Abstract

It has been suggested that MMPI-2 scoring requires removal of some items when assessing patients after a traumatic brain injury (TBI). Gass (1991. MMPI-2 interpretation and closed head injury: A correction factor. Psychological assessment, 3, 27–31) proposed a correction procedure in line with the hypothesis that MMPI-2 endorsement may be affected by symptoms of TBI. This study assessed the validity of the Gass correction procedure. A sample of patients with a TBI ($n = 242$), and a random subset of the MMPI-2 normative sample ($n = 1,786$). The correction procedure implies a failure of measurement invariance across populations. This study examined measurement invariance of one of the MMPI-2 scales (Hs) that includes TBI correction items. A four-factor model of the MMPI-2 Hs items was defined. The factor model was found to meet the criteria for partial measurement invariance. Analysis of the change in sensitivity and specificity values implied by partial measurement invariance failed to indicate significant practical impact of partial invariance. Overall, the results support continued use of all Hs items to assess psychological well-being in patients with TBI.

Keywords: Traumatic brain injury; MMPI-2; Measurement invariance; Assessment; Correction procedures

Introduction

It has been suggested that people with a traumatic brain injury (TBI) may endorse some items on the MMPI-2 because the content of these items refers to neurological symptoms associated with their injury (Alfano, Neilson, Paniak, & Finlayson, 1992; Gass, 1991). Therefore, elevated profiles that result from neurological distress may be incorrectly interpreted as reflecting psychological distress. While neurological symptoms are not incompatible with a psychological illness, the potential influence of these symptoms on MMPI-2 scores must be considered in patients suffering neurologic conditions (Bachna, Sieggreen, Cermak, Penk, & O’Connor, 1998). Correction procedures recommend removing some items that refer to neurological symptoms to minimize the risk of falsely elevated MMPI-2 profiles in patients suffering a TBI. The most frequently used correction procedure was described by Gass (1991) who recommended removing 14 items from the scoring process.

Individuals who suffer a TBI are at risk of developing a variety of psychological disorders such as anxiety disorders (Moore, Terryberry-Spohr, & Hope, 2006), posttraumatic stress disorder (Moore et al., 2006), depressive disorders (Kreutzer, Seel, & Gourley, 2001), personality disorders (Koponen et al., 2002), and perhaps psychotic disorders (Fujii, 2005). Accurate diagnosis of these types of psychopathology is vital for the care and rehabilitation of a patient who has sustained a TBI. A mood disorder can inhibit recovery from a TBI in both the young and the elderly, with treatment of the mood disorder being crucial to effective rehabilitation (Davis, Reeves, Hastie, Graff-Radford, & Naliboff, 2000; Fann, Katon, Uomoto, & Esselman, 1995). Evaluating eligibility for treatment and likelihood of a successful outcome for a patient who sustained a TBI relies on accurate diagnosis (Handel, Ovitt, Spiro, & Vani Rao, 2007). The importance of psychopathological treatment in patient recovery illustrates the critical nature of psychological evaluation for those experiencing a TBI.
Typically, advocates of correction procedures recommend scoring the MMPI-2 profile of a person with TBI twice if necessary (Gass 1991, 2009). On the first occasion, the profile is scored according to the standard procedure and reviewed for signs of psychopathology. If the profile indicates psychological distress, then the profile is rescored with the 14 items referring to neurological symptoms removed. Accordingly, it is assumed that a person suffering a TBI who is experiencing psychological distress would continue to generate an elevated profile despite the removal of items from a clinical scale. Some validity studies support the correction approach (Gass & Wald, 1997; Rayls, Mittenberg, William, & Theroux, 1997). In contrast, other studies question the correction approach (Arbisi & Ben-Porath, 1999; Bruilot, Strauss, & Spellacy, 1997; Edwards et al., 2003; Glassmire et al., 2003) a debate that remains unresolved (Glassmire et al., 2003; La Chapelle & Alfano, 2005). The scoring correction is included in some contemporary clinical assessment guides (Butcher, 2006; Graham, 2006).

