    0
```

```r
pander(alpha2)
```

```
## Warning in pander.default(alpha2): No pander.method for "psych", reverting
## to default.No pander.method for "alpha", reverting to default.
```

- **total**:

raw_alpha	std.alpha	G6(smc)	average_r	S/N	ase	mean	sd
0.8547	0.8539	0.9187	0.3276	5.845	0.04195	2.76	0.8516

- **alpha.drop**:

	raw_alpha	std.alpha	G6(smc)	average_r	S/N	alpha se
x_o2_i1	0.8413	0.8396	0.893	0.3224	5.234	0.04566
x_o2_i2	0.8386	0.8379	0.9067	0.3197	5.168	0.04698
x_o2_i3	0.8286	0.8289	0.899	0.3057	4.844	0.05015
x_o2_i4	0.8538	0.8533	0.9057	0.346	5.819	0.04223
x_o2_i5	0.8334	0.8325	0.8961	0.3113	4.972	0.04838
x_o2_i6	0.8298	0.8276	0.895	0.3039	4.802	0.04938
x_o2_i7	0.8401	0.8379	0.8998	0.3196	5.168	0.04629
x_o2_i8	0.8498	0.8499	0.9111	0.3398	5.662	0.04361

	raw__alpha	std.alpha	G6(smc)	average__r	S/N	alpha se
x__o2__i9	0.8298	0.8298	0.905	0.3072	4.877	0.04977
x__o2__i10	0.8533	0.852	0.9129	0.3435	5.755	0.04244
x__o2__i11	0.8637	0.8632	0.9098	0.3644	6.308	0.03957
x__o2__i12	0.8541	0.854	0.9086	0.3472	5.85	0.04241

- **item.stats**:

	n	raw.r	std.r	r.cor	r.drop	mean	sd
x__o2__i1	25	0.6515	0.6574	0.6468	0.5583	3.64	1.44
x__o2__i2	25	0.6925	0.6778	0.6552	0.5958	3	1.607
x__o2__i3	25	0.7888	0.7807	0.7769	0.7236	2.52	1.475
x__o2__i4	25	0.4678	0.4831	0.4405	0.3637	2.36	1.254
x__o2__i5	25	0.746	0.7397	0.7333	0.6814	3.08	1.288
x__o2__i6	25	0.7932	0.7945	0.7939	0.7399	3.56	1.261
x__o2__i7	25	0.6639	0.6779	0.6623	0.59	2.12	1.201
x__o2__i8	25	0.5283	0.5287	0.4828	0.4303	2.48	1.262
x__o2__i9	25	0.7758	0.7702	0.7484	0.7046	2.92	1.525
x__o2__i10	25	0.5114	0.5014	0.4548	0.3939	2.84	1.463
x__o2__i11	25	0.3384	0.3464	0.2974	0.2158	2.12	1.333
x__o2__i12	25	0.4709	0.4741	0.4316	0.3635	2.48	1.295

- **response.freq**:

	1	2	3	4	5	miss
x__o2__i1	0.16	0.04	0.16	0.28	0.36	0
x__o2__i2	0.24	0.24	0.08	0.16	0.28	0
x__o2__i3	0.4	0.12	0.12	0.28	0.08	0
x__o2__i4	0.28	0.36	0.16	0.12	0.08	0
x__o2__i5	0.12	0.2	0.36	0.12	0.2	0
x__o2__i6	0.08	0.16	0.12	0.4	0.24	0
x__o2__i7	0.44	0.2	0.16	0.2	0	0
x__o2__i8	0.28	0.24	0.28	0.12	0.08	0
x__o2__i9	0.24	0.2	0.2	0.12	0.24	0
x__o2__i10	0.24	0.2	0.24	0.12	0.2	0
x__o2__i11	0.48	0.16	0.2	0.08	0.08	0
x__o2__i12	0.28	0.28	0.2	0.16	0.08	0

- **keys**: *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1*, *1* and *1*
- **scores**: *3.083*, *2.667*, *1.25*, *1.667*, *1.583*, *3.667*, *3.167*, *2.167*, *3.833*, *3.25*, *2.583*, *4.417*, *3*, *2.583*, *3.917*, *3.583*, *3.5*, *2.583*, *3.25*, *2.833*, *2.167*, *1.083*, *1.917*, *2.417* and *2.833*
- **nvar**: *12*
- **boot.ci**:
- **boot**:
- **Unidim**:
    - **Unidim**: *0.6722*

- **Fit**:
  - **Fit.off**: *0.9272*
- **call**: alpha(x = data_o2, keys = NULL, cumulative = FALSE, title = NULL, max = 10, na.rm = TRUE, check.keys = TRUE, n.iter = 1, delete = TRUE)
- **title**:

