Psy 600K Midterm Exam

To whom it may concern,

Thank you for your submission. Below, I have listed 8 Flaws within your manuscript that I would like you to address. I have outlined each flaw into 3 sections. The first, I create a **Description of** the flaw, next I offer an **Explanation** ofthe flaw, and lastly, I offer a **Suggestion** to fix the flaw.

**Flaw #1: Domain sampling of the items**

* **Description:**
  + The items of this scale (referred to in Table 1; page 4) are completely polarizing to the entire female gender. Furthermore, it makes assumptions that the reader is a male. For example, in item 4, “A good sales***man*** knows how to make a woman like ***him”***.This may not have been a problem if it was appropriate to assume that the sample, and population of interest, was 100% male (and their sexual orientation leads them to like women) – However, I assure you this is not the case. In fact, the majority of your sample was female (67%).
* **Explanation**
  + I would like to explain the idea of *domain sampling* to you from Raykov & Marcoulides (R & M;2011; pg. 5) textbook. R & M explain that creating a measure construct, it requires the construct to appropriately test a sample from that domain. A problem with your measure, is that your test is meant to embody *people*, but the entire measure is focused on males. According to Robert DeVellis (2012; 3rd edition; pg 81), items such as these leave individuals in a quandary. Think about the consequences of this, women who answered this question cannot appropriately respond on your provided scale because they are not a ‘him’. Therefore, 67% of your sample has been completely struck down by your measure.
* **Suggestion**
  + I would highly encourage you to completely re-word all items that polarize a gender. In context of your construct of a “*People* Person”, you need to accurately lay out that everyone that would be included in your sample (to represent your population) can answer these items appropriately. In fact, please reword all analyses containing gender specific pronouns as they are polarizing towards the test taker. Please use words such as “people”, “individuals”, and “others”. I would also advise that you, at bare minimum, mention these gender polarizing items in your limitations of your paper. However, I would even go on further and say you should completely redo the analyses without items that contain these items. Even consider resampling with newly written items overall. This suggestion, and description applies to items: 3, 4, 9, 10, and 14.

**Flaw #2: Factor loading reports**

* **Description**
  + Please report *ALL* factor loadings. No factor loadings are *trivial* as you stated in this manuscript. More importantly, your statement of, “Although a number of items had substantial loadings on more than one factor, it seemed clear to us that the three factors represented distinct constructs” does not have me sold. What was your criteria for determining distinction? Although this can be subjective to a degree, I am fairly confident that anyone in the field would argue that a difference in factor loadings of 4.51 to 4.88 in factor 1 & 2 respectively is fairly close.
* **Explanation**
  + Unfortunately, the problem with items loading onto multiple factors, as your items do, is that now it is hard to tell which factor your item responses are being influenced by. Factor loadings can be defined as a reflection of the, “…important relationships between factors and observed measures” (R & M, 2012; pg 43). Therefore, even though you defined clearly distinct factors from your analysis, the relationship between some of them are much too close to really be able to accurately define which factor is influencing the item responses.
* **Suggestion**
  + I really recommend that you assess these items once more. It is very likely, that due to these very similar factor loadings (specifically in items, 1, 2, 4, & 15) you do not have an accurate understanding of how your factors are influencing the item response. I really would not feel comfortable interpreting these items overall and you should consider having trepidations in your interpretations of these items as well. Additionally, consider removing these items entirely from your scale. They are not great at making a prediction towards any distinct factor in your analysis.

**Flaw #3: Interpretation of screeplot**

* **Description**
  + You interpret the scree plot by indicating that the three factor solution all yielded Eigen values over 1 and therefore, that is the best solution. However, looking at the scree plot, I do not think this is the interpretation that you should have gone with.
* **Explanation**
  + The problem with this interpretation is that yielding Eigenvalues over 1 is not the primary way to understand the number of factors. Although it is one way, it is definitely not the primary way. It appears you did not even look at your scree plot because after 1 factor, your factors do not explain much explained variance as you increase the number of factors. I noticed that you went with the SPSS default in factor extraction – Which I will tell you is not the best method for factor extraction. Many papers often find that the default methods used in SPSS (and SAS) have been found to be extremely inaccurate (EFA slides, slide 7, 02/27/18). It is important for you, as a knowledgeable researcher to make decisions yourself. To be honest, there are multiple methods for choosing the correct factors for your measure when running an exploratory factor analysis. However, choosing only one method, and that method being a computer driven method with no thought, does not seem to serve you well.
* **Suggestion**
  + Please re-look at your screeplot. Then, assess how it looks. Personally, as I view it, I see that moving from one to two factors does not help to explain much more of the variance in your measure. Often, a goal in building a model through factor analyses is creating a model that is both parsimonious and explains the most amount of variance (EFA slides, page 3, 02/27/18). As you can see, this is a tradeoff. Looking at your screeplot, I see that there is a large “bend” going from the eigen value of 1 – 2 factor numbers. This may serve as a better cutoff point. Because your model is quite parsimonious and explains a large amount of the variance. At least, going from 1 – 2 factors does not contribute much more. A problem I see is that you are at risk for having a lot of unexplained variance overall in your model. You must express this in your limitations. It is possible that some of the bad items (Please refer to Flaw #1) in your scale are contributing to this unexplained variance and you may want to consider removing those and re-running the EFA. Lastly, another potential strategy to help guide your factor analyses interpretation would be to conduct a parallel analysis against the scree plot data. This compares the eigen values of your data matrix a randomized dataset with random values. However, because these are so highly correlated this is may not be the best method, but it is worth a consideration (EFA slides, page 7, 02/27/18).