Whether the MMPI-2 requires score correction to most accurately diagnose psychopathology in patients’ with a TBI invokes the ‘rubric of invariance testing’ (Widaman & Reise, 1997). Measurement invariance evaluates the equivalence across populations of the measurement model underlying a test. Testing for measurement invariance evaluates the precise quantitative generalizability of construct validity and is essential to determine whether a psychological test is appropriate for use across different populations (Brown, 2006; Meredith, 1993; Widaman & Reise, 1997). Although the use of correction procedures implies a failure of measurement invariance, no study of MMPI-2 correction has examined invariance, to date.

Measurement Invariance Testing

Measurement invariance testing follows a sequential approach where numerical parameters of the measurement model are initially allowed to be freely estimated across groups in the first step, then parameters are constrained to equality in subsequent steps of analysis. When adequacy of model fit is maintained after the additional imposition of parameter equality then invariance is established. When all parameters meet the requirements for invariance, this is known as strict measurement invariance. When some parameters fail and some parameters meet the requirements for invariance, this is known as partial measurement invariance (Brown, 2006; Meredith, 1993; Widaman & Reise, 1997).

Computational and model-fitting challenges abound when modelling item level data, particularly with larger numbers of test items (Bandalos, 2008). Therefore, in one of a series of studies, this study will model items from Hs of the MMPI-2 and test for invariance between a TBI sample and the MMPI-2 normative data. The 32 items on Hs are related to individual perception of physical health and bodily functioning. Only eight items are unique to Hs sharing items with Clinical Scales D, Hy, and Sc. Hs has 20 items repeated on Scales D and Hy leading some to suggest Hs can be considered a measure of general psychological distress (Friedman, Lewak, Nichols, & Webb, 2001). Five of the Gass (1991) correction items (items 101, 149, 175, 179, and 247) are found on the Hs scale. Therefore, Hs is an important scale in the assessment of the Gass correction model. Despite the wide use of the MMPI-2, review of the literature found no previous item level factor analysis of the Hs scale.

How to Establishing Measurement Invariance

A full latent variable model comprises both a structural model and a measurement model (Byrne, 1998). The structural model describes relationships between the latent factors and constructs. The measurement model describes the relationship between the latent factors and the observed test scores. Testing measurement invariance involves testing the similarity of the measurement model across populations. When the observed test scores are dichotomous items, such as with the MMPI-2, there are three parameters in the measurement model that are required to be equivalent to establish strict measurement invariance, namely, factor loadings, item thresholds, and item residuals (Millsap & Yun-Tein, 2004). The factor loadings and thresholds describe the relationship between endorsement of each MMPI-2 item and the respective latent variable. Residuals reflect item-unique variance and item measurement error. Importantly, with dichotomous items, the results of the measurement invariance testing with theta parameterization are equivalent to a test of differential item functioning under item response theory (Glöckner-Rist & Hoijtink, 2003).

Millsap and Yun-Tein (2004) recommend when using models with dichotomous items to initially constrain thresholds to equality for identification purposes. In subsequent steps, item loadings, then residuals, are constrained to equality. Alternatively, Muthén and Muthén (1998–2010) propose residuals are constrained to equality in the first step with loadings and thresholds freely estimated and then in a subsequent step, loadings and thresholds are constrained to equality with residuals freely estimated.

With the invariance estimation approach described by Millsap and Yun-Tein (2004), there is no test of threshold invariance (Bontempo & Hofer, 2007). Investigating invariance of threshold parameters is valuable because such a test allows evaluation of the item difficulty parameter or differential item function, a particular concern in item response theory (Kim & Yoon, 2011). In contrast, the approach advocated by Muthén and Muthén (1998–2010) does not include a test of strict invariance which is required to implement the practical sensitivity and specificity analysis available should partial measurement invariance
be observed (Millsap and Yun-Tein 2004). Therefore, a hybrid approach to invariance analysis was examined in this study, as follows:

Step 1: Determine a baseline measure of model fit across the two samples of interest.

Following the first stage of the Muthén and Muthén (1998–2010) approach, the residuals were constrained to equality, along with one loading from each factor, and the factor means (which were set to zero) for identification purposes.

Step 2: Test for strict invariance.

