Confirmatory factor analysis of the Maslach Burnout Inventory among Florida nurses

Jason W. Beckstead*

College of Nursing, University of South Florida, 12901 Bruce B. Downs Boulevard, Tampa, FL 33612-4766, USA

Received 1 November 2001; received in revised form 1 February 2002; accepted 22 February 2002

Abstract

Burnout among human service professionals, such as nurses, has been studied in various countries for years using the Maslach Burnout Inventory (MBI). This paper reports on confirmatory factor analyses using LISREL that examined the factorial validity of the MBI. The sample consisted of 151 registered nurses from west-central Florida. Modifications of the initial hypothesized three-factor structure were necessary to adequately fit the data. Findings are compared to the published normative values for the MBI and to similar studies of European nurses. Recommendations for measurement models of the MBI in future studies that use structural equation modeling techniques are offered.

Keywords: Maslach Burnout Inventory; Nurses; LISREL

1. Introduction

1.1. Burnout: its definition and measurement

Burnout is a condition of emotional exhaustion (EE), depersonalization (DP), and a reduced sense of personal accomplishment (PA) that can occur among individuals who work with people in some capacity. The term “burnout” was first used to describe a syndrome of exhaustion observed among mental health professionals (Freudenberger, 1974). Maslach (1982) provided a comprehensive definition of the term, incorporating the physical as well as the mental exhaustion observed in professionals whose work requires continuous contact with other people. According to Maslach, EE, having no capacity left to offer psychological support to others, emerges first. This is followed by the professional’s attempt to defend him or herself psychologically through isolation from the source of such affect. In this way the professional develops very impersonal relationships with his or her patients in an attempt to avoid stress and emotional fatigue.

Many theories of burnout attempt to explain its development through an interplay of job-related (environmental) and personality factors (e.g., Golembiewski and Munzenrider, 1988; Leiter and Maslach, 1988; O’Brien and Page, 1994). Burnout has been shown empirically to be related to certain characteristics of the job environment. For example, the stressful conditions prevalent in the health care setting, including exposure to death and dying, interpersonal conflict, and noise pollution have been found to increase burnout among nurses (Schmitz et al., 2000; Topf and Dillon, 1988). Personality factors including psychological hardness, locus of control and empathy have been shown to moderate such environmental influences in nurses (Astrom et al., 1991; Schmitz et al., 2000; Stechmiller and Yarandi, 1993).

Various instruments exist for measuring job-related burnout in human service professionals. Among these are Jones’ (1980) Staff Burnout Scale, Pines’ (1981) Burnout Measure, and the Maslach Burnout Inventory (MBI) (Maslach and Jackson, 1981). By far the most widely used is the MBI, due, in part, to the wealth of
research in support for its reliability and validity (see Maslach et al., 1996).

The MBI, containing 22 items, assesses burnout on three dimensions: EE, DP, and lack of PA. Maslach et al. (1996) reported internal consistency estimates of reliability: 0.90, 0.79, and 0.71, for the EE, DP, and PA subscales, respectively. Normative values for various human service professionals have also been established by the authors. The group norms published for nurses and physicians are 22.19, 7.12, and 36.53, for the three subscales, respectively (Maslach et al., 1996).

1.2. The purpose of the study

Nurses compose the largest segment of employees in the US healthcare industry, an estimated 2 million jobs in 1996 (Bureau of Labor Statistics, 1998). This number increased to 2.7 million as of March 2000 (US Department of Health and Human Services, 2001). Despite this increase, the US Department of Health and Human Services (1990) projected a shortage of 600,000 nurses in the industry by the year 2005. By 2020 the supply of fulltime RNs is projected to have fallen 20% below projected requirements (Buerhaus et al., 2000). Consistent with findings in other professions, nurses are susceptible to burnout (Topf and Dillon, 1988; Lewis and Robinson, 1992). Oehler et al. (1991) showed that greater work-related stress is often linked to increased EE in hospital nurses. Nurses working in community settings also experience burnout (Chung and Corbett, 1998). Understanding the process and consequences of burnout extends our understanding of people. Understanding the measurement of burnout, specifically in nurses, has the potential for making a significant contribution to research, policy, and interventions aimed at increasing, both the number, and the professional longevity of nurses working in direct health care jobs. The purposes of this study were: (a) to examine the factorial validity of the MBI by testing its hypothesized three-factor structure in a sample of Florida registered nurses; and (b) in the event that the hypothesized structure does not fit the data, to propose and test alternative factorial structures.