```
summary(alpha2)
```

```
##
## Reliability analysis
##  raw_alpha std.alpha G6(smc) average_r S/N   ase mean   sd
##       0.85      0.85    0.92      0.33 5.8 0.042  2.8 0.85
```

```
pander(summary(alpha2))
```

Reliability analysis
raw_alpha std.alpha G6(smc) average_r S/N ase mean sd 0.85 0.85 0.92 0.33 5.8 0.042 2.8 0.85

raw_alpha	std.alpha	G6(smc)	average_r	S/N	ase	mean	sd
0.8547	0.8539	0.9187	0.3276	5.845	0.04195	2.76	0.8516

For this sample $\alpha_1 = 0.873$ and $\alpha_2 = 0.854$

6) Using Stata, calculate the average score of participants from the first occasion, then calculate the average score of participants from the second occasion. Then, calculate the correlation between the two average scores using code like pwcorr avgscr1 avgscr2. Report this correlation and, in a sentence or two, provide an interpretation (see also Question 16).

```
avg_o1 <- rowMeans(data_o1)
avg_o2 <- rowMeans(data_o2)

cor(avg_o1, avg_o2)
```

```
## [1] 0.790936
```

7) Reload the data and reshape it for analysis in Stata. Although it is a pain, I am requiring you to use some of the code that we have presented in the past .do files to reshape the data from "double-wide" format. See, for example, the Class03.do and Class04.do files. As one way to check your work, submit a screenshot of the output from code like table person item occasion, contents(mean score) and/or simply table person item occasion .

```
data_raw$person <- factor(data_raw$person)
colnames(data_raw) <- c("person", paste0("1_", 1:12), paste0("2_", 1:12))