**Flaw #4: Taking the total scale mean, rather than item mean**

* **Description**
  + When you are reporting the scale statistics, please do not report the mean of the entire measure, along with those standard deviation units. Please indicate what the average item response score was *per scale* (which will yield a number from 1-5). Furthermore, it would even make sense to see what the mean score for each item was as well. Especially considering that this is a multidimensional construct.
* **Explanation**
  + Because you are claiming that you have a multidimensional construct, assessing the overall mean of the pooled items is not appropriate. It is very possible that individuals may score high on subscales of your construct, which will contribute to their overall score, but very low in other aspects, thus, making it hard to understand what aspects of the test taker contribute to their “people person” attribute. R & M (2012; pg 11-12) make a great argument as to why taking a composite score is completely inappropriate in a multidimensional construct because it is simply the sum of values that are additions of only partially related factors. Additionally, by taking the item means, as opposed to the overall scale mean, there is a better understanding of the item difficulty. R & M (2012; pg 15) explain item difficulty as a good parameter estimate from your sample. Please keep in mind, that in a Likert scale such that you have, this difficulty estimate helps to get at the extremeness of the item, if an item has a higher mean, then the people that are responding at a very high level and are less likely to score on the lower end and vice versa (Critical Statistics slides, pg. 2; 02/06/18). As your reviewer, I need to see how people on average score on the items in order to understand how individuals are likely to respond. This provides insights to how the item reflects your construct. Additionally, I think you will find some interesting insights to your own scale because, there may be instances where all Reponses for certain items are very high or very low, thus creating an understanding as to how individuals are responding in general.
* **Suggestion**
  + I would like this to be reported in two ways. Please show the item means individually and the average response number for each of the 3 factors you determined via you analysis (again, these should both fall between 1 and 5). Then, please comment on any means that seem extreme and explain why you think that these means are as so. Doing this can help to understand what it is about what kind of predictions can be made about the population you are trying to capture using this scale.

**Flaw #5: Reliability inference**

* **Description**
  + Your reliability interpretation is wrong. You are not considering all of the factors associated with reliability. Furthermore, there are other ways of looking into reliability as well.
* **Explanation**
  + To start, you need to specify the type of reliability you are looking at when writing a manuscript. From what I can tell, you are interested in *Test-retest reliability* by getting the correlation of item scores from time 1 to time 2. However, you have failed to understand that there may be other aspects related to the correlation of your two scores. I am curious how you determined the appropriate length between time 1 and time 2. It appears it was the beginning and end of the class, but, is that necessarily a good length of time to be determining the reliability of the measure? R & M (2012; pg 148) explain that depending on the construct of interest, time plays a major role in the understanding of this coefficient reliability. Furthermore, there are more reliability coefficient measures to consider than just the test-retest reliability. For example, the Cronbach’s alpha reliability coefficient should be reported as well. The Cronbach’s alpha is a measure of the proportion of explained variance within your items (DeVellis, 2012; 3rd ed.; pg. 36) and therefore helps the reader understand the reliability of the test. Furthermore, this is a common standard in the literature, and you should follow suit to be taken seriously in the academic research field.
* **Suggestion**
  + Please list the limitations of your test retest reliability coefficient and, if possible, please assess the how the population may be affected by taking the measure at different time points as well. Additionally, it is appropriate to report the Cronbach’s alpha, so that I may see the internal consistency of your items. Please report the alpha separately for each scale. This can help guide your own understanding of your scales as well.

