

Point-by-point responses to reviewers' comments in italics

Nursing Research Manuscript # NRES-D-09-00080 "A multilevel confirmatory factor analysis of the Practice Environment Scale (PES)"

Reviewer #2:

This case study brings up a very interesting and important analytical issue by using multilevel instead of single level approach to conduct a confirmatory factor analysis for instrument validation. The authors justify the importance of using this multilevel approach. They demonstrated conceptually and empirically, the similarity and difference between using single level and multilevel confirmatory factor analysis. The paper is informative and well written.

We thank reviewer for positive feedback.

In addition to the above strengths of this paper, a few comments and suggestions to authors for further improvement.

Reviewer brings up important points that we have attempted to address under the page restrictions of the Methodology section of Nursing Research.

1. Even though the authors used empirical data to compare the findings between single level and multilevel CFA, the primary focus of this paper is on conceptual level rather than practical level of how to analyze the data. Readers who are interested in this topic may be also interested in knowing how to actually conduct the analysis. The authors only used one sentence (lines 4-5 p.7) to describe how they analyze it. In addition to say that Mplus was used to analyze the data, there was no information as to how the data set was formatted, what models were used and how data were analyzed. If the space allows, it will be helpful to expand this part a little bit more; otherwise, references regarding "how to" issues will help reader to know how to conduct such analysis. For example, Muthen (1991), Journal of educational measurement or similar analyses by using different software.

Reviewer requests more information on practical training for fitting models proposed in our paper so readers can come away knowing how to fit such models. This is an important point that we address within the page limitations by referencing Muthén (1994) for more theory and www.statmodel.com for "how to." The latter is an excellent resource for learning multilevel CFA in general with principles that can be applied in any standardized latent variable software.

2. A very important part of this paper is the interpretation of the factor loadings (Bs) from a single level model and a two-level (within and between) model. The authors used a full page (p.8) to discuss this issue; however, the second half of p.8 needs clarifications. (a). In order to explain the difference in inference for B's of the two-level and single level models, the authors led the readers to Figure 4. But only one sentence was used to explain what do those light and dark squares mean without further explaining the major ideas the authors would like to convey to

readers in Figure 4. (b). In the following sentence, the authors talked about the trend, which has nothing to do with Figure 4 but moved to Table 3. It will be helpful to add a sentence or two to explain what important information does Figure 4 provide, otherwise, it is not necessary to have this Figure. (c). I would like the authors double check on the next statement. I understand that the Zs from between should be smaller than those Zs from within because the sample size and standard errors. Would the sample size also influence on the Bs? Why? (d). I am not clear what empirical evidence the authors was used to state that between convergent validity is established? If statistical significance of the factor loadings is the only reason, the authors need to consider such significance is due to the large sample size. ($J=4,783$). With a small number of units, the same loadings may not be statistically significant. This leads to my next suggestions.

We agree that this portion of the paper could be written more clearly and attempt to do so by addressing each comment separately.

(a) Reviewer feels that Figure 4 needs more explanation to become useful to the reader.

We agree. So a decision needs to be made whether to expand on Figure 4 or drop it and focus on Table 3 since it already has Figure 4 information.

(b) Given the page constraints of the Methodology section, we have decided to drop Figure 4 and center the discussion surrounding Table 3.

(c) Reviewer challenges our statement regarding the influence of sample size on B. We agree that this is not the case, thus we have deleted this relationship from the page by deleting B from the sample size explanation (keeping its change due to correlation).

(d) Reviewer wanted more information on the convergent validity. Yes, reviewer is correct that we are using statistical significance of the factor loadings as the sole justification of convergent validity. We agree that sample size plays a major role and so we have added a sentence to this paragraph regarding this issue.

3. I will suggest the authors to expand their discussion on situations when the second level sample size is not large enough, or give some recommendations under what situation the two-level CFA will not be recommended.

Reviewer's advice on addressing sample size recommendation is important. We have added our suggested guidance of 5-10 second level subjects (units) per parameter to the Discussion section.

4. The authors may give some thoughts about the interpretations on these within and between level's Bs? Readers may be interested in knowing it is always true that within loadings be higher than between loadings or other way around is possible and what do these mean to a researcher? Or how to interpret the situation when within loadings are significant and not between loadings?