data_long <- melt(data_raw, id.vars = c("person"))
data_long <- separate(data = data_long, col = variable, into = c("occasion", "item"),
    sep = "_")
```

8) 8. Note the code available to you in the .do files, and include a) a discrete histogram of all 25x12x2 scores, b) a histogram of marginal person scores, c) a histogram of marginal item scores, and d) a histogram of marginal occasion scores. Use discrete histograms where you think they are appropriate, or substitute tables if histograms are not informative, for example, tabulate occasion, summarize(score) . Histograms of interactions are not necessary.

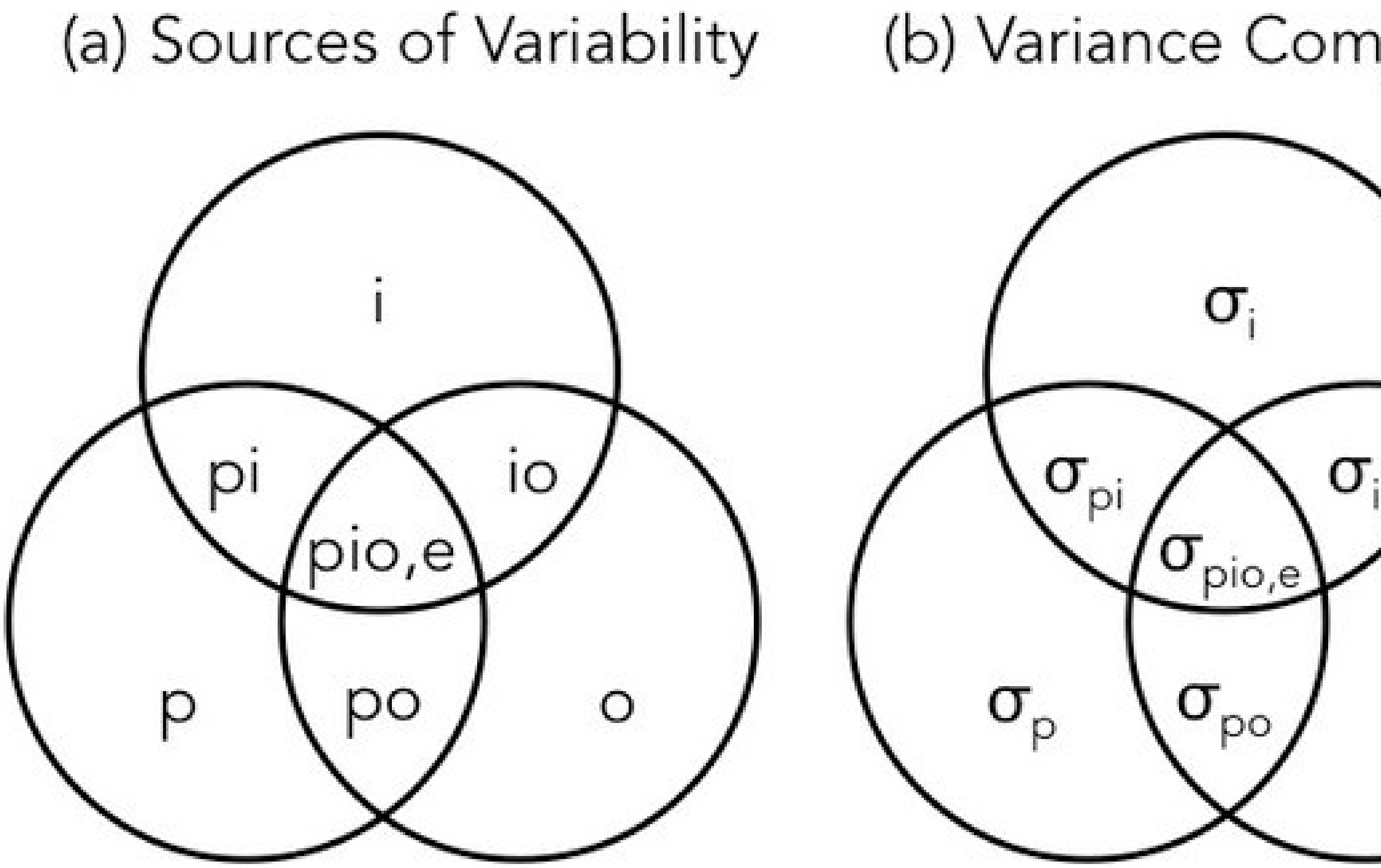
Part 3: The Generalizability Study

9) Write out the model implied by the data collection design under the tenets of Generalizability Theory. Draw the Venn diagram for this design.

The model implied here can be written as

$$
\begin{aligned}
X_{pi} &= \mu + \nu_p + \nu_i + \nu_o + \nu_{pi} + \nu_{po} + \nu_{oi} + \nu_{pio,e} \\
\nu_p &\sim N(0, \sigma_p^2) \\
\nu_i &\sim N(0, \sigma_i^2) \\
\nu_o &\sim N(0, \sigma_o^2) \\
\nu_{pi} &\sim N(0, \sigma_{pi}^2) \\
\nu_{po} &\sim N(0, \sigma_{po}^2) \\
\nu_{io} &\sim N(0, \sigma_{io}^2) \\
\nu_{pio,e} &\sim N(0, \sigma_{pio,e}^2)
\end{aligned}
$$

The venn diagram for the variances is seen below:



10) Estimate the variance components for this model using the mixed or xtmixed command. Feel free to go get coffee while this runs. Don't forget to create interactions using commands like egen pXi = group(person item). Include a table with four columns, the source of variance, the estimated variance components, their square roots, and their percentage of total score variance.