1.3. Confirmatory factor analysis

This study employed confirmatory factor analysis (CFA) to examine the factor structure of the MBI in a sample of nurses. In contrast to exploratory factor analysis, the aim of which is simply to identify the factor structure present in a set of variables, the aim of CFA is to test an hypothesized factor structure or model and to assess its fit to the data. CFA may be viewed as a submodel of the more general structural equation modeling (SEM) approach to analysis. Specifically, it is a measurement model of the relations of indicators (observed variables) to factors (latent variables) as well as the correlations among the latter. CFA is generally based on a strong theoretical and/or empirical foundation that allows the analyst to specify an exact factor structure in advance. The CFA approach usually restricts which variables will load on which factors, as well as which factors will be correlated. In CFA each observed variable has an error term, or residual, associated with it that expresses the proportion of variance in the variable that is not explained by the factors. These error terms also contain measurement error due to the lack of reliability in the observed variables. The typical research question in CFA is: Are the covariances (or correlations) among variables consistent with an hypothesized factor structure? As such, CFA is quite useful for studying the factorial validity of multi-item, multi-subscale instruments such as the MBI.

2. Method

2.1. Participants

A convenience sample of 151 registered nurses attending a major university in west-central Florida between 1999 and 2001 participated in this study. The data were collected during research methods classes attended by nurses pursuing advanced practitioner degrees. Most of the participants were female (85.4%). The majority were Caucasian (83.7%), 6.8% were Asian, 5.4% were African American, and 4.1% were Hispanic. Most worked in hospitals (86%), 7.3% worked in private practices, and 6.4% worked in home health settings. The average age of the participants was 36.5 years (range 22–57). The average number of years experience working as a nurse was 10.3 years (range 1–30).

2.2. Measurement

Professional burnout was measured using the MBI (Maslach et al., 1996). The MBI uses multiple items to measure burnout on three dimensions: EE, DP, and PA. The nine items of the EE subscale describe feelings of being emotionally overextended and exhausted by one’s work. The DP subscale contains five items assessing an unfeeling and impersonal response towards the recipients of one’s care (i.e., patients). The eight items of the PA subscale describe feelings of competence and successful achievement in one’s work with people. Participants rated the frequency of experiencing feelings related to each subscale using a 7-point scale with the verbal anchors: Never, A few times a year or less, Once a month or less, A few times a month, Once a week, A few times a week, and Every day, centered under the
The factorial validity of the MBI was assessed using SEM. The models below were tested with LISREL 8.30 (Jöreskog and Sörbom, 1996). Evaluation of each model was based on considering a variety of fit measures, and model comparisons are based on incremental differences in fit. These measures are now briefly discussed. The $\chi^2$ minimum fit function test is an inferential test of the plausibility of a model explaining the data. The root mean square error of approximation (RMSEA) expresses the lack of fit due to reliability and also model (mis)specification (Browne and Cudeck, 1993). RMSEA expresses fit per degree of freedom of the model and should be $<0.08$ for acceptable fit, with $0.05$ or lower indicating a very good fitting model. The goodness of fit index (GFI) and adjusted goodness of fit index (AGFI), which adjusts for the number of parameters estimated, range from $0$ to $1$ with values of $0.9$ or greater indicating a good fitting model (Jöreskog and Sörbom, 1996). These two indices are analogous to $R^2$ in multiple regression. The comparative fit index (CFI) assesses fit relative to a null model using noncentrality parameters (Bentler, 1988). The CFI also ranges from $0$ to $1$ with values of $0.9$ or greater indicative of good fitting models. The standardized root mean square residual (RMR) is the average of the differences between the sample correlations and the estimated population correlations. The RMR has a range from $0$ to $1$; values of $0.08$ or less are desired (Hu and Bentler, 1999).

### 3. Results

The means, standard deviations, and confidence intervals of the three subscales are shown in Table 1. Using one-sample $t$-tests to compare the current sample to the published norms discussed above, this sample of nurses scored significantly higher on PA subscale ($t = 4.906$, $df = 150$, $p < 0.001$). The sample did not differ from the normative values on either the EE or DP subscales. The reliability estimates (i.e., Cronbach’s alphas) reported in Table 1 were compared to the published norms using the test of significance developed by Alsawalmeh and Feldt (1999) for testing the equality of independent alphas. The Cronbach’s alpha coefficients calculated on this sample of nurses were not significantly different from those published by Maslach et al. (1996).