Reviewer would like us to further discuss the interpretations of the differences in Bs within and between. While this is an important issue, the page restrictions of the methodology papers keep us from elaborating. We hope that readers gain an appreciation of the connection of ANOVA and multilevel CFA in the case of clustering and will motivate to read and think more about the issues such as further interpretation of Bs.

Minor suggestions.

1. Provide a reference for $ICC > .001$ use multilevel (line 23, p. 4).

This was a misprint as it should've said 0.10 as noted (with a reference) in the results section. This has been fixed.

2. Line 17, p. 5 ".enjoy good statistical properties, " such as.

We indicate now that it is "fast and accurate."

3. Line 20, p. 7, more accurate for n should be n subscript of \bar{j}
Change = 15.24 to ? 15.

Reviewer is technically correct but we have indicated that this sample size is approximately equal to 15.24 so we preserve the original statement.

Reviewer #3: Review of Manuscript # NRES-D-09-00080

"A multilevel confirmatory factor analysis of the Practice Environment Scale (PES)"

This paper is intended to describe multilevel confirmatory factor analysis using data from the NDNQI for the Practice Environment Scale and the Job Enjoyment Scale. The investigators describe the measuring instruments and then the step by step use of a multilevel process to show validity at an individual and a group level. Strict adherence to this purpose varies through the article and the conclusion seems to be more of an argument for the tool than the method. This article could be a very valuable methods piece for the readers of Nursing Research. Understanding these methods is difficult but very necessary for nursing researchers using advanced statistical analysis for their work. This reviewer does not fully understand the method but approached the article to learn more about this method. To that end, the following suggestions are made (by someone with imperfect understanding seeking to learn).

Reviewer recommends that we make strict adherence to the purpose of the article "... we aim to demonstrate a multilevel CFA using NDNQI data as a case study." In an attempt to respond to this reviewer's suggestion, we have revised the paper being mindful of the page restrictions set by the methodological section of Nursing Research and comments from other reviewers.

Ways in which this manuscript could be improved.

1. Clarify the purpose - is it the method or the PES - and stick to that purpose throughout. For example, the introductory paragraph indicates that the crucial problem is the measurement of RN satisfaction and the quality of care. While this is a very important problem - it is not the focus of this paper.

Reviewer would like us to clarify the purpose of this paper upfront and stick with it throughout. Embedded in the original version was the purpose, but it was located several paragraphs into the paper and perhaps needed to be upfront to make it clear. We have moved the purpose sentence to the first paragraph that explicitly states the aim of the paper. We discuss a multilevel CFA using a case study to hook the reader. Therefore we feel we needed to preserve the discussion of measurement of RN satisfaction and quality of care – a vitally important issue to researchers in nursing.

2. The initial explanation of multilevel confirmatory factor analysis is excellent. I would suggest re-wording the sentence from line 12 - 14 on page 5 to be "First, a multivariate within covariance matrix is calculated by summing the within covariance matrices from all of the units."

We have replaced this sentence as reviewer suggested. This new sentence is clearer.

3. Also, the explanation of different types of validity is good and could be expanded a bit to help the reader apply this to the CFA.

We thank reviewer for positive feedback but have no room under the methodology restrictions to add more validity information.

In the results section

4. Please explain (or perhaps delete) the statement about the STD range being under what would be expected. Is the short range relevant to this paper? If it is, explain how you determine this expected range and why this finding is important.

We have cut the portion of the sentence referring to STD.

5. Explain the meaning of intra-class correlation coefficients (ICCs) larger than 0.10 requiring multilevel modeling. What exactly does an ICC >0.10 mean? From the earlier explanation I would guess that it mean increasing similarity among RN respondents from the same unit. But, earlier the cut-off was given as 0.01 (pg 4 line 23). What is the range of an ICC? If it only ranges between 0.0 and 1.0 and an ICC of <.01 indicates a need for multilevel modeling, is it possible to get an ICC that doesn't indicate the need for multilevel modeling?

Reviewer questioned the inconsistencies of the quote of ICC. We fixed the inconsistency was a typo) to say that 0.10 is the cut-off in all parts of the manuscript. In the original paper we stated that "(ignoring the ICC will result in 50% inflation of perceived information, see for example Campbell, Donner, & Klar, 2007)." Perhaps this is too late in the paper for the reader to appreciate. Therefore we moved this parenthetical statement earlier in the manuscript (where we first mention ICC cut-point. We believe these two modifications have helped clarify the paper.