```
data_long$pxi <- as.factor(100 * as.numeric(data_long$person) + as.numeric(data_long$item))
data_long$pxo <- as.factor(100 * as.numeric(data_long$person) + as.numeric(data_long$occasion))
data_long$oxi <- as.factor(10 * as.numeric(data_long$occasion) + as.numeric(data_long$item))
```

```r
mixed <- lmer(value ~ 1 + (1 | person) + (1 | item) + (1 | occasion) + (1 | pxi) +
    (1 | pxo) + (1 | oxi), data = data_long)

summary(mixed)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: value ~ 1 + (1 | person) + (1 | item) + (1 | occasion) + (1 |
##     pxi) + (1 | pxo) + (1 | oxi)
##    Data: data_long
##
## REML criterion at convergence: 1900
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.42134 -0.58655 -0.02843  0.54729  2.64540
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  pxi      (Intercept) 0.48748  0.6982
##  pxo      (Intercept) 0.10228  0.3198
##  person   (Intercept) 0.57337  0.7572
##  oxi      (Intercept) 0.09958  0.3156
##  item     (Intercept) 0.26306  0.5129
##  occasion (Intercept) 0.01380  0.1175
##  Residual             0.74172  0.8612
## Number of obs: 600, groups:
## pxi, 300; pxo, 50; person, 25; oxi, 22; item, 12; occasion, 2
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.6562     0.2476   10.73
```

11) A novice psychometrician with no sense of the context observes from the percentages, "it looks like items are a much greater source of variance than occasions!" Explain the flaw in this reasoning.

12) Estimate the Mean Squares for this model using the anova command. You will first need to set the maximum matrix size to a large number, using code like set matsize 1000. Write out the equation for the estimated variance component, $\hat{\sigma}_p^2$, in terms of mean squares, MS, and confirm that this calculation corresponds to your results from mixed or xtmixed. Recall that $n_p = 25, n_i = 12$ and $n_o = 2$.

```r
anovlm <- lm(value ~ person + item + occasion + pxi + pxo + oxi, data = data_long)

anova(anovlm)
```

```
## Analysis of Variance Table
##
## Response: value
##           Df Sum Sq Mean Sq F value    Pr(>F)
## person    24 400.92 16.7049 22.5668 < 2.2e-16 ***
```

```
## item       11 160.34 14.5762 19.6912 < 2.2e-16 ***
## occasion    1   6.20  6.2017  8.3779  0.004115 **
## pxi       264 453.20  1.7167  2.3191 8.453e-12 ***
## pxo        24  47.26  1.9690  2.6600 7.521e-05 ***
## oxi        11  39.62  3.6017  4.8655 7.547e-07 ***
## Residuals 264 195.42  0.7402
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

13) Calculate and report the mean and standard deviation of marginal person scores, averaging over items and occasions. Following the code from class, you could obtain this using code like, summarize pmean if ptag . Explain why the term $\hat{\sigma_p}$, is less than the standard deviation of marginal person means.

14) Describe the o, po, and io variance components in words, and include whether they are good, bad, or neutral with respect to relative error in a $p \times i \times o$ design. There is no need to reference the actual values, here.

The o variance component describes variance across occations.

Part 4: The Decision Study

15) Write out the full equation for the relative error variance, σ^2_δ

16) Calculate the generalizability coefficient for relative error, $E\hat{\rho}^2$, when there are 12 items administered on one occasion. Explain the differences between this coefficient, the coefficients from Question 5, and the coefficient from Question 6. Explain the differences between the questions that these different coefficients answer.

17) Use the "pxixr D Study Template" to include a graph of a) the standard error of measurement and b) the generalizability coefficient for relative error. Relabel and rescale where appropriate.

18) If the scale is administered on 1 occasion, how many items are required to achieve a reliability of 0.75? You can use the template to answer this.

19) Compare the benefits of doubling the number of items from 6 to 12 versus doubling the number of occasions from 1 to 2. Compare the benefits of doubling the number of items from 12 to 24 versus doubling the number of occasions from 1 to 2. How could you use this information to address the question of whether items are a greater source of error than occasions?