# Problem Set 1

*Thomas Kelley-Kemple, Joanna Moody, Vinh Nguyen*

*September 18, 2017*

**Part 1**

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?**

The researcher is referring to the reliability of the score generating processes by the population of random e-Raters with different scoring algorithms for this specific essay prompt, from which e-Raters A and B came. We can think of this as the inter-rater reliability

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?**

The replication is the scoring process involving a random e-Rater (or, scoring algorithm). The e-Rater/scoring algorithm is random, but the students, form, and testing occasion are fixed.

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

```
#create function to calculate Spearman-Brown
SB_Rho<-function(k,rho){
  return(k*rho/(1+(k-1)*rho))
}

SB_Rho(2,.7)
```

```
## [1] 0.8235294
```

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**

```
#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).**

```r
#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
alpha2 <- alpha(data_o2, keys=NULL, title=NULL, cumulative=FALSE, max=10, na.rm = TRUE, check.keys=TRUE
#get just the alpha numbers
alpha_time1 <- alpha1$total$std.alpha
alpha_time2 <- alpha2$total$std.alpha
#output alpha data
alpha1[c(1,2)]
```

```
## $total
##  raw_alpha std.alpha   G6(smc) average_r      S/N        ase     mean
##  0.8804046 0.8732209 0.9383743 0.3646671 6.887736 0.03320417 2.556667
##         sd
##  0.9115133
##
## $alpha.drop
##          raw_alpha std.alpha   G6(smc) average_r      S/N   alpha se
## x_o1_i1  0.8569859 0.8489743 0.9194098 0.3382022 5.621391 0.04023086
## x_o1_i2  0.8664800 0.8571515 0.9171808 0.3529573 6.000424 0.03655268
## x_o1_i3  0.8790375 0.8718499 0.9256352 0.3821387 6.803350 0.03356221
## x_o1_i4  0.8587405 0.8503834 0.9219498 0.3406758 5.683749 0.03957368
## x_o1_i5  0.8637962 0.8554854 0.9307687 0.3498710 5.919718 0.03794051
## x_o1_i6  0.8668948 0.8601196 0.9257334 0.3585619 6.148965 0.03707616
## x_o1_i7  0.8734048 0.8641382 0.9303787 0.3663748 6.360421 0.03516011
## x_o1_i8  0.8743661 0.8658129 0.9323785 0.3697099 6.452281 0.03502604
## x_o1_i9  0.8797215 0.8705569 0.9294884 0.3794218 6.725405 0.03281351
## x_o1_i10 0.8619843 0.8544676 0.9216150 0.3480061 5.871324 0.03862585
## x_o1_i11 0.8878134 0.8870726 0.9403549 0.4166080 7.855246 0.03260048
## x_o1_i12 0.8756217 0.8676766 0.9222698 0.3734778 6.557240 0.03454801
```

```r
alpha2[c(1,2)]
```

```
## $total
##  raw_alpha std.alpha   G6(smc) average_r      S/N        ase mean       sd
##  0.8547132 0.8539146 0.9187441 0.3275545 5.845313 0.04194918 2.76 0.851646
##
## $alpha.drop
##          raw_alpha std.alpha   G6(smc) average_r      S/N   alpha se
## x_o2_i1  0.8413197 0.8395884 0.8929899 0.3224083 5.233965 0.04565599
## x_o2_i2  0.8385635 0.8378787 0.9066881 0.3196530 5.168219 0.04698382
## x_o2_i3  0.8286076 0.8288941 0.8989762 0.3057455 4.844333 0.05015026
## x_o2_i4  0.8538483 0.8533409 0.9057036 0.3459597 5.818536 0.04223314
## x_o2_i5  0.8334079 0.8325484 0.8961089 0.3112893 4.971873 0.04837585
## x_o2_i6  0.8298088 0.8276441 0.8950010 0.3038833 4.801947 0.04937522
## x_o2_i7  0.8400521 0.8378697 0.8997610 0.3196386 5.167878 0.04628575
## x_o2_i8  0.8497800 0.8498859 0.9110713 0.3397993 5.661601 0.04360574
## x_o2_i9  0.8297841 0.8298422 0.9050349 0.3071695 4.876898 0.04977468
## x_o2_i10 0.8533153 0.8519660 0.9128966 0.3434877 5.755207 0.04243731
```