6. Indicate what the "PesSR01" (on page 7) is to help the reader quickly grasp the sentence; e.g. "as illustrated with the first item in the staffing resources scale, PesSR01", . .

This has been edited.

7. Little assists to help the reader quickly grasp things could include:
In the title of Table 1 indicate that the n of 15.24 means average unit size and J is the number of units. This is explained in the footnote to figure 3 - but needs to be in that first table as well. In the title of Table 2 - say Item PesSR01,

We have edited as suggested.

8. Last sentence of each paragraphs on page 8 refers quickly to the CFI and RMSEA evidence for factorial validity. I suggest including the CFI and RMSEA in the Table 3 and adding an explanation in the paragraphs.

We have edited as suggested.

9. I suspect that the graphs in Figure 4 could be very helpful with more explanation. What does it mean that the "two-level model reduced the within Bs and Zs"? Perhaps that the size of the correlations between the items and the domain are smaller and the statistical significance is smaller? With a Z cutoff of 1.96 for statistical significance dropping from 180.3 to 173.8 doesn't make much difference.
Also - what does it mean that "The within versus the between Vs and Zs drop substantially"? Is that what is shown in the graph? More description and explanation is needed.

Our explanation of Figure 4 was awkward; also indicated by another reviewer. We dropped Figure 4.

10. Regarding the correlations in Table 4 - purportedly showing discriminant validity - what is a reasonable cutoff for convergent/discriminant validity? The authors point to only one correlation, of .92, as an indicator of non-discrimination; however, there are other high correlations. More information about making these kinds of judgments would be useful if you are going to include this as evidence.

Reviewer questioned our lack of a cutoff for discriminant validity. We have fixed this by stating a correlation less than 0.85 has discriminant validity. We have also noted that both between and within correlations of the latent variables Hospital Affairs and Quality of Care.

11. If the authors retain the argument that correlations among the PES subscales and Job Enjoyment measure indicate evidence for validity- more information is needed here. At the least a correlation matrix among PES and JE subscales. It is also possible that this part of the analysis is not needed for a paper with the purpose of describing multilevel CFA.

Since we are correlating the PES score as a second-order factor when going against PES the correlation of just these two latent variables sufficiently describes the correlation matrix and its correlation with subscales is not necessary.

12. The Discussion section needs to discuss the results. Currently the first paragraph of the discussion presents considerations for the method that should come before the presentation of results. The second paragraph could fit into an implications section.

We have added considerations of sample size in a relabeled section called “Discussion and/or Implications” and believe addresses reviewers critique.

Reviewer #4: The publication of a methods paper using multilevel CFA is inevitable and relevant to nursing research given the inherent hierarchical structure of nurses' data, which is nested within units, nested within hospitals/facilities.

This manuscript has the potential to make an important contribution to methods used in nursing there are several key points that the authors need to attend to before this paper should be considered for publication.

Thanks to reviewer for comments. In an attempt to respond to this reviewer’s suggestion, we have revised the paper being mindful of the page restrictions set by the methodological section of Nursing Research and comments from other reviewers.

The introduction and background of this paper present information about the data sources used in this study and the rationale for using this methods. Several sentences in this section are very unclear and require rewriting.

This was a difficult comment to act on. Perhaps this reviewer was commenting the lack of purpose upfront. Attempting to respond we decided on the following. Embedded in the original version was the purpose, but it was located several paragraphs into the paper and perhaps needed to be upfront to make it clear.

The relationship of Job Enjoyment to PES is not clearly described, and the inclusion of JE appears cursory rather than purposeful. JE should be further described along with the reason for inclusion or it should be deleted. I suspect that the authors were attempting to compare these two for purposes of construct validity but this is never stated. It is not clear to the reader if the authors intended the references on page 2, line 20-21 to refer to the OES and JE scales or to the psychometric methods.

Reviewer questioned the lack of purpose for simultaneous inclusion of JE and PES upfront. Therefore in the first page of paper we have added the statement “PES and JE together allow further exploration of their construct validity” to clear the point of their simultaneous inclusion. We reviewed the references that reviewer was uncertain was PES and JES or psychometric. All references have the word “reliability” or “psychometric” in their title so we believe are cited properly next to the term “psychometric methods.”

