# THE IMPACT OF FACTOR NONINVARIANCE ON OBSERVED COMPOSITE SCORE VARIANCES

**W. Holmes Finch[1,*] & Brian F. French[2]**
[1,*]Department of Educational Psychology, Ball State University, Muncie, IN 47306, USA
Phone: (765) 285-8500; Fax: (765) 285-3653; Email: whfinch@bsu.edu
[2]Washington State University

**ABSTRACT**

Assessments are often comprised of several subscales measuring separate but correlated skill sets. Various methods exist to mathematically merge the subscales to create total or factor scores. Two of the more popular methods include unit and regression weighted factor scores. Often these scales are not inspected for factor invariance across groups for which comparisons will be made. The extent to which a lack of measurement invariance influences these different scoring methods is not clear. This study investigated the impact of the absence of factor loading invariance on factor score variances. Results demonstrate that a lack of measurement invariance can have a dramatic impact on group variances for the composite scores, thus leading to a potential violation of an assumption of many parametric analyses. Discussion of the results includes suggestions for researchers to consider measurement invariance when composite scores are employed to examine group differences.

**Keywords:** *Measurement Invariance, Factor Analysis, Group Differences, Factor scores*

## 1. INTRODUCTION

Educational and behavioral research involves examining the interrelationship of variables to inform models of behavior. Generally, these variables are indicators of broader latent abilities, sometimes referred to as factors. In some instances, these factors are given a numerical value following a specified scoring system, which results in observed composite or factor scores. These factor scores are derived based on observed indicators to relate the factors to other variables in a straightforward manner and construct a test (Gorsuch, 1983). Most common is the use of such scores in subsequent analysis (e.g., analysis of variance, regression; Allen, 1999; Berger & Milem, 1999) to answer questions related to group differences or to explain patterns of behaviors, for example. A search of the PsycINFO database using the term factor scores found over 700 articles in which factor scores were used in some fashion over the last ten years (1997-2007), including studies comparing factor score means. Therefore, it seems clear that this practice is commonly in use among psychological researchers.

Factor score is being used as a general term to this point. However, we recognize that it may have a specific meaning for certain researchers working in various educational or psychological settings. There are multiple ways (methods and formulas) to create factor scores (e.g., Gorsuch, 1983; Harman, 1976; Kaiser, 1962; Kestelman, 1952; Ten Berge, Krijnen, Wansbeek, & Shaprio, 1999; Thurstone, 1935). Under certain data conditions different methods may be more appropriate than others (e.g., Dobie, McFarland, & Long, 1986). Perhaps two of the more popular methods include the regression based approach and the sum score, or unit weighting approach. This popularity is most likely due to their relative ease of computation. For instance, a summed (or averaged) score can be efficiently created and understood by just about any person. Likewise, the regression based method is the standard technique used in major statistical software packages, such as SAS and SPSS, when the user requests that the factor scores be saved from an exploratory factor analysis. Latent factor means are another alternative, yet generally require more specialized software and technical knowledge. In addition, researchers using popular scales that were originally created as unit weighted or factor scores may not realistically have the option to use latent factor means if they want their work to be associated with prior studies using the instrument. Regardless of how the score is computed, it is meant to give the researcher or practitioner a concrete estimate of a subject's ability on the trait being measured. That said, the inferences that can be drawn from these scores are dependent, in part, on the score reliability and validity evidence that have been provided.

Factor scores created using the regression method can be expressed as (Gorsuch, 1983):

$$F_{nk} = Z_{nj} R_{jj}^{-1} S_{jk} \quad (1)$$

Where

$Z_{nj} = \text{Matrix of standardized observed variable scores}$

$R_{jj}$ = Correlation matrix for observed variables

$S_{jk}$ = Factor structure matrix

The resulting values are *z*-scores and can be used in subsequent analyses, such as to make group comparisons on the factors. However, comparison cannot be made within the data set, as all factors have identical means (0) and standard deviations (1). Thompson (1993) provided a solution for this problem through the use of noncentered standardized factor scores. Gorsuch suggested that the main criterion for selecting a factor score estimation method is the ease of computation, which the regression method certainly meets given that its calculations are built into common statistical software.

The unit weighted method for obtaining composite scores typically involves the summation of a set of all observed variables or items associated with a common measure in order to obtain a total score, or summation of subsets of items to create subscale scores. This is given as:

$$F_i = x_{i1} + x_{i2} + x_{i3} \ldots x_{in}$$

$$\text{Where } F_i \text{ is the score of for the } i_{th} \text{ person} \qquad (2)$$

$$X_{i1\ldots n} = \text{the item response for the } i_{th} \text{ person on the } n_{th} \text{ item.}$$

A variation of this approach involves using the mean of these variables, rather than the sum, as the total score, where $F_i$ is simply divided by $N_{items}$ or the total number of items to create an individual's factor score. In either case, the combined variables are each given a weight of 1. This approach is very commonly used by researchers, practitioners, and test developers in conjunction with items on a scale that represent index scores (e.g., IIS, Pascarella & Terenzini, 1980; WISC-III, Wechsler, 1991; WJ-R, Woodcock-Johnson, 1990). There have been suggestions that scores be standardized before summing or averaging variables to create the unit-weighted factor score (Wainer, 1976), as is the case for the regression method. And, as with the regression scores described above, once created these scores are often used in further analyses such as linear regression, ANOVA, and correlation to examine a variety of research questions.

An important issue associated with creating factor scores and using them in subsequent analyses is measurement/factor invariance (e.g., Little, 1997). Factor invariance refers to the situation where parameters (e.g., loadings) linking the observed indicators and the latent factors are not equivalent across groups. If factor invariance does not hold, interpretation of reported group differences can be called into question (Rock, Werts, & Flaugher, 1978), as they could be due to either measurement differences or real ability differences. Thus, the importance of factor invariance with respect to these factor scores cannot and should not be minimized. The potential problems when creating composite scores from non-invariant indicators of factors have been discussed (e.g., Cheung & Rensvold, 1999). Cheung and Rensvold argue that of the three approaches (i.e., summated scores, latent means, factor scores) for creating construct means across groups, the use of factor scores (based on the regression method) leaving out items not associated (i.e., zero factor loading) with any of the factors may be the most appropriate and accurate procedure to follow. That is, each factor would be regressed with items having a non-zero loading associated with it. The real data example they use illustrates how different conclusions regarding statistical significance of mean differences between groups can be reached given factor score calculation choice and lack of invariance. If such differences can be generalized beyond this single example, the implications for practice would be wide-spread. The thought of incorrect conclusions being drawn for many studies in many areas at the group level and for the individual is quite disconcerting in light of its potential influence on theory development or decisions regarding individuals based on composite scores. However, given that Cheung and Rensvold's conclusion was based on one real data set where the level of invariance was unknown, simulation work controlling this and other factors is warranted.