All measurement parameters (loadings, thresholds, and residuals) are constrained to equality in line with the final stage of the Millsap and Yun-Tein (2004) approach. The measures of model fit provided in this step are then compared with those derived from Step 1. If measures of model fit do not deteriorate significantly then measurement equivalence is established, otherwise Step 3 is completed.

Step 3: Defining a partial measurement invariance model.

A model with the least number of parameters allowed to differ across groups is defined to meet statistical criteria of no significant loss of fit compared with the model in Step 1.

If partial measurement invariance is observed, Millsap and Kwok (2004) recommend reviewing the changes in sensitivity and specificity values of diagnostic classifications using respective factor scores across the partial and strict invariance conditions. Any significant reduction in sensitivity or specificity from partial to strict invariance thus reflects the practical impact of failure of strict invariance. A change in classification sensitivity or specificity may be considered insubstantial if there is no significant change in the sensitivity or specificity of diagnostic classification under partial versus strict invariance (Millsap & Kwok, 2004). Partial measurement invariance reflects the performance of the instrument in a real clinical application while strict invariance reflects the performance of the instrument in the ideal condition of complete measurement model equivalence across groups.

Although the method outlined by Millsap and Kwok (2004) provides an important means by which to evaluate the impact of partial measurement invariance, there appear to be few if any published examples of this method to date. If the observed measurement model fails the test of strict invariance, but does meet the requirements for partial measurement invariance, then the sensitivity and specificity values of the partially invariant model will be reviewed to determine the clinical impact on diagnostic accuracy.

Method

Participants

The study used six samples. The study included a clinical sample of 259 TBI patients (162 male and 97 female) assessed at a private practice specializing in forensic TBI evaluations who were tested using the MMPI-2 from 1995 to 2005. After removing 17 cases with missing data or invalid profiles, the final sample ($n = 242$) had an average age of 35.7 years ($SD = 12.9$). In contrast to full-information maximum-likelihood with continuous test items, the categorical item factor-analysis algebra requires exclusion of cases with any missing data. In addition, conventional MMPI-2 validity rules were used to exclude patients who had invalid profiles (i.e., CNS > 30 raw score, VRIN or TRIN T-scores > 80, or $F$ or $F_{Hs}$ T-scores > 110). Data on TBI severity were available for 212 participants, derived from the examining neuropsychologists report and stratified according to the criteria of Williamson, Scott, and Adams (1996); 117 (55.2%) of participants had experienced mild injuries, 46 (21.7%) moderate, 32 (15.1%) moderately severe, and 17 (8%) severe injuries. A Spearman’s rho correlation analysis revealed no significant correlation between TBI severity and Hs scale score ($r = .047, p = .502$).

After removing cases with missing data, five samples were generated from the MMPI-2 normative data. Norm A ($n = 1,275$) and Norm B ($n = 1,273$) were randomly generated nonoverlapping subsets of the entire normative sample. An all-female sample (female norm, $n = 1,432$) with an average age of 40.4 years ($SD = 15.1$), and an all-male samples (male norm, $n = 1,116$) with an average age of 41.6 years ($SD = 15.2$) were also generated from the MMPI-2 normative data. The TBI, Norm A and Norm B samples were each used to generate a potential baseline factor models. The purpose of using several independent samples was to increase that likelihood that the baseline factor model generalized across populations. The female norm and male norm samples were included in the replication procedure to ensure that the preferred baseline factor model generalized across gender.

Finally, on the assumption that a satisfactory baseline model was observed across all samples, an MMPI-2 Community sample was created for the measurement invariance testing ($n = 1,786$, male 1,116 and female 670) which matched the proportion of males and females observed in the overall TBI sample. This sample was relatively large, compared with the TBI sample, because a smaller Community sample produced missing item covariances, due to skewed item responses. This study was approved by the Human Research and Ethics Committee of St Vincent’s Hospital, Melbourne.
Exploratory factor analysis (EFA), confirmatory factor analysis (CFA), and invariance testing were conducted using Mplus Version 6.11. The weighted least squares means and variance adjusted (WLSMV) estimator was employed for dichotomous indicators.