The reference to the organization as the unit on page 3, line 9 is not clear. Presumably the authors could at some point conduct a 3 level hierarchical model with the organization at one level, units at another and nurses at the third. That would mean changing the definition of organization away from the unit. It would be easier to clarify these definitions now.

Reviewer makes a good point that clarifying the definition of units of analysis is important in the context of a multilevel model. We have added a sentence per reviewer's suggestion.

Methods/Analysis:

The ICC in Table 1 should be defined as ICC1 or ICC2.

Reviewer would like clarification of whether ICC1 was fitted or ICC2. We stay away from this terminology because the ICC we used is defined in Figure 3 using the ANOVA analog.

Multilevel CFA (or SEM) can be analyzed based on the separate covariance matrix for each unit and then reformulated so that it can be treated as a test of covariance structures across two models (i.e. two levels - individual (within) and cluster (Between) level). From this point of view, it is necessary to present the test result (chi-squared test) for model estimation. The authors did not provide this test result; only two fit measures-CFI & RMSEA- which do not provide a test of model fit. The specific results for CFI and RMSEA suggest significant differences still exist between the matrix implied by the 5-factor model underlying the unit structure, and the data matrix. These differences suggest misspecifications in the model, not the data. The statement that the authors "fit" the model (in the abstract and methods) is more appropriately referred to as "estimated the model to determine its fit" with the data.

Reviewer points out that we did no testing across models. We stayed away from testing across models because with such large number of RNs and such large number of units, slight deviations in the models would provide statistical significance. Regarding our use of the term "fit" we believe is a matter of opinion and have seen use of this term in both the way we have described and the manner in which reviewer suggests. We have decided to keep the terminology as is.

The authors showed correlations between 4 latent variables in Table 4. But if we can see the latent intra-class correlation to check the proportion of factor variance that is due to between clusters, then this could be useful to understand the results.

This reviewer and a previous reviewer also suggested showing all of the correlations. But due to a lack of space and the same explanation we gave to a previous reviewer "Since we are correlating the PES score as a second-order factor when going against PES the correlation of just these two latent variables sufficiently describes the correlation matrix and its correlation with subscales is not necessary." we have decided to not show this particular correlation matrix.

The authors indicate that evidence exists that supports the PES as a valid and reliable instrument to measure the nursing practice environment (Lake 2002). Yet they do not present Cummings et al (2005) Nursing Research study where the validity of the PES and other practice environment analytic structures were questioned based on the replication of factor structures that showed that the theoretical model underlying the PES did not fit the data.

We checked Cummings et al (2005) paper that did indeed include nurse working but we could not see its direct relevance to the multilevel CFA described in our paper. For this reason we did not include it in the revision of our paper.

Cummings, G, Hayduk, L, Estabrooks, C (2005), "Mitigating the Impact of Hospital Restructuring on Nurses: The Responsibility of Emotionally Intelligent Leadership," Nursing Research, 54(1), 2-12.

Discussion:

The discussion of the results should be expanded substantially to include the significance of the results, and the additional uses of this method in nursing data. Limitation of study procedures should be presented and discussed.

We appreciate that reviewer would like expansion but page limitations of the methodology section of Nursing Research has limited us to the discussion presented.

And as a minor comment, in page 4 (line 23), authors indicated 0.01 as the cut-off value of high ICC. This number should actually be 0.1.

We have fixed this mistake.

There are several other grammatical terms that need attention throughout the manuscript. For example in the abstract, it should read the "data" were collected... rather than the PES is collected...

We fixed this.

- page 5 line 17 "enjoys" is an emotive term applied to statistical properties.

We changed to "has."

CHECKLIST FOR STYLE

REFERENCES --

After the 6th author's name and initial, use et al. to indicate the remaining authors of the article (e.g., Taunton; Li).

This has been done.

TABLES: Define all abbreviations in tables in a note below each table.

This has been done.

Note: Revisions done in this paper based on reviewers' comments forced us to drop the sentences "One question a reader may have is: given the high between correlation of JE and PES, why not just use the simpler JE scale? One important reason is that the PES scales provide more actionable information for nurse administrators than the conceptually singular JE scale," in the discussion due to page limitations.