No clear direction regarding the optimal approach for accurately dealing with non-invariant indicators in creating composite scores exists. One suggestion is to leave such indicators in the composite when just a few exist (e.g., Marsh & Hocevar, 1985). This argument is based on the idea that a few items or variables, given that there are many, will not accumulate to a high enough level so as to influence group mean scores and hence mean differences. This may be due to differences in both directions (e.g., noninvariance favoring different groups for different items) so that either the non-invariant indicators essentially cancel each other out or the differences are so small as to not make a difference when combined with all other variables and thus are not problematic. On the other hand, the presence of many non-invariant indicators or a small number exhibiting large differences on factor loadings may result in the construct being measured differently across groups. Another approach suggested for dealing with noninvariance is to estimate factor scores under the condition of partial factorial invariance (PFI; Cheung & Rensvold, 1999). See also Byrne, Shavelson, and Muthén, 1989 for more information on PFI. Additional research focusing on the influence of PFI on the estimation of factor scores and resulting mean estimates is needed.

Particularly, researchers should compare results including and excluding the non-invariant indicators as well as explore the magnitude of such loadings on factor score methods (Cheung & Rensvold). In related research, including other types of invariance (i.e., differential item functioning) in scoring of tests also can create group mean differences in observed mean scores when differences do not exist on the latent variable (e.g., Li & Zumbo, 2008).

Of particular interest in regards to group non-invariance is its influence on the variances of observed composite scores, as these are most commonly use in practice in comparison to latent mean comparisons. Specifically, when factor loadings are not invariant across groups, the resulting impact may be most marked on the factor variances. To illustrate this point, consider the case where observed and latent variables are related such that absolute simple structure holds; i.e. each observed variable is related to only one factor. If the factor is standardized using the reference variable method (Thompson, 2004), then the variance of the factor can be expressed as:

$$\phi_{11} = \lambda_{x11}^2 \sigma_{11}^2$$

where

$$\phi_{11} = \text{Variance for factor 1} \qquad (3)$$

$$\lambda_{x11} = \text{Loading of reference variable } x \text{ on factor 1}$$

$$\sigma_{11}^2 = \text{Variance for reference variable } x$$

Thus, taking an example from Brown (2006, p. 133), the variance estimate for the latent variable for a group of participants with a reference variable variance of 32.49 and a standardized factor loading of 0.8848, would be $(0.8848)^2(32.49) = 25.44$. Now consider if a second group of participants were included in the study, with a variance on the reference variable identical to that of the first group, 32.49. However, factor loading invariance did not hold for this second group such that the standardized loading was 0.4, compared to 0.8848 in the first. The factor variance for this second group would be calculated as $(0.4)^2(32.49) = 5.20$. Therefore, despite the fact that the variance of the observed reference variable is the same for both groups, the lack of invariance on the loading leads to a difference in variances of the latent variable, so that the latent variable variance of the group with the smaller loading is lower. This issue has been demonstrated in the context of nonuniform differential item functioning (DIF), which is equivalent to differences in factor loadings across groups, where the standard deviations, and thus variances, have been shown to be reduced for one group in comparison to the another (French & Maller, 2006). Given these results, and the fact that composite scores are estimates of latent variables, it is of interest to ascertain the influence of factor loading non-invariance on the variance of these composites. This issue is particularly important when researchers plan to use the composites in subsequent analyses, as many standard methods for comparing group means (e.g., *t*-test, analysis of variance) require the assumption of homogeneity of variance across groups.

The primary purpose of this study was to examine the impact of non-invariant factor loadings on the variance of each of two types of factor scores: (1) regression method (Gorsuch, 1983) and (2) unit weighted method (i.e., simple sum of items). The influence of non-invariance is assessed in terms of the null hypothesis rejection rates for an *F*-test comparing factor score variances between two groups, as well as the average group variances across replications for each combination of conditions. While the simple example associated with equation 3 demonstrates the basic problem, it does not shed any light on the actual impact of this non-invariance that might be seen in practice. For instance, from this example we cannot ascertain to what extent the differences in variance might be associated with such things as sample size, group size ratios, or complexity of the factor structure. Nor is it possible to determine the impact of sampling variation on the magnitude of group variance differences. In addition, prior research in this area (Li & Zumbo, 2008) utilized a simulation study in order to examine the impact of a type of non-invariance (differential item functioning) on mean comparisons using the *t*-test. For these reasons a Monte Carlo study in conjunction with the analytic result described above will assist to clarify the impact of non-invariance on group variance equality. .

## 2. METHOD

To ascertain which manipulated variables influenced the comparison of group variances using different methods for obtaining composite scores, a Monte Carlo simulation study was conducted. Replications ($N$ = 1000) for each combination of manipulated conditions ensured stable results. The data were simulated using SAS with known model differences across groups under varying conditions of sample size, number of factors, proportion of non-invariant factors, magnitudes of loading differences and magnitude of mean differences.

*Number of Factors and Indicators*

Continuous and normally distributed subtest level data were simulated from both 2- and 3-factor models, with interfactor correlations set at 0.50 to represent moderately correlated factors and simple structure. The number of

indicators per factor was either 3 or 9 representing a minimum number and a sufficient number of such observed variables.

*Sample Size*

Adequate power and estimate stability in factor analysis can vary depending on the data conditions. Therefore, two group sample sizes were employed, 50 and 500, resulting in 2 sample size combinations: 50/50 and 500/500. Sample sizes are consistent with previous similar simulation research (e.g., Rensvold & Cheung, 2001; Lubke & Muthén, 2004; Meade & Lautenschlager, 2004).

*Magnitude and Percent of Difference Between Groups in Factor Loadings*

Five levels of factor loading differences were simulated. A baseline condition was established where no differences in loadings were present (true factorial invariance). The remaining 4 conditions included differences from 0.10 to 0.40 increasing in increments of 0.10. Currently, an effect size and associated guidelines for what represents a meaningful factor loading difference is not available (Millsap, 2005) and 0.25 has been used in previous simulation work related to factor analysis and invariance (e.g., French & Finch, 2006; Meade & Lautenschlager, 2004). The percent of loadings simulated to be different between the groups included 0% (invariant), 33%, and 100%. The magnitude of group 1 factor loadings also was varied (0.5 and 0.9) to represent moderate to strong relationships between the variables and the factors. That is, in one condition all loadings were 0.5 for group 1 and in another condition loadings were all 0.9 from group 1. The group 2 loadings differed from these by the amounts above.