**Flaw #6: Interpretation problem #1, class grade outcome is not reflective of interpersonal skills**

* **Description**
  + I have concerns understanding how your criterion variable to determine the validity of this scale is in any way sensible. Doing well in a marketing skill seems quite independent from ability to be a “People Person”.
* **Explanation**
  + The problem here is your choice to determine the validity of your measure. There is a large difference between doing well in a class and being able to properly be a “People Person”. I am questioning how you even got to this conclusion. A criterion variable is chosen based on a “gold standard” empirically related to your latent variable of interest (DeVellis, 2012; 3rd ed.; pg 61). I even would go as far to say that you are not necessarily looking at the correct type of validity in your analysis. Please review *concurrent validity*, which is focuses on the time effect of studying in a class then relating that to the scale (DeVellis, 2012; 3rd ed.; pg. 62). In general, I have problems believing that what you have identified as a variable to identify validity is truly appropriate. In a sense, you are comparing apples and oranges, by comparing being a people person to class performance.
* **Suggestion**
  + I would highly recommend removing these sections altogether. The conclusions that are made because of your relationship between the outcome and scale of interest are hard to believe. Unless you are able to find me a theoretical framework and a series of scientific research that identifies this as a good comparison of being a “People Person” then I will not accept this in the manuscript. I want to remind you, that even though you are running an *exploratory* factor analysis, your explanation of the construct needs to make sense in accordance to theory. Or, at minimum, you need to input that this is *not* basedon theory and ensure that readers of this manuscript are aware.

**Flaw #7: There is no evidence to support your factor extracts**

* **Description**
  + Based on your factor analyses, you created 3 factors: *Popularity factor, People skills, & Sociability.* However, you have no theory or real rationale to support these factors that you have created. I truly have a problem with this, just because this is an exploratory method, does not mean it should not be grounded in practical theory.
* **Explanation**
  + Starting with your popularity factor, I need to disagree with this. “People seldom don’t do what I ask” and “People like me” certainly do not seem to be getting at the same latent variable. Have you considered the idea that some of these items might map better to different constructs? It’s possible that all of these are correlated, but that is because the underlying latent variables that are influencing these item responses may be correlated. This is related to the concept of *Content Validity*, which is, “…inanimately linked to the definition of the construct being examined” (DeVellis, 2012; 3rd ed., pg 60). Therefore, it is important that these items truly match what the underlying construct is. I don’t exactly understand how doing what someone asks has to do with popularity. Or really any of the other items in your scales from a theoretical perspective.
* **Suggestion**
  + One of the best pieces of advice I can give you is to do your research. Look into articles that describe “People Person’s” and report back with more evidence to explain to me (and any other potential reader) how it fits into this factor. DeVellis (2012, 3rd ed., pg 60) also emphasizes that an important way to assess how these items fit would be to work with a colleague that is well versed in the domain you are considering.

**Flaw #8: Interpretation problem #2, r of 0.09 not good. Despite significance**

* **Description**
  + Having a significant correlation is not necessarily indicative of a good correlation. Your p-value may be less than 0.05, but your correlation is only a 0.09.
* **Explanation**
  + Therefore, if you want to understand the amount of explained variance from your correlation, then you really are only explaining about 0.81% of the variance in your prediction of PPT to course performance. Significance testing aside, is that truly a good amount of explained variance? In measurement, we are interested in the effect size of our correlation rather than the p-value (Critical Statistics Slides; pg. 6; 02/6/18). Unfortunately, there is no true agreed upon value of a “good” effect size but understanding the background literature in your domain is a helpful start. Furthermore, I would like to reference DeVellis’ (2012, 3rd ed.; pg 62) explanation of validity in which he proposes that having a good correlation with the criterion variable does not necessarily infer predictive accuracy. Thus, I have problems with your conclusion over all. Just because the measure is correlated with the outcome in a positive fashion does *not* necessarily mean it is doing a good job of predicting the outcome.
* **Suggestion**
  + Overall, I would not use p-values to guide your interpretations of this test. P-values are not quite as useful in the testing and measurement world as identifying items that are completely uncorrelated (which serves as your null hypothesis) is not a realistic scenario. Therefore, please report effect sizes and talk with people in the field what their interpretations of your statistics are. This will help to lead you to making more appropriate interpretations. Furthermore, DeVellis (2012, 3rd ed.; pg. 59-60) and I would both advise you to talk with subject matter experts to look at your items and consider completely revising your scale. This is an extremely important aspect of building a measure to understand your latent variables and will give more credibility towards the understanding of these factors and items.

*Your manuscript has been denied.*

Thank you for your time,



Neil Yetz