Procedure and Results

Invariance testing requires a defined baseline model meeting the criteria of configural invariance (Byrne, 1998; Widaman & Reise, 1997). In view of the paucity of relevant published factor models for MMPI-2 Scale Hs, it was necessary to define a baseline model. The first step was to define a candidate baseline model the Norm A, Norm B, and TBI samples (from now on called a candidate model). The next step was a replication process using CFA, which permitted a review of the candidate model across multiple samples. In this stage, the female norm and male norm sample were included along with the samples used for the candidate models. The addition of these two samples meant there were five samples included in the replication procedure, two of which were gender specific which helps ensure the generalizability of the selected factor model across genders.

Baseline Model

Admissibility criteria. Factor models were assessed against four admissibility criteria before evaluation of fit (Brown, 2006). Firstly, a minimum of three indicators is required per factor for identification. Secondly, standardized item loadings cannot be >1. Thirdly, to ensure model factors must be clearly articulated the standardized factor covariance must be >2 standard errors <1. Fourthly, all items should load significantly on respective factors.

Admissible models were subjected to χ² comparison tests for nested models and the root-mean-square-error approximation (RMSEA), the Tucker-Lewis Index of fit (TLI), and the Comparative Fit Index (CFI) fit indices were compared for both nested and non-nested models. When modelling item level data, it is more difficult to obtain good fit than when modelling parcels or composites (Bandalos, 2008; Little, Cunningham, Shahar, & Widaman, 2002; Wirth & Edwards, 2007). Therefore, criteria for acceptable model fit will be slightly less stringent than is currently recommended for models based on aggregate or parceled scores. Briefly, current practice for aggregate score models is as follows. Multiple fit indices were evaluated as recommended by Vandenberg and Lance (2000) and Brown (2006). The RMSEA was required to be <0.05 which indicates the data fits the model well (MacCallum & Browne, 1993). In addition, TLI and CFI values close to or >0.95 indicate a well-fitting model, values >0.90 indicating acceptable model fit. If the values of TLI or CFI fall <0.90 then the model will be rejected (Brown, 2006; Hu & Bentler, 1999).

Performance of a factor model across samples is crucial (Thomson, 2004). A well-replicated model is shown to fit the data well despite individual sample variance. Furthermore, the admissibility of a candidate model is assessed in all samples. If a candidate model is not admissible in any sample, then it will be removed from consideration as the baseline model. In addition to examination of model fit criteria, item R², factor articulation, and review of item content of items for model interpretability will also be considered (Henson & Roberts, 2006; Kline, 2010).

Procedure for finding a baseline model. Following recommendation from Henson and Roberts (2006) and Floyd and Widaman (1995), EFA was conducted to find the model with the least number of factors that was statistically nonsignificant using a χ² test (p > .05). The next step was to run a simple-structure CFA using the EFA results. Modification indices and correlation residuals were examined to identify any source of model misspecification. Residual correlations were also examined for evidence of method effects, for example, items referring to similar content (Brown, 2006; Kline, 2010).

A four-factor model was found from EFA to fit in the TBI sample (WLSMV χ² [374, N = 242] = 390.43, p = .27). Re-specification as a simple-structure CFA also supported the four-factor model as best-fitting but with three modifications, re-locating both items 45 from the 3rd factor and 53 from the 2nd factor to Factor 4. Additionally specifying a correlated residual between pair 3 (sleep and energy) and 39 (sleep) was supported both theoretically and empirically, The resulting four-factor CFA model fitted the Hs data from the TBI sample (WLSMV χ² [457, N = 242] = 510.55, p = .058).

Replication procedure. Table 1 shows fit statistics for the TBI candidate model in the other samples. Re-specification of factor loadings for items 45 and 53 and the correlated residual between item pair 3 and 39 was supported in the other samples, all re-specified parameter estimates being highly significant. This supports the model meeting the requirement of configural invariance across a variety of samples. In terms of overall fit, the TBI model fitted reasonably well in the normative samples, with excellent RMSEA values that were either on the edge of or included in the 90% confidence interval (CI) for the TBI sample and best judged as not statistically significantly different in terms of interval estimation (Cumming & Finch, 2005). The TLI
and CFI values were slightly poorer in the normative samples (Table 1) but poorer overall fit is to be expected with item level data (Bandalos, 2008; Little et al., 2002).