*Data Generation and Analysis*

Observed data consistent with two and three factors were generated and exploratory factor analysis (EFA) was conducted with maximum likelihood extraction and PROMAX rotation to obtain regression based factor scores. In order to determine whether mean differences between groups on the observed variables would influence the variances of the composites, simulated observed variable means were varied between groups (0, 0.2, and 0.8) as expressed by Cohen's *d* (Cohen, 1988). For all simulation conditions two groups were used, with differences in loadings for corresponding variables occurring on only one factor. For example, in the 100% loading difference condition for the 3-factor model, all loadings differed on factor 1, and loadings on factor 2 and 3 were equal. Results from the EFA were employed to estimate factor scores for each simulee using the regression method (see Thompson, 2004), and the standard summated (i.e., unit weight) method was also used to create composite scores. These methods were selected because they are commonly seen in the literature, and each has support in prior research (see Grice and Harris, 1988 for a review of this issue).

*Evaluation Criteria*

Group variances for both types of factor scores were obtained and used to conduct a test of equality of group variances using an *F*-test. The outcome variables in the simulation study, across replications for both types of factor scores were the rejection rate for this *F*-test as well as the average variance by group. The groups' variances were simulated to be the same on the observed variables across all conditions, therefore any differences in the composite score variances would be a function of non-invariance in the factor loadings. Analysis of variance (ANOVA) was employed to examine the effects of the manipulated factors on the outcomes. Note in the results, F1 and U1 refers to the factor where differences were simulated and scores were based on the regression method and the unit-weighted method, respectively.

## 3. RESULTS

*2 Factors*

The results of the ANOVA revealed that the interaction of factor by sample size by degree of group loading difference by loading size by number of indicators was the highest order statistically significant term ($\omega^2$=0.307). Other significant interactions and main effects were subsumed in this higher order term, with the exception of the interaction between factor and proportion of loadings that were invariant ($\omega^2$=0.523). Table 1 includes rejection rates for the test of group variance equality by the sample size, base factor loading, difference between group factor loadings and number of observed indicators for both the regression and unit weighted methods. For samples of size 50($N_{small}$), the rejection rates for equality of the first regression-based composite score (F1) variances were consistently between 0.01 and 0.15. In addition, for this sample size there was no apparent relationship between group loading differences and the rejection rates. In contrast, for samples of 500($N_{large}$), the F1 rejection rates for the invariant condition were elevated in the 3 indicators per factor condition, but not for the 9 indicators per factor condition. Furthermore, this elevation in rejection rates rose concomitantly with differences in group factor loading values when the reference group loading was 0.5, but not when it was 0.9. In the case of 9

indicators, the rejection rates for $N_{large}$ were elevated above 0.1 for a reference group loading of 0.9 but not for 0.5. An examination of Table 2, which contains the differences in variances between groups in scores averaged across replications, shows that under most conditions the groups' average variances did not differ by more than 0.05, except for $N_{large}$, the reference group loading was 0.5 and the group loadings differed by 0.2 or more. This combination of conditions was associated with the highest rejection rates in Table 1.

The pattern of rejection rates for the second regression calculated composite score (F2) differed from that of F1 in several ways. First of all, these rejection rates were above 0.2 for 3 indicators with $N_{small}$ and the reference group loadings were 0.5. On the other hand, when the reference group loadings were 0.9 and $N_{small}$, the equal variances rejection rates for F2 in the 3 indicators condition were comparable to those for F1. When each factor had 9 indicators, the rejection rates for F2 with $N_{small}$ were uniformly below 0.05. For $N_{large}$, the rejection rates for the test of F2 variance equality across groups was elevated across number of indicators, reference group loadings, and differences between group loadings. This rejection rate increased with increasing differences in group loadings, particularly when the reference loading was 0.5 for 9 indicators. Finally, results in Table 2 reveal that, as with F1, the average variances across replications were very similar for the two groups with $N_{small}$, but not with $N_{large}$. Indeed, the F2 variance for group 1 declined with increasing differences in factor loadings, particularly when the reference loading was 0.9 with 3 indicators. When the reference loading was 0.5, the F2 variance for group 1 appears to have declined initially from the invariance condition to a difference of 0.1 and then largely stabilized, regardless of differences in group loadings.

With respect to the first unit weighted (or summed) composite score (U1), Table 1 reveals a steady increase in rejection rates for the test of equal group variances concomitant with increases in factor loadings differences across all other simulated conditions. In the 9 indicator condition, these increases in rejection rates were greater when the reference loading was 0.5, while for 3 indicators no such pattern was evident. As was true for F2, the U1 variances for group 1 generally declined as the difference in loadings increased, which apparently led to the increase in rejection rates for the test of equal variances.

While the group variances for U1 became more different as the population factor loadings became increasingly different, no such pattern was evident for the variances of the second factor, U2. A review of Table 1 reveals that the rejection rates were actually highest for the invariant condition when the reference loading was 0.5 and the factors had 3 indicators. Otherwise, they were between 0.08 and 0.11 for the remaining conditions. The average differences in U2 variances (Table 2) were near 0, indicating a similarity between groups.

Rejection rates by proportion of loadings simulated to be different appear in Table 3. This term was found to be statistically significant by the ANOVA. These results demonstrate the influence of the magnitude of non-invariance on differences between group variances. Specifically, for F2 and U1, the impact is very clear: As the proportion of non-invariant loadings increases, the rejection rate for the test of equal variances increases as well. In other words, more invariant loadings were associated with larger differences in group variances. On the other hand, both F1 and U2 appear to have been impervious to the proportion of non-invariant loadings. Finally, the results in Table 3 also provide a rough baseline regarding the proportion of significant tests for variance differences that would be expected in the ideal case; i.e. equal variance on the observed variables and invariance on the factor loadings, with values ranging between 0.11 and 0.15.