CFA was performed on the covariance matrix of the MBI items. The model parameters were estimated using maximum likelihood. The initial model tested was one in which each item loaded on only one of three latent factors corresponding to its composite subscale or dimension. Somewhat surprisingly, this hypothesized three-factor model did not fit the data well; neither from a statistical ($\chi^2 = 452.55$, $df = 206$, $p = 0.001$) nor a practical ($GFI = 0.78$, $AGFI = 0.73$, $CFI = 0.82$, $RMSEA = 0.09$, and $RMR = 0.11$) perspective. This model was therefore rejected. These results are summarized in Table 2.

Other studies that have examined all 22 items of the MBI have reported cross-loadings for Item 12 and Item 16 (Byrne, 1993; Leiter and Durup, 1994; Schaufeli and Van Dierendonck, 1993). Item 12, “I feel very energetic”, is a PA item that has been consistently found to cross-load strongly and negatively on the EE factor. Item 16, “Working with people directly puts too much stress on me”, is an EE item that has been found to cross-load onto the DP factor. In light of this, a second model (Model 2) was tested that explicitly incorporated these two cross-loadings. Model fit statistics are shown in Table 2. A comparative test of this model against the previous more restrictive model, achieved by contrasting the difference in their $\chi^2$ values relative to the difference in their degrees of freedom, confirmed that freeing these two parameters made a significant improvement in the

### Table 1

Means, standard deviations, 95% confidence intervals, and internal consistency estimates on subscales of the Maslach Burnout Inventory

<table>
<thead>
<tr>
<th>Sample of 151 registered nurses</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>95% CI</th>
<th>Cronbach’s $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotional exhaustion</td>
<td>22.81</td>
<td>11.08</td>
<td>20.9–24.6</td>
<td>0.88</td>
</tr>
<tr>
<td>Depersonalization</td>
<td>6.97</td>
<td>5.77</td>
<td>5.9–7.9</td>
<td>0.80</td>
</tr>
<tr>
<td>Personal accomplishment</td>
<td>38.81</td>
<td>9.01</td>
<td>37.3–40.3</td>
<td>0.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Published norms ($N = 1104$)</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>95% CI</th>
<th>Cronbach’s $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotional exhaustion</td>
<td>22.19</td>
<td>9.53</td>
<td>—</td>
<td>0.90</td>
</tr>
<tr>
<td>Depersonalization</td>
<td>7.12</td>
<td>5.22</td>
<td>—</td>
<td>0.79</td>
</tr>
<tr>
<td>Personal accomplishment</td>
<td>36.53</td>
<td>7.34</td>
<td>—</td>
<td>0.71</td>
</tr>
</tbody>
</table>
fit of the model to the data ($\chi^2 = 42.71, df = 2, p < 0.05$). This test is known as the likelihood-ratio test and is quite useful in testing nested structural equation models (Gonzalez and Griffin, 2001). Nonetheless, inspection of the fit statistics indicated that some degree of model misfit still remained. Indeed, a review of the modification indices (MI) revealed some abnormally large values representing error covariances between various items. Despite common findings of correlated error variance terms, there remains considerable controversy in the CFA literature regarding their interpretability and cause. Bentler and Chou (1987) have remarked that model specification that forces all error terms to be uncorrelated is rarely appropriate with real data.
Incorporation of these correlated error terms into CFA models does not otherwise undermine the factorial validity of the MBI, but rather, it provides a more realistic factorial representation of the observed data structure (Byrne, 1993).

Based on inspection of MIs associated with the correlated error variances, specific error covariance terms were freed sequentially. That is, one parameter was freed and then the likelihood-ratio test was used to assess the significance of improvement in the fit of the model. This process continued until freeing additional parameters did not produce a significant improvement in model fit. The resulting model (Model 3) was consistent with the observed data ($\chi^2 = 192.78$, df = 180, $p = 0.244$). The various fit indices also reflected an acceptable fit of the model to the data (see Table 2). The model and its parameter estimates are shown in Fig. 1. The path coefficients connecting the items to the factors are factor loadings and may be interpreted as standardized regression coefficients. The correlations among the three latent factors: EE, DP, and PA are depicted as curved arrows between the circles. Residual variance terms for each item are also shown. For clarity of presentation, the values of the 23 correlations among the residual variances are not shown in Fig. 1, however, all were significant ($p < 0.05$).

