

Reviewer Comments:

Reviewer #2: The aim of this paper is to demonstrate how to use multilevel confirmatory factor analysis (CFA) to investigate the validity of the Practice Environment Scale (PES). In introduction section, the authors pointed out even though the PES and Job Enjoyment (JE) scales have been tested via best practice psychometric methods, it is necessary to ensure continued support for validity by further re-evaluation. They demonstrated a method which has not been used in nursing literature, multilevel confirmatory factor analysis to validate PES. The concept of using multilevel CFA is similar to the ANOVA by breaking down the total variance covariance matrix into between and within variance covariance matrices and then to examine the factor loadings using between and within information.

This case study brings up a very interesting and important analytical issue by using multilevel in stead 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.

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

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.
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.

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.

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?

Minor suggestions.

1. Provide a reference for $ICC > .001$ use multilevel (line 23, p. 4).
2. Line 17, p. 5 "...enjoy good statistical properties, " such as...
3. Line 20, p. 7, more accurate for n should be n subscript of j bar.
Change = 15.24 to ~ 15 .

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).

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.
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."
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.

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.
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?
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", . .
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,
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.
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.

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.

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.

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.

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.

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.

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.

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 nursers at the third. That would mean changeing the definition of organization away from the unit. It would be easier to clarify these definitions now.

Methods/Analysis:

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

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.

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.

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.

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.

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.

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...

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

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).

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