*3 Factors*

The results of the ANOVA for the 3 factor results indicated that the interaction of factor score by number of indicators by proportion of non-invariant loadings was statistically significant ($\omega^2$=0.96). In addition, the interactions of difference in factor loadings by factor score ($\omega^2$=0.885) and sample size by factor score ($\omega^2$ = 0.805) also were significant. Table 4 contains the rejection rates for equality of variance for the score by proportion of non-invariant loadings by number of indicators. Given these results, it appears that the baseline rates of rejection when all loadings were invariant, is similar to that for the 2-factor condition. Furthermore, as the proportion of non-invariant loadings increased, the rejection rates for F1 and U1 in the 3 indicator case, and U1 and F3 in the 9 indicator case increased steadily. On the other hand, for 3 indicators per factor, the rejection rates for the other composite scores did not change markedly until all of the loadings for the factor were different across groups. When factors were characterized by 9 indicators, rejection rates for the first two scores determined by the regression weight method (F1, F2) increased by approximately 0.06 from a proportion of 0% to 33%. The increase in rejection rates for these two factors was more notable from proportions of non-invariant loadings of 33% to 100%. On the other hand, the variance equality rejection rates for U2 and U3, the unit weighted composites, did not increase until the proportion of non-invariant loadings was 100%.

Table 5 contains the rejection rates for the composite score variances by the difference in group loadings and the sample size. Rejection rates for all of the composite scores increased concomitantly with differences in factor loadings. The greatest such change occurred for U1, though the increases in rejection rates for U2 and U3 were both

greater than for any of the regression weighted factor scores. In terms of sample size, higher rejection rates were evident for $N_{large}$ across all composite scores with the unit weighted score having the highest value.

The average difference in variances between groups by number of indicators and proportion of non-invariant loadings appear in Table 6. When the loadings were invariant, the regression weighted composite score variances for the two groups did not differ by more than 0.01 in the 3 indicators condition. However, as the proportion of non-invariant loadings increased, the difference in average variances did as well for the third composite with the proportion of non-invariant loadings equal to 1. Indeed, when one-third of the loadings were non-invariant, the average group variances did not differ by more than 0.02. When there were 9 indicators, the average variances for the regression weighted scores of group 2 were actually larger in the sample than were those of group 1, though the very low rate of significant differences would suggest that these differences were not present in the population as a whole. As the proportion of non-invariant loadings increased, the variances for the first group remained largely unchanged, while those of the second group declined by roughly 0.17 to 0.20 for each of the three scores. The average variances of the unit weighted composites for group 1 displayed a similar pattern to the regression based scores in that they did not change dramatically even as the proportion of non-invariant loadings was increased. This outcome was particularly evident when there were 9 indicators. For 3 indicators, all of the average variances decreased for both groups as non-invariance increased from a proportion of 0% to 0.33. The average variances for the second group were smaller still when the proportion was 100%, while they remained unchanged for group 1. In the 9-indicator condition, the variances for group 1 were unchanged across all proportion settings, while they declined for U1 from a proportion non-invariant of 0% to 33%, and for U2 and U3 for proportion non-invariant 33% to 100%.

Table 7 contains the average differences in variances between groups by difference in group loadings and sample size. When factor invariance held, the group variances for the regression weighted composite scores never differed by more than 0.04. However, as the degree of non-invariance increased, the variance in F3 for the focal group declined, while that of the reference group remained largely unchanged from a loading difference of 0.1 to 0.4. A somewhat similar pattern was evident for all of the unit weighted composite variables, with the greatest such decline in focal group variances occurring for the first composite. In addition, it is noted that for both the regression weighted and the unit weighted scores, the variance of all variables declined when any invariance was introduced. In terms of sample size, variances for the regression weighted scores were smaller with $N_{large}$, and the difference between the two groups' variances was never more than 0.1. However, the pattern of differences for the unit weighted scores was virtually the same regardless of the sample size, with the greatest group difference occurring for the U1. This demonstrates that the unit-weighted composites are most influenced by the lack of invariance compared to the regression weighted composites.

## 4. DISCUSSION

The findings presented in this study have practical implications for educational and psychological researchers and practitioners using composite scores created using items or subscales from an instrument to compare groups. Such scores are utilized frequently, sometimes in a post-hoc fashion based upon the results of an exploratory factor analysis. The resulting composites may take the form of true factor scores created using a technique such as the regression-based method described here, or they may simply be unit-weighted sums (or averages) of the constituent items or subscales. In either case, these scores are then often used in subsequent analyses or for making decisions about individuals in a clinical or educational context (e.g., individual educational plan). In the majority of such studies, researchers do not report investigating the factor invariance of these scales. Indeed in the case of sum scores, factor analysis may never actually be used in the construction or validation of the scales' scores, especially for locally developed or research-specific measures. The results presented here demonstrate that using either type of factor score without investigating invariance might lead to problems with subsequent statistical analyses and the possibility of inaccurate conclusions. Interestingly, these problems were perhaps most evident when unit-weighted sum scores were created with no direct application of factor analysis.

A quick thought experiment that considers many areas in educational and psychological research that employ measures scored in such a way accompanied by the lack of empirical evaluation of invariance leads a person to many conclusions. A warning is issued that continuing with this experiment may cause sleepless nights. Two immediate conclusions exist for us. The first and positive conclusion would be that little difference would exist if a lack of invariance existed. The second, negative, and more troubling conclusion, would be that it does matter if a lack of invariance exists. If the latter is the case, many conclusions in areas making score comparisons would be misleading and theory and interventions have been be developed based on such inaccurate conclusions. We certainly hope this is not the case!

In considering the hypothesis tests comparing the group variances that were used in this study, it is important to note that the nominal Type I error rate of 0.05 was frequently surpassed, even when the loadings were simulated to be

invariant and the group variances on the individual observed variables were essentially equal. As a result of this general inflation of the rejection rates we have attempted to avoid the use of the term Type I error rate, since the empirical rate is not known exactly. Nonetheless, it is possible to examine patterns of rejection rates from this study, and thereby gain an understanding of the relative changes in variance equality associated with the various study conditions. In addition, by presenting the average variances, we hope to elucidate when significant results for the *F*-test are practically important and when they are not.

Given the current results, it seems clear that a lack of factor invariance at the population level is associated with increased inequality of variances across the groups. As the degree and proportion of non-invariance increased, the rejection rate for a test of equality of group variances also increased under many of the conditions simulated in this study. This outcome was particularly marked for unit-weighted scores, though it manifested itself for both types. Indeed, these rejection rates for the non-invariant conditions were at least twice as high as those for the invariant data, and in some cases over five times the size of those in the invariant case. A similar pattern was found for the factor scores as well, though the degree of group variance difference in the non-invariant conditions was not as great as it was for the sum scores. This pattern of results was apparent in both the 2-factor and 3-factor conditions, and the rejection rates for the equality of variances were generally higher for larger sample sizes.