Freed 23 of the 231 error covariance terms significantly improved the fit of the model but called into question the necessity of the cross-loading of Items 12 and 16 that had been observed in previous studies. Consequently, the initial restrictive factor model (Model 1) was re-estimated in the presence of these correlated error terms (Model 4 in Table 2) and compared to Model 3. The deletion of the two cross-loading parameters produced a significant detriment in the model fit ($\chi^2 = 40.24$, df = 2, $p < 0.05$). Thus, even in the presence of relaxing the assumption of uncorrelated error variances the cross-loading of Items 12 and 16 observed in previous studies persists in the factorial structure of the MBI.

Based on these analyses, the hypothesized three-factor structure of the MBI did not fit the data in this sample of Florida registered nurses. An alternative model that explicitly modeled the cross-loading of two items seen in previous studies showed a significant improvement in fit, but was still not adequate to explain the observed data. By sequentially incorporating correlated error variance terms into the cross-loading model, a consistent fit to the data was achieved.

4. Discussion

This study was designed to collect data on registered nurses and test the factorial validity of the Maslach Burnout Inventory using CFA. The hypothesized three-factor model that allowed items to load on one, and only one latent factor corresponding to the MBI subscales, did not fit the observed data. Explicitly modeling the dual association of two items that had been shown in other populations to cross-load onto unintended subscales resulted in a significant improvement in the fit of the factor structure to the data from this sample. However, there was still significant misfit in the model. This discrepancy was reduced by empirically (and conservatively) freeing fewer than 10% of the error covariance terms on the basis of the likelihood-ratio test. Even with these correlated residual terms, the cross-loading of Items 12 and 16 was necessary for an acceptable fit of the three-factor model to the data.

The degree of misfit observed in the hypothesized three-factor model is not unlike that reported in samples of European nurses. Schaufeli and Janczur (1994) found significant discrepancies in the fit of the three-factor model in their data ($\chi^2 = 458.03$, df = 206), and ($\chi^2 = 464.03$, df = 206) from 200 Polish and 183 Dutch nurses, respectively.

In the current study, Item 12, “I feel very energetic”, loaded more strongly (−0.49) on the EE subscale than on PA, its intended subscale (0.17). Compared to the normative exploratory factor analysis reported in Maslach et al. (1996) which used a composite sample of professionals ($N = 1104$), the cross-loading of Item 12 onto the EE factor was more pronounced in this sample (−0.30 vs. −0.49, respectively). When the intended loadings are also compared (0.43 vs. 0.17, respectively), it would appear that a complete reversal in the behavior of this item occurred in this sample of nurses. The behavior of Item 16 relative to the published norms was less dramatic. Item 16, “Working with people directly puts too much stress on me”, loaded 0.53 on EE (its intended subscale) and cross-loaded slightly (0.12) on the DP subscale. The sample loading on the EE factor was comparable to the normative loading reported (0.54). The cross-loading in the sample was not as strong as that reported by Maslach et al. (1996), 0.12 vs. 0.31, respectively.

Item 8, “I feel burned out about my work,” had the highest factor loading (0.86) in the sample; this was similar to the published normative value of 0.81. The estimate of residual variance in Item 8 was 0.25; approximately 25% of the variance in this item was not explained by its relationship to the EE factor. The items with the largest residual variances were Item 13, “I feel frustrated by my job,” and Item 4, “I can easily understand how my recipients feel about things.”

Fitting the model in Fig. 1 to the observed data also provided estimates of the correlations among the MBI subscales (factors) that may be compared to the factor-analytic results published by Maslach et al. (1996). The data factor analyzed initially to support the validity of the MBI was collected on several human service
professions including: 142 police officers, 132 nurses, 125 agency administrators, 116 teachers, 97 counselors, 91 social workers, 68 probation officers, 63 mental health workers, 43 physicians, 40 psychologists and psychiatrists, 31 attorneys, and 77 others. In the current sample of nurses the correlation between the DP and PA factors was \(-0.25\) (CI\(_{95\%}\) -0.39 to -0.09) nearly equal to the norm reported by the authors (\(-0.26\)). However, a one-sample z-test using Fisher’s transformation revealed the sample correlation between EE and DP (0.65, CI\(_{95\%}\) 0.55 to 0.73) was significantly stronger (\(p<0.05\)) than the normative value (0.52) reported. The observed correlation between EE and PA (\(-0.16\), CI\(_{95\%}\) -0.31 to 0.00) was weaker (although not significantly so) than the published normative value (\(-0.22\)). These differences suggest that there maybe a systematic difference in the burnout experience among nurses compared to the majority of other human service professionals. In particular, EE may play a stronger role in the development of feelings of DP for nurses. However, such a conclusion goes beyond the limits of this correlational study.

