S-061 Assignment #1

One person per partnership should submit this completed assignment as a .pdf, .doc, or .docx to Canvas. Make sure you have pasted your annotated .do file to the end. The assignment is due on Tuesday, September 19, at 12PM sharp. We will not grade the .do file or your R code (we'll use it to understand errors if they arise), but of course we recommend annotation regardless.

Collaboration across different partnerships is strongly encouraged. However, code, written work, and calculations should be those of listed pair of authors alone. I encourage you to meet face-to-face and not simply exchange answers via email. Try to work in the true spirit of partnership. You may type answers directly into this document, or you may write in a separate document with numbers referring to the prompts below.

Please scroll to the end of the document and write your names there, to facilitate blind grading.

Part 1: Classical Test Theory and Validation

This scenario is motivated by a real question that came up at a Technical Advisory Committee in Iowa City. I've changed some of the details due to confidentiality agreements, so let me be clear that this specific scenario does not necessarily apply to any real company. Nonetheless, the core principles in this scenario apply.

An essay component of a college admissions test consists of a single essay prompt whose response is scored by a proprietary computer algorithm—an "electronic rater." Let's call this algorithm, "e-Rater A." The testing company then acquires a startup company that has developed a second algorithm; let's call it, "e-Rater B." For our purposes, *assume that e-Rater A and e-Rater B are equally accurate and drawn from a population of similar e-Raters*; they function essentially like two equally-trained human raters. The testing company decides to have both e-Raters rate every written essay. As you would expect, the ratings are not all the same. A researcher at the company reports that the correlation between the two ratings is 0.7.

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 implements an audit policy whereby a random 10% of all essays are rated by a well-trained human rater. The purpose of this procedure is to audit and track e-Rater performance. For this question, assume human raters are essentially like e-Raters, with ratings equally accurate and equally covarying with those of e-Raters and other human raters (although, of course, determining this is the audit).

   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

In a pilot test for a new scale for *grit*[1], a group of 25 participants fills out a 12-item scale on 2 occasions separated by three months. The item score scale ranges from 1 to 5 (see a sample questionnaire in the footnote, though note that the data you have are already "polarized" so that 5 is always "grittier"). The data are included in Assignment1.dta. The same 25 participants complete all 12 items on both occasions, and the same 12 items are used on both occasions.

Note that the data are "double wide," where every row is a person, and x_o2_i8, for example, corresponds to scores on the eighth item on the second occasion.

[1] Duckworth, A. L., Peterson, C., Matthews, M. D. & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology. 92*, 1087-1101. Link. You've filled out a shorter (8-item) version already, but see a sample 12-item questionnaire here:

5. Using Stata, calculate coefficient alpha for the first occasion and the second occasion separately. Use code like, `alpha x_o1_i1-x_o1_i12`. Report both coefficients. In a sentence or two, interpret coefficient alpha for the first occasion (see also Question 16).

6. Using Stata, calculate the average score of participants from the first occasion, then calculate the average score of participants from the second occasion. Use code like, `egen avgscr1 = rowmean(x_o1_i1-x_o1_i12)`. 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).

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`.

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.

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.

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 $n_o = 2$.

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.

Part 4: The Decision Study

15. Write out the full equation for the relative error variance, σ_δ^2.

16. Calculate the generalizability coefficient for relative error, $\mathbf{E}\hat{\rho}^2$, 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?

Write your name(s) here:

_____ _____