While the patterns described above were indeed apparent in the data, they did not occur with equal frequency across the two (or three) factor scores. For example, in the 2-factor condition, only the first unit-weighted score exhibited an increase in variance equality test rejection rates with increasing levels of loading non-invariance. Given that the non-invariance occurred on the first factor, this result would be expected. However, when 3 factors were present, the increase in variance inequality was present for all of the unit-weighted scores, though most notably in the first. In contrast, the findings for the regression weighted composite scores were much more complex and difficult to summarize succinctly. A combination of high reference group factor loadings (0.9) and a larger number of indicators (9) seemed to mitigate variance inequality quite substantially in the 2- factor condition. A similar pattern with respect to group 1 loading values was evident in the 3-factor case, though the interaction with number of indicators was not statistically significant.

In terms of sample size, the larger sample was associated with a higher rejection rate for detecting variance differences for the scores examined here. However, the reasoning for this difference due to sample size appears different for regression weighted and unit weighted sum scores. For example, referring to Table 1, with a sample of 500 with 9 indicator variables, a reference loading of 0.5 and a group loading difference of 0.1, the null hypothesis of no variance difference for U1 was rejected at a rate of 0.30, while for N of 50 the rejection rate was 0.96. An examination of the average differences in variances between groups in Table 2 reveals that for this combination of conditions, the actual group variances are nearly identical across sample sizes. Thus, the difference in rejection rate can be attributed to a difference in sample size. On the other hand, for the same combination of conditions, the rejection rate for F2 when N=500 was 0.04 and for N=50 it was 0.22. An examination of the actual group variances, however, reveals a very different picture than was evident for U1. Specifically, the average F2 variance estimates for the two groups were identical in the smaller sample size condition, while they differed by 0.17 for the larger sample. These differing patterns for the regression and unit-weighted variables appeared consistently, for two factors. Thus, while for samples of 50 and 500 differences between group variances of the unit weighted variables were essentially the same, for the regression calculated scores smaller samples were associated with little or no decline in group 1 variances, unlike the result for the larger sample size.

This finding of inflated rejection rates for equality of group variances in the non-invariant condition is a definite cause for concern for researchers and data analysts, because many standard parametric statistical analyses such as ANOVA, regression, and MANOVA, to name but a few, operate under the assumption of homogeneity of variance. It would appear, however, that when invariance of factor loadings in the population does not hold, this assumption with regards to commonly used factor scores may be invalid. This violation would be detected with standard assumption checking of the said analyses if such checking is actually carried out. But again, rejection of the assumption would not be accurate. This fact, taken in conjunction with the general lack of evidence that researchers actually consider the invariance (or lack thereof) for their instruments, should be a cause for concern. Refer to the thought experiment above but proceed with caution.

*Implications for practice*

The results of this study have several interesting implications for researchers to consider when they employ factor scores. Perhaps most importantly, a lack of factor invariance could potentially have a dramatic impact on the homogeneity of group variances on these scores. Whether the researcher uses factor scores in the form of values obtained by first conducting an exploratory factor analysis, or through the simpler unit-weighted score method, if the population factor loadings are not invariant, it appears that variance inequality is very likely in many of the cases studied here. Thus, researchers should first establish factor invariance prior to creating and using these scores.

Failure to do so may well result in compromised statistical analyses and in turn lead to incorrect interpretations of what is happening in the population. In addition to formally testing for factor invariance with their sample, data analysts can also investigate prior research with the measure in question to ascertain whether prior invariance testing work has been conducted on similar samples they will be working with in terms of demographic or other group categorization variables. Moreover, a simple examination of exploratory factor analysis results for the groups in question can provide very useful information regarding the comparability of factor loadings (Finch & French, 2007). If a lack of invariance is found using one of these methods, researchers must decide on a course of action in creating and using the scores The results of this study suggest that if the degree of invariance is small, the impact on factor variances may not be very large. Nonetheless, it is clear that some group variance differences should be expected. One suggestion that has been offered previously (Cheung & Rensvold, 1999) is to conduct the desired analyses (e.g., *t*-test) both with and without the non-invariant variable/loading included when creating the factor scores. If there is a difference in results with respect to the comparison of group variances, the researcher could then conclude that the lack of invariance most likely will influence the variances. They would then need to decide whether to continue with the analysis excluding the non-invariant variables from the calculation of factor scores, or to refrain due to the lack of invariance. An alternative approach would be to calculate the factor scores allowing the groups' loadings to differ for the non-invariant variables. Li and Zumbo (2008) found that Type I error rates for mean comparisons were not severely inflated when as much as 10% of the indicators were found to lack invariance (i.e., DIF items). This result might provide some guidance for the applied researcher who elects to continue with analyses while including a few non-invariant loadings. However, given the results reported here which showed an impact on group variances when a third of the indicators were non-invariant, we suggest that this approach be undertaken with extreme caution because the resulting factor scores would not be strictly comparable, meaning that any group differences that were discovered may be due to the divergence in measurement rather than differences on the construct.

The rejection rate differences described in the results were not homogeneous across score calculation methods, or data conditions such as sample size, number of indicators, and so on. One major implication of these complex outcomes is that the practitioner may not be able to pinpoint precisely how a lack of invariance will influence the analyses, though it seems clear that some influence will occur. It does appear that for smaller samples with more factor indicators and fairly large loadings (0.9 in this case), the influence of non-invariance is somewhat less severe compared to many of the other conditions examined in this study.

The results with regard to a comparison of unit and regression weighted scores are complex and do not appear to definitely support one approach over the other. When the loadings are invariant, both methods produced results in which average group variances were very similar, notwithstanding the somewhat inflated rejection rates for the *F*-test in this case. On the other hand, when group loadings differed, the rejection rates for the regression weighted approach were lower in some instances, but higher in others. Furthermore, whereas the variances for both regression weighted factors (F1 and F2) appear to have been affected by non-invariance, at least in some instances, only the variances for the first weighted score (U1) was influenced when there were two factors. The three factor results follow this pattern weakly, but there the results as to which approach might be optimal appear to be even less clear.

The complexity of this study's results appears to be in accord with the general literature on the creation of factor scores. As Grice and Harris (1998) note in their review of the issue of creating factor scores, there is evidence both for and against the use of unit weighting as a method for creating composite scores. Prior research on this issue has not directly examined the performance of these methods in the context of comparing variances on the composite scores, so that no direct comparison of the current study's results can be made with earlier work. We look forward to future work in this area so such comparisons can be made, and more importantly, stronger conclusions can be stated about the subject matter. Such work should help guide researchers and practitioners in their work and assist with making more accurate group comparisons.

