Responses to Final Editorial Comments from Rob Gould

Comparison of Learning Outcomes for Simulation-Based and Traditional Inference Curricula in a Designed Educational Experiment

\* **Responses in bold**

Let me begin with a few high level things before moving on.  These high  
level issues are my own opinions that might result in some changes in the  
paper. I'm not insisting on these, but would like to be assured that they've  
all been considered.  
1) Big Question: This is our first randomized study and the most empirical  
we've published.  There just aren't that many coming our way.  For this  
reason, we have the opportunity to set a precedent.  So would you consider  
supplying one or all of the following:  (a) your data set (b) your R code  
for the analysis?  Not sure what your IRB would say about (a), but there  
might be ways of assuring them.  These would be attached as supplementary  
files, and not folded into the paper itself.

**a) The IRB documentation stated that the data would be deidentified then stored on a secure computer. Based on this wording I don’t believe we are allowed to disseminate it freely.**

**b) The R code is all embedded in an .Rnw document which is used to compile the paper using the knitr package. I am happy to include this file that contains the code that cleans the data, generates the tables/figures, and conduct all modeling. The simulation study was housed in a separate R code file (because it took far too long to run each time the pdf document is compiled) which has also been included as a supplemental file.**

2) Have you done a simple t-test (or perhaps more appropriately, a  
permutation test) between the two groups for each outcome variable?  This is  
a nice thing to do and provide.  If the results are significant, then it  
removes suspicion that findings might have been due to a "fancy" model.  If  
they are not significant, but the model finds some significance it leads, at  
least in my mind, to some deeper understanding about the role of the  
covariates.  (As I recall, this was one of David Freedman's favorite 'smell  
tests'.  If a model predicted that eating nuts was good for you, but you  
couldn't see the difference in a t-test or some simple procedure, then the  
effect either didn't exist or was small enough not to adjust your lifestyle  
for.)

**We checked the simple t-tests out of curiosity and they were not significant at alpha=.05. However, the research plan from before the experiment was to use a model based approach since we knew that we would have a half semester of information that we could use to control for pre-treatment differences in each curricula group. It turned out to be the right choice because the random assignment had placed noticeably stronger students in the traditional group and the final models found midterm and lab 5 to be significant pre-treatment covariates to account for this imbalance going into the inference unit treatment.**

3) Have you considered fitting a random effects model (or hierarchical model  
or nested model or whatever you want to call it)?  If not, you might try.  
Ise are rather commonly applied to classroom data.  And the first question  
that arose to my mind when you added random effects to your simulation was  
Why not account for them in the first place?

**With only one classroom for each treatment we are unable to estimate of both the fixed effect of treatment and random effect for classrooms. The simulation study was able to assess error under various inter/intra classroom covariance structures only because we assumed the values of the random effects that we cannot estimate from the study data.**

4) Your discussion in the conclusion was very nice and thought-provoking.  
My own thoughts, while reading the paper, was I wonder just what we \*are\*  
expecting students to learn from this approach. Not "what are they learning  
compared to traditional" but "what are they learning"?  This approach  
involves computation: if students had a computer before them during the exam  
and could run simulations, would they have done better?  (Or worse? ) Do  
students understand what is happening when they run the computations anymore  
than someone who pressed the "return" button to fit a model?  You found that  
 Simulation students who were given traditional problems did ok, but that's  
lopsided. What about giving traditional students simulation problems to see  
how they do? I suspect much worse.

**This is an interesting question for a future study, but we don’t wish to postulate on the hypothetical outcomes using the results of this study.**

OK, here are smaller things.  
1) Consider moving the test items to a supplementary file.  People will want  
to print this out, and with the appendices, there's lots to print. I would  
keep the diagnostic reports on your models, though.

**Done, the exam question are now attached in the file labeled *SupplementalMaterialsMaurerLockStudyTISE.pdf***

2) page 3, first full paragraph: There's something funny about the tense in  
this pragraph. Seems like when you're talking about older studies you should  
use the past tense. AS in "This study did not however attempt" rather than  
"does not attempt". This also makes it easier to know which study you're  
referring to when you say "this" study.  Throughout this paragraph, you  
could use some careful checking of the verb tensse.

**Edits - done**

3) Same place, and throughout:  When you have more than two authors in a  
citation, use et. al.  So, I'd said something like Tintle et al (2011) found  
weak evidence....  Tintle, etl al (2012) then found....  
Two authos get an ampersand.

**Edits - done**  
  
4) Please use data as a plural.  I try to be consistent in this myself,  
though sometimes fail. It may be an archaic editorial distinction, but I am  
fond of it. So first paragraph under Methodology: "..there data were  
omitted..."

**Edits - done**

5) Same paragraph under Metholodgy:  You present the cohort labels out of  
alphabetical order, and I couldn't see a reason for it. IF there is a  
reason, please state it. If not, change to alphabetical.

**The ordering of the cohort labels was challenging due to the fact that there was not arrangement that would present the cohorts in alphabetical order simultaneously with respect to groupings within curricula, classrooms and lab rooms the mixture of class rooms and times. The reordering the labels in the first paragraph leads to a more confusing organization in the diagrams and discussion that follows.**

6) pg 4 AND throughout: you might make graphs slightly larger so they are  
mor legible. You might also take out the shading (or use lighter shading).  
The words in Figure 1 ("9. Sampling Dist. 10. Boostrap..." ) are hard to  
make out.

**I have changed the box shading to make it clearer to read.**

7) the sentence that begins on pg 4 and ends on 5:  I don't like the phrase  
"display concepts".  HOw do you display an abstraction.  Please be a little  
more precise in what is being displayed.

**Edits - done**

8) I think this graph could be enhanced by labeling the upper graphic  
"Lectures" and the lower "Labs", or something similar. I spent some time  
searching through the text to figure out the difference.

**Edits - done**

9) Figure 3: I can't read the words in the margins.  Maybe try lighter  
shading or bigger graphs.  (The advantage of an electronic press is we don't  
worry about page counts so much.)

**Edits - done**

10) pg 8 first paragrpahn under "Analysis":  "our data include..."

**Edits - done**

11) pg 8 send paragraph under Analysis:  I suggest a new paragraph at "The  
bivariate MANCOVA...."

**Edits - done**

12) pg 9 first paragraph.  A small thing, but maybe the word "acknowledged"  
is better than "understood" for "The assumption of independence....is  
understood to be.."?

**Edits - done**

13) Can you please say a bit more about how covariates were removed  
according to the collinearity criterion?  Did you look at variance  
inflation? Or something else? Was there a cutoff?

**We checked the final model with both with and without hw5 due to our concern for collinearity. When hw5 was present, neither hw5 or lab5 were significant; when hw5 was removed, the magnitude of effect for lab5 more than doubled and became significant. This was taken as an indication of collinearity and hw5 was removed. There was no formal test conducted in this model decision. It is noteworthy that the p-value for the treatment effects were *weakened* by removing homework 5, so the decision was not favoring a significant result.**

14) pg 10, 2nd paragraph of 4.2:  Table 4 shows -- based \*on\* the Pillai's  
...."  Also, in the next sentence, I think "less" should be "fewer" as in  
"had three weeks fewer of coursework..."

**Edits - done**

15) Page 11: sentences after items 1 and 3 (the formula) in your list.

**I am not sure what needs to be addressed here.**

16) pg 13, 2nd paragraph:  hyphenate "between-labmate covariance".

**Edits - done**

17) Please proofread yourself.  Make sure references are correct (and that  
all citations have references and vice versa).  After this, I'll proofread  
one more time, give you a chance to look-over, and then we publish.

**Sounds good.**