The correlations observed among the MBI subscales may also be compared with those observed in other studies involving nurses. Schmitz et al. (2000) administered the MBI to a sample of 361 German nurses and estimated the correlations to be 0.45 between EE and DP, -0.13 between EE and PA, and -0.07 between DP and PA. Florida nurses showed a significantly stronger correlation between EE and DP compared to the German nurses (0.65 vs 0.45; t = 2.99, df = 510, \(p<0.01\)). The other two correlations did not differ significantly between the two groups of nurses. Although Schaufeli and Janczur (1994) used CFA to examine the factorial validity of the MBI in samples of Dutch and Polish nurses, they did not report the correlations among the three factors. Working not with nurses but with samples of Canadian school teachers, Byrne (1993) found that structural equation models of the three-factor structure of the MBI also required correlated error variance terms in order to fit the observed data.

Based on common rules of thumb for conducting factor-analytic studies, the sample size used here \((N = 151)\) may, at first, seem small. However, recent advances in factor analysis theory (MacCallum et al., 1996) have demonstrated a lack of validity for such rules of thumb, and have provided a basis for establishing new guidelines for sample size based on the number of variables being analyzed, their communalities, the number of factors extracted, and the number of variables representing each factor. Good recovery of population factor structure can be achieved with samples well below 100 provided the communalities are consistently high (\(\geq 0.6\)). With communalities in the range of 0.5, it is quite possible to achieve good recovery of population factor structure with samples between 100 and 200, provided the ratio of variables to factors is 20:3 or better. In the current study, the average communality was 0.449 and the ratio of variables to factors was 22:3 suggesting that the results obtained represent a close match to population factor structure.

Power in CFA has a different focus than in traditional analysis. Effect size is defined as the degree of model discrepancy in the population and is expressed as the confidence interval around the various fit indices which make use of noncentral \(\chi^2\) distributions (see Steiger et al. (1985) for the theoretical basis for the use of this distribution). Researchers can take into account the imprecision in their sample estimate of fit by testing an hypothesis about the population value of the fit index. In other words, if the fit of a given model was actually mediocre in the population, and we test the hypothesis that the fit of our model is exact, or close, what is the likelihood of rejecting the null hypothesis? Relatively simple methods of estimating the power for tests of exact- and close-fit based on the RMSEA and other fit indices have been developed by MacCallum and others (MacCallum et al., 1996; MacCallum and Hong, 1997). These methods allow researchers to estimate power for detecting the lack of fit in the population for models of varying complexity (degrees of freedom) and various sample sizes. In the current study, the power for tests of exact- and close-fit was estimated for the final model to be 0.92 and 0.98, respectively. Thus, it is with considerable confidence that the model fit to the sample data also holds for the population.

Some limitations must be kept in mind concerning the results reported above. First, the sample was limited to only one region of the United States (Florida). Therefore the findings should not be considered representative of the entire country. Second, the data analyzed were obtained from nurses’ self-reports and consequently, may reflect bias in reporting certain feelings. Third, all the nurses in this study were enrolled in advanced practice programs at a major university. This may have contributed to the higher scores on the PA subscale.

Given the demonstrated cross-loading behavior of certain items on the MBI, calculation of composite subscale scores should not be done blindly, but with caution. Examining the factor structure, using CFA or even exploratory factor analytic techniques prior to working with the subscale scores is recommended in order to avoid spurious correlations among subscales. To maintain comparability with the vast amount of published research employing the MBI in nursing and other professions it is recommended that the subscales be calculated as intended using all 22 items. Maslach et al. (1996, p. 11) suggested that future studies employing SEM and the MBI should omit Items 12 and 16 from their analyses because of the ambiguity in factor loading patterns. I disagree and suggest that
researchers make full use of the strength and flexibility of SEM to assess the MBI’s interrelated components. Specifically, I recommend that four different measurement models be tested in future SEM studies using the MBI: (a) the original hypothesized three-factor model in which each of the 22 items loads only on its intended subscale or factor, (b) the model that allows Items 12 and 16 to cross-load as in this study, (c) the model omitting Items 12 and 16 altogether as suggested by Maslach et al. (1996), and (d) a model that judiciously incorporates correlated error variances between items. The likelihood-ratio test may then be used to determine if there are significant differences in fit among these measurement models and to identify which model best fits the data within any particular study (see Gonzalez and Griffin (2001) for discussion of model identification issues). In the event that no best measurement model can be identified in this way, the relationships among the items and latent factors estimated from all four models of the MBI should be examined for consistency and the results interpreted with caution.

References