The results of this study further reinforce the need for researchers to examine the factor invariance of their measures, and to make adjustments to their analyses using factor scores when invariance does not hold. While the issue of factor invariance does not appear to be widely considered when researchers use unit-weighted sum (or average) scores or factor scores obtained from exploratory factor analysis, it is clear that non-invariance can have an impact on the subsequent scores, particularly in terms of group variances. In turn, inequality of group variances can affect the tests used to compare means on these composite scores. Future research in this area should focus on the impact of non-invariance on other relationships among factor scores, in addition to comparisons of means, such as correlation and regression. Such results would be very important because many researchers use composite scores by the methods studied here, without examining the issue of factor invariance, and thus ignoring potential problems in analysis.

**REFERENCES**


[1]. Allen, D. (1999). Desire to finish college: An empirical link between motivation and persistence. *Research in Higher Education, 40*, 461-485.

[2]. Berger, J. B., & Milem, J. F. (1999). The role of student involvement and perceptions of integration in a causal model of student persistence, *Research in Higher Education*, *40*, 641-664.

[3]. Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 3*, 456-466.

[4]. Brown, T.A. (2006). *Confirmatory Factor Analysis for Applied Research.* New York: The Guilford Press.

[5]. Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*, 1-27.

[6]. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

[7]. Dobie, T., McFarland, K. & Long, N. (1986). Raw score and factor score multiple regression: An evaluative comparison. *Educational and Psychological Measurement,46*, 337-347.

[8]. French, B. F. & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling, 13, 378-402.*

[9]. Finch, W.H. & French, B.F. (2007). Using exploratory factor analysis for locating invariant referents in factor invariance studies. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

[10]. French, B. F. & Maller, S. J. (2006, April). *The influence of differential item functioning on internal consistency reliability.* Paper presented at the American Educational Research Association, San Francisco, CA.

[11]. Gorsuch, R. L. (1983). *Factor analysis* 2nd edition. Hillsdale, NJ: Lawrence Erlbaum

[12]. Grice, J. W., & Harris, R. J. (1998).A comparison of regression and loading weights for the computation of factor scores. *Multivariate Behavioral Research, 33* ,221-247.

[13]. Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.

[14]. Kaiser, H. F. (1962). Formulas for component scores. *Psychometrika,* 27, 83-87.

[15]. Kestelman, H. (1952). The fundamental equations of factor analysis. *British Journal of Psychology, 5*, 1-6.

[16]. Li, Z. & Zumbo, B. D. (2008). *Impact of differential item functioning on statistical conclusions.* Paper presented at the National Council on Measurement in Education conference, New York, NY.

[17]. Little, T. D. (1997). Mean and covariance structures (MACS) analysis of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53–76.

[18]. Lubke, G.H. & Muthén, B.O. (2004). Applying multi-group confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11*, 514-534.

[19]. Marsh, H.W. & Hocevar, D. (1985). Confirmatory Factor Analyses of Multitrait-Multimethod Data: Many Problems and a Few Solutions. Psychological Bulletin, 97, 562-582.

[20]. Mead, A. W., & Lautenschlager, G. J. (2004). A monte-carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling, 11,* 60-72.

[21]. Millsap, R.E. (2005). Four unresolved problems in studies of factorial invariance. In A. Maydeu-Olivares & J.J. McArdle (Eds.) *Contemporary Psychometrics* (pp. 153-172). Mahwah, NJ: Lawrence Erlbaum Associates.

[22]. Pascarella, E. T., & Terenzini, P. T. (1980). Predicting persistence and voluntary dropout decisions from a theoretical model. *Journal of Higher Education, 51,* 60-75.

[23]. Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim & L. L. Neider (Eds.), *Research in management: Equivalence in measurement* (pp. 25-50). Greenwich, CT: Information Age Publishing.

[24]. Rock, D. A., Werts, C. E., & Flaugher, R. L., (1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research, 13*, 403-418.

[25]. Ten Berge, J., Krijnen, W., Wansbeek, T., & Shapiro, A. (1999). Some new results on correlation-preserving factor scores prediction methods. *Linear Algebra and its Applications, 289*, 311-318.

[26]. Thompson, B. (2004). *Exploratory and Confirmatory Factor Analysis.* Washington, DC: American Psychological Association.

[27]. Thompson, B. (1993). Calculation of standardized, noncentered factor scores: An alternative to conventional factor scores. *Perceptual and Motor Skills, 77*, 1128-1130.

[28]. Thurstone, L. L. (1935). *The vectors of mind.* Chicago: University of Chicago Press.

[29]. Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind.

*Psychological Bulletin, 83,* 213-217.
[30]. Wechsler, D. (1991). *Wechsler Intelligence Scale for Children – Third Edition.* San Antonio, TX: Psychological Corporation.
[31]. Woodcock, R. W., & Johnson, M. B. (1989/1990). *Woodcock- Johnson Psycho-Educational Battery-Revised.* Itasca, IL: Riverside Publishing.

Table 1
*Rejection Rates for Group Differences on Factor Score Variances by Sample size, Difference in Group Loading Values, Loading Value of Reference Group, and Number of Indicators, Invariance/non-invariance: 2-Factor Condition*