```
## x_o2_i11 0.8637079 0.8631580 0.9097513 0.3644446 6.307696 0.03957437
## x_o2_i12 0.8540983 0.8540116 0.9085511 0.3471755 5.849858 0.04240916
```

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) **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.**

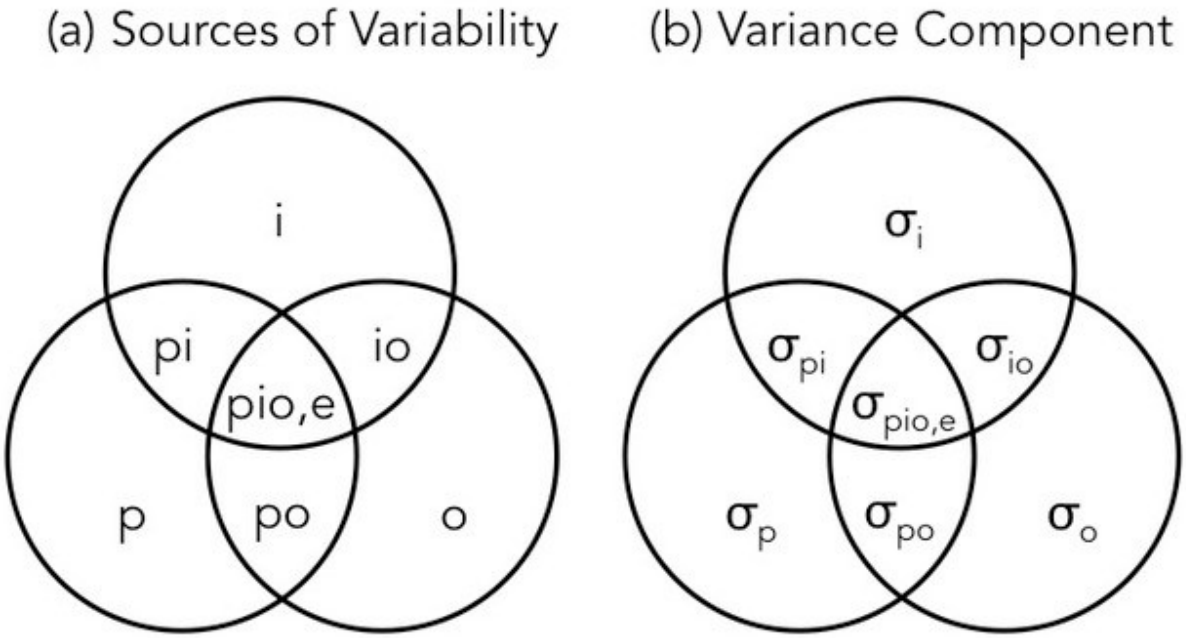
**Part 3**

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))

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.**

The novice is not taking into account the fact that there are many more items than occasions. Moreover, the cost of adding additional items may be quite low, especially as compared to adding another test.

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 $n_o = 2$.**

```
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
```

The equation for finding $\hat{sigma_p^2}$ is

$$\frac{MS_p - MS_{pi} - MS_{po} + MS_{pio,e}}{n_i n_o}$$

plugging in the Mean Squares and $n$s from our ANOVA, we find that the estimated variance component is

$$\hat{\sigma}_p^2 = \frac{16.7 - 1.72 - 1.97 + .74}{12 \cdot 2} = .5729$$

This is basically the same as the answer we recieved from the mixed model.

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.**

```
#get marginal person scores

marg_person <- ddply(data_long,"person",summarise,mean_p=mean(value))