| 3 indicators | | | | | | |
|---|---|---|---|---|---|---|
| N | Difference | Loading | F[a]1 | U[b]1 | F2 | U2 |
| 50 | 0 | 0.5 | 0.14 | 0.10 | 0.32 | 0.19 |
| | | 0.9 | 0.06 | 0.10 | 0.08 | 0.11 |
| | 0.1 | 0.5 | 0.15 | 0.10 | 0.24 | 0.10 |
| | | 0.9 | 0.05 | 0.10 | 0.07 | 0.09 |
| | 0.2 | 0.5 | 0.15 | 0.13 | 0.27 | 0.08 |
| | | 0.9 | 0.05 | 0.15 | 0.08 | 0.08 |
| | 0.3 | 0.5 | 0.14 | 0.20 | 0.27 | 0.08 |
| | | 0.9 | 0.05 | 0.26 | 0.10 | 0.08 |
| | 0.4 | 0.5 | 0.15 | 0.25 | 0.27 | 0.08 |
| | | 0.9 | 0.05 | 0.40 | 0.13 | 0.08 |
| 500 | 0 | 0.5 | 0.22 | 0.13 | 0.26 | 0.14 |
| | | 0.9 | 0.24 | 0.10 | 0.24 | 0.09 |
| | 0.1 | 0.5 | 0.29 | 0.25 | 0.43 | 0.10 |
| | | 0.9 | 0.22 | 0.28 | 0.28 | 0.09 |
| | 0.2 | 0.5 | 0.41 | 0.55 | 0.50 | 0.09 |
| | | 0.9 | 0.19 | 0.60 | 0.46 | 0.09 |
| | 0.3 | 0.5 | 0.53 | 0.73 | 0.49 | 0.09 |
| | | 0.9 | 0.18 | 0.75 | 0.57 | 0.09 |
| | 0.4 | 0.5 | 0.64 | 0.86 | 0.48 | 0.09 |
| | | 0.9 | 0.14 | 0.86 | 0.57 | 0.09 |
| 9 indicators | | | | | | |
| N | Difference | Loadings | F[a]1 | U[b]1 | F2 | U2 |
| 50 | 0 | 0.5 | 0.01 | 0.09 | 0.01 | 0.09 |
| | | 0.9 | 0.03 | 0.11 | 0.04 | 0.10 |
| | 0.1 | 0.5 | 0.01 | 0.30 | 0.04 | 0.09 |
| | | 0.9 | 0.04 | 0.14 | 0.04 | 0.09 |
| | 0.2 | 0.5 | 0.01 | 0.49 | 0.04 | 0.10 |
| | | 0.9 | 0.04 | 0.15 | 0.04 | 0.10 |
| | 0.3 | 0.5 | 0.01 | 0.55 | 0.03 | 0.09 |
| | | 0.9 | 0.04 | 0.17 | 0.04 | 0.09 |
| | 0.4 | 0.5 | 0.01 | 0.67 | 0.04 | 0.09 |
| | | 0.9 | 0.03 | 0.21 | 0.04 | 0.09 |
| 500 | 0 | 0.5 | 0.02 | 0.10 | 0.02 | 0.09 |
| | | 0.9 | 0.21 | 0.10 | 0.21 | 0.10 |
| | 0.1 | 0.5 | 0.01 | 0.96 | 0.22 | 0.09 |
| | | 0.9 | 0.19 | 0.45 | 0.20 | 0.10 |
| | 0.2 | 0.5 | 0.01 | 1.00 | 0.37 | 0.11 |
| | | 0.9 | 0.16 | 0.65 | 0.22 | 0.10 |
| | 0.3 | 0.5 | 0.02 | 0.97 | 0.55 | 0.10 |
| | | 0.9 | 0.14 | 0.80 | 0.25 | 0.10 |
| | 0.4 | 0.5 | 0.02 | 1.00 | 0.57 | 0.10 |
| | | 0.9 | 0.11 | 0.90 | 0.31 | 0.10 |

[a.] Factor score based on regression method.
[b.] Factor score based on unit-weighted method.

Table 2

*Average Between Group Factor Score Variance Differences by Sample Size, Difference in Group Loading Values, Loading Value of Reference Group, and Number of Indicators, Invariance/non-invariance: 2-Factor Condition*

| 3 indicators | | | | | | |
|---|---|---|---|---|---|---|
| N | Difference | Loading | F[a]1 | U[b]1 | F2 | U2 |
| 50 | 0 | 0.5 | 0.00 | -0.01 | 0.01 | 0.00 |
| | | 0.9 | 0.00 | -0.02 | 0.00 | 0.00 |
| | 0.1 | 0.5 | 0.01 | -0.34 | -0.01 | 0.00 |
| | | 0.9 | 0.01 | -0.61 | -0.04 | 0.02 |
| | 0.2 | 0.5 | 0.01 | -0.62 | 0.00 | -0.01 |
| | | 0.9 | 0.01 | -1.19 | -0.04 | -0.03 |
| | 0.3 | 0.5 | 0.03 | -0.89 | -0.01 | 0.03 |
| | | 0.9 | 0.01 | -1.72 | -0.03 | 0.01 |
| | 0.4 | 0.5 | 0.01 | -1.01 | 0.01 | 0.01 |
| | | 0.9 | 0.00 | -2.18 | -0.03 | -0.03 |
| 500 | 0 | 0.5 | 0.00 | -0.01 | -0.01 | 0.01 |
| | | 0.9 | 0.00 | 0.00 | -0.01 | -0.01 |
| | 0.1 | 0.5 | 0.01 | -0.33 | -0.05 | -0.01 |
| | | 0.9 | 0.03 | -0.63 | -0.06 | 0.01 |
| | 0.2 | 0.5 | 0.07 | -0.63 | -0.06 | 0.00 |
| | | 0.9 | 0.03 | -1.18 | -0.10 | -0.02 |
| | 0.3 | 0.5 | 0.13 | -0.85 | -0.05 | -0.01 |
| | | 0.9 | 0.02 | -1.70 | -0.14 | 0.01 |
| | 0.4 | 0.5 | 0.19 | -0.98 | -0.03 | -0.01 |
| | | 0.9 | 0.02 | -4.44 | 0.00 | 0.00 |
| 9 indicators | | | | | | |
| N | Difference | Loadings | F[a]1 | U[b]1 | F2 | U2 |
| 50 | 0 | 0.5 | 0.00 | 0.10 | 0.00 | 0.33 |
| | | 0.9 | 0.00 | 0.43 | 0.00 | 0.20 |
| | 0.1 | 0.5 | 0.00 | -10.94 | 0.00 | -0.06 |
| | | 0.9 | 0.01 | -7.37 | -0.01 | 0.29 |
| | 0.2 | 0.5 | -0.01 | -13.31 | 0.01 | -0.02 |
| | | 0.9 | 0.01 | -9.16 | -0.01 | 0.29 |
| | 0.3 | 0.5 | 0.00 | -9.89 | 0.01 | -0.18 |
| | | 0.9 | 0.00 | -11.43 | -0.01 | 0.25 |
| | 0.4 | 0.5 | 0.00 | -11.98 | 0.01 | 0.00 |
| | | 0.9 | 0.00 | -13.05 | -0.01 | 0.01 |
| 500 | 0 | 0.5 | 0.00 | -0.03 | 0.00 | 0.01 |
| | | 0.9 | 0.00 | -0.03 | 0.01 | 0.15 |
| | 0.1 | 0.5 | -0.01 | -11.03 | -0.17 | -0.06 |
| | | 0.9 | 0.03 | -7.25 | -0.04 | 0.07 |
| | 0.2 | 0.5 | -0.01 | -13.44 | -0.23 | -0.05 |
| | | 0.9 | 0.02 | -11.54 | -0.04 | -0.05 |
| | 0.3 | 0.5 | -0.01 | -10.04 | -0.16 | -0.01 |
| | | 0.9 | 0.01 | -15.23 | -0.05 | -0.06 |
| | 0.4 | 0.5 | -0.01 | -12.17 | -0.15 | -0.06 |
| | | 0.9 | 0.01 | -18.67 | -0.07 | 0.12 |

Note: A negative value indicates group 2 variance was larger compared to group 1 and a positive value indicates the opposite.