#mean person score
round(mean(marg_person$mean_p),3)
```

```
## [1] 2.658
```

```
#variance of person scores
round(sd(marg_person$mean_p),3)
```

```
## [1] 0.834
```

The estimate for $\sigma_p$ is smaller than our calcluated standard deviation here (.757 vs .834) because the variance across people also includes the variance of people interacted with items, people interacted with occasions, and people interacted with both (plus random error). The whole point of G-theory is to separate these components out and so we would be remiss to assume that this standard deviation of person scores is all attributable to true differences in their scores.

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 occasions (constant across persons and items). This is ostensibly neutral as long as it affects all people equally. At the same time, we would prefer it to be smaller as it adds undesirable noise if we want to use the test score in an absolute setting. In contrast, the $po$ variance component is definitely bad, as it obfuscates both relative and absolute positions of persons across testing occasions. Finally, $oi$ variance is fairly benign. While it's not a good thing that items change across time, if it does not affect different people differently, then it will not alter peoples' relative scores.

**Part 4**

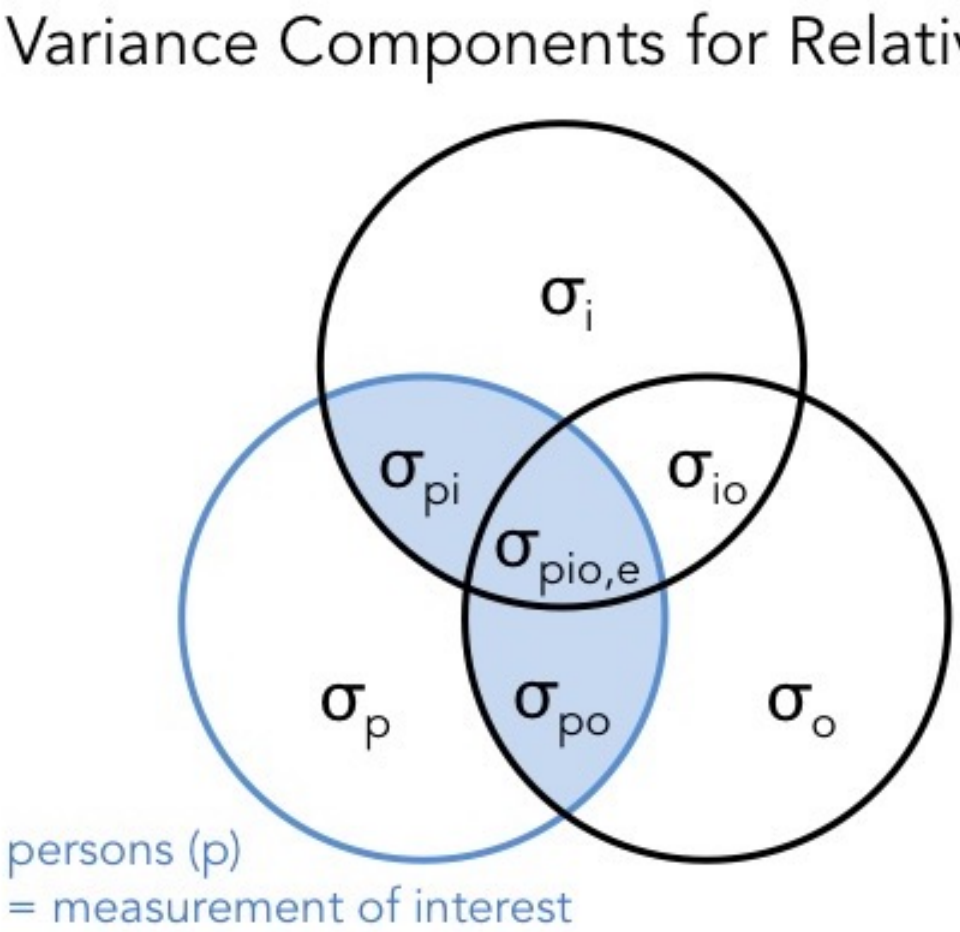
15) **Write out the full equation for the relative error variance, $\sigma^2_\delta$**

As a rule, we know that any source of variability that intersects with the object of measurement (in this case persons) will change the measurement's relative position. In the case of a $p \times i \times o$ design, the relative error variance will include $\sigma^2_{pi}$, $\sigma^2_{po}$, and $\sigma^2_{pio,e}$ (see Figure). These variance components refer to single-unit replications, so we must divide by the relevant number of items and occasions to obtain error for average scores (over items and occasions). So the full equation for relative error variance is:

$$\sigma^2_\delta = \frac{\sigma^2_{pi}}{n'_i} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pio,e}}{n'_i n'_o}$$

where $n'i$ and $n'o$ are the number of items and number of occasions of a hypothetical (') test design (what *could be* rather than what *was* in the G-study data).

From Question 10, we know that $\sigma^2_{pi} = 0.487$, $\sigma^2_{po} = 0.102$, and $\sigma^2_{pio,e} = 0.741$. So we can calculate $\sigma^2_\delta = \frac{0.487}{12} + \frac{0.102}{2} + \frac{0.741}{12 \times 2} = 0.122$ for 12 items and two occasions.




The relative error does not include error terms $\sigma_i^2$ or $\sigma_o^2$ because variation across items or occasions, respectively, are the same for every person. Items that are more difficult will be more difficult for all persons and occasions whith distractions will be more difficult for all persons. Therefore, they do not affect the relative position of one person to another. The same is true for variability for item-occasion interactions ($\sigma_{io}^2$).

16) **Calculate the generalization 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.**

The generalization coefficient for relative error, $E\hat{\rho}^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2}$.

And from Question 15, we know $\sigma_p^2 = 0.573$. Plugging in these numbers into the above equation we get:

$$E\hat{\rho}^2 = \frac{0.573}{0.573 + 0.122} = 0.824$$

17) **Use the "$p \times i \times r$ 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.**

```
relative_error <- function(mod,raters,items) {

  sdvar <- as.data.frame(VarCorr(mod))
  sdvar$varComp <- sdvar$sdcor^2
```

```
  p <- sdvar$varComp[sdvar$grp=="person"]
  o <- sdvar$varComp[sdvar$grp=="occasion"]
  i <- sdvar$varComp[sdvar$grp=="item"]
  pxi <- sdvar$varComp[sdvar$grp=="pxi"]
  pxo <- sdvar$varComp[sdvar$grp=="pxo"]
  oxi <- sdvar$varComp[sdvar$grp=="oxi"]
  err <- sdvar$varComp[sdvar$grp=="Residual"]

  rel_err <- (pxi/items)+(pxo/raters)+(err/(items*raters))
  return(rel_err)

}
```

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?**