[a.] Factor score based on regression method.
[b.] Factor score based on unit-weighted method.

Table 3
*Rejection Rates for Group Differences on Factor Score Variances by Proportion of Non-Invariant Loadings: 2-Factor Condition*

| Proportion | F[a]1 | U[b]1 | F2 | U2 |
|---|---|---|---|---|
| 0 | 0.12 | 0.12 | 0.15 | 0.11 |
| 0.3 | 0.13 | 0.42 | 0.20 | 0.09 |
| 1.0 | 0.13 | 0.61 | 0.38 | 0.09 |

[a.] Factor score based on regression method.
[b.] Factor score based on unit-weighted method.

Table 4
*Rejection Rates for Group Differences on Factor Score Variances Number of Indicators and Proportion of Non-invariant Loadings: 3-Factor Condition*

| Indicators | Proportion | F[a]1 | U[b]1 | F2 | U2 | F3 | U3 |
|---|---|---|---|---|---|---|---|
| 3 | 0 | 0.14 | 0.10 | 0.17 | 0.10 | 0.18 | 0.09 |
| | 0.3 | 0.21 | 0.27 | 0.17 | 0.11 | 0.19 | 0.11 |
| | 1.0 | 0.41 | 0.55 | 0.43 | 0.55 | 0.48 | 0.55 |
| 9 | 0 | 0.01 | 0.09 | 0.01 | 0.09 | 0.02 | 0.09 |
| | 0.3 | 0.06 | 0.16 | 0.07 | 0.10 | 0.31 | 0.10 |
| | 1.0 | 0.32 | 0.79 | 0.34 | 0.79 | 0.35 | 0.79 |

[a.] Factor score based on regression method.
[b.] Factor score based on unit-weighted method.

Table 5
*Rejection rates for Group Differences on Factor Score Variances by the Magnitude of Difference in Groups' Loadings, and Sample Size: 3-Factor Condition*

| Difference | F[a]1 | U[b]1 | F2 | U2 | F3 | U3 |
|---|---|---|---|---|---|---|
| 0 | 0.13 | 0.13 | 0.14 | 0.11 | 0.16 | 0.10 |
| 0.1 | 0.17 | 0.35 | 0.17 | 0.25 | 0.20 | 0.25 |
| 0.2 | 0.23 | 0.57 | 0.24 | 0.37 | 0.32 | 0.37 |
| 0.3 | 0.28 | 0.70 | 0.28 | 0.44 | 0.38 | 0.45 |
| 0.4 | 0.32 | 0.79 | 0.32 | 0.48 | 0.44 | 0.48 |
| N | | | | | | |
| 50 | 0.05 | 0.41 | 0.07 | 0.28 | 0.10 | 0.28 |
| 500 | 0.45 | 0.73 | 0.43 | 0.48 | 0.54 | 0.47 |

[a.] Factor score based on regression method.
[b.] Factor score based on unit-weighted method.

Table 6

*Average Between Group Variance Differences on Factor Score Variances Number of Indicators and Proportion of Non-invariant Loadings: 3-Factor Condition*

| Indicators | Proportion | F[a]1 | U[b]1 | F2 | U2 | F3 | U3 |
|---|---|---|---|---|---|---|---|
| 3 | 0 | 0.0 | 0.0 | 0.0 | 0.01 | 0.01 | 0.02 |
| | 0.3 | -0.02 | 0.62 | 0.0 | 0.03 | -.0.01 | -0.03 |
| | 1.0 | -0.03 | 1.44 | 0.06 | 1.5 | 0.09 | 1.43 |
| 9 | 0 | -0.09 | 1.14 | -0.08 | 1.49 | -0.06 | 0.55 |
| | 0.3 | 0.01 | 18.37 | 0.01 | -0.04 | 0.09 | 0.03 |
| | 1.0 | 0.09 | 16.36 | 0.10 | 16.86 | 0.09 | 15.92 |

Note: A negative value indicates group 2 variance was larger compared to group 1 and a positive value indicates the opposite.

[a.] Factor score based on regression method.

[b.] Factor score based on unit-weighted method.

Table 7

*Average Between Group Variance Differences on Factor Scores by the Magnitude of Difference in Groups' Loadings, and Sample Size: 3-Factor Condition*

| Difference | F[a]1 | U[b]1 | F2 | U2 | F3 | U3 |
|---|---|---|---|---|---|---|
| 0 | -0.04 | 0.56 | -0.04 | 0.79 | -0.03 | 0.28 |
| 0.1 | 0.0 | 14.52 | 0.01 | 12.32 | 0.02 | 12.24 |
| 0.2 | 0.02 | 18.47 | 0.04 | 14.37 | 0.06 | 14.47 |
| 0.3 | 0.02 | 21.66 | 0.05 | 16.51 | 0.09 | 15.84 |
| 0.4 | 0.01 | 22.03 | 0.05 | 15.27 | 0.11 | 14.94 |
| N | | | | | | |
| 50 | 0.01 | 17.57 | 0.01 | 13.34 | 0.02 | 13.19 |
| 500 | 0.01 | 16.63 | 0.05 | 12.53 | 0.09 | 12.28 |

Note: A negative value indicates group 2 variance was larger compared to group 1 and a positive value indicates the opposite.

[a.] Factor score based on regression method.

[b.] Factor score based on unit-weighted method.

Table 8

*Rejection rates for Group Differences Variances on Factor Score Variances by the Reference Group Loading: 3-Factor Condition*

| Loading | F[a]1 | U[b]1 | F2 | U2 | F3 | U3 |
|---|---|---|---|---|---|---|
| 0.5 | 0.30 | 0.61 | 0.29 | 0.41 | 0.42 | 0.41 |
| 0.9 | 0.19 | 0.53 | 0.20 | 0.35 | 0.23 | 0.35 |