To ensure that a better-fitting alternative model had not been overlooked, extensive examination of simple-structure CFA’s derived from EFA analysis in Norm A and Norm B were also examined. Alternatives models were examined for overall fit, item level explained variance, and theoretical coherence defined as better models of homogeneous constructs. No better-fitting model to that described in Table 2 from the TBI sample was found through this exhaustive process (details available on request from the authors).

As a consequence the best-fitting model derived from the TBI sample was adopted as the baseline model for subsequent invariance analysis. Factor 1 is labeled ‘gastrointestinal complaints’, Factor 2 ‘orofacial symptomology’, Factor 3 ‘sleep quality/energy levels’, and Factor 4 ‘general health’.

### Measurement Invariance Testing

In Step 1, the residuals are constrained to equality across groups with loadings and thresholds allowed to vary (except for loadings on marker variables). In Step 2, all residuals, loadings, and threshold are constrained to equality across groups. If the second test fails to meet the requirement for measurement invariance, then item parameters (loadings and thresholds) will be freed one pair, per item, at a time to find a model that meets the requirements of partial measurement invariance. Invariance is established if the CFI does not increase by $>0.002$ despite the imposition of equality on the additional measurement model parameters (Meade, Johnson, & Braddy, 2008). If failure to find strict invariance is found, the backwards elimination approach will be applied to determine which parameters violate the hypothesis of measurement invariance as recommended (Cheung & Rensvold, 1999). Review of the modification indices will show which item parameters are to be freed (Brown, 2006; Kline, 2010).

Fit measures for the invariance tests are shown Table 3. Baseline invariance testing (Step 1) produced acceptable fit. When all loadings and thresholds were constrained to equality across groups (Step 2) a decrease in the CFI value of 0.015 was observed, failing the test of strict invariance. The backwards elimination procedure was employed to define the partially invariant model. The modification indices for the loadings and threshold parameters were reviewed. The item with the highest aggregation of the modification indices was selected as the item to be freed across groups. This process was repeated until a model was defined which met the criteria for partial measurement invariance, namely, a decrease in CFI of $<0.002$.

This procedure showed that sequentially freeing loadings and thresholds for items 10, 47, 3, and 39 was sufficient to meet the requirements for partial measurement invariance. These items loaded on Factors 3 and 4 in the factor model shown in Table 2. That is, when the loadings and thresholds for these four items were freely estimated across groups, the RMSEA was unchanged in terms of alternative values falling within either 90% CI, and the CFI decreased by 0.001 when compared with the fit indices generated for Step 1. Overall, the evaluation of measurement invariance showed that Factors 1 and 2, described above (Table 2), satisfied the criteria for strict invariance. In contrast, Factor 3 contained three items that failed the test of strict invariance and Factor 4 contained one item that failed the test of strict invariance. All five items (101, 149, 175, 179, and 247) from the Gass (1991) correction procedure met the criteria for partial invariance (see Table 2). Factor loadings and explained variance for the partially invariant measurement model for Community and TBI samples are also shown in Table 2.

### Practical Impact Analysis

Practical impact analysis after finding a partially invariant measurement model was completed as suggested by Millsap and Kwok (2004). The interested reader is directed to the original paper cited for complete details of the procedure. Briefly to complete

---

**Table 1.** Four-factor TBI model replication analysis using CFA

<table>
<thead>
<tr>
<th>Sample</th>
<th>$\chi^2$</th>
<th>$df$</th>
<th>$P$</th>
<th>RMSEA</th>
<th>CFI</th>
<th>TLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full female</td>
<td>954.499</td>
<td>457</td>
<td>$&lt;.0001$</td>
<td>0.028</td>
<td>0.025–0.030$^a$</td>
<td>0.937</td>
</tr>
<tr>
<td>Full male</td>
<td>696.796</td>
<td>457</td>
<td>$&lt;.0001$</td>
<td>0.022</td>
<td>0.018–0.025$^a$</td>
<td>0.938</td>
</tr>
<tr>
<td>Norm A</td>
<td>837.057</td>
<td>457</td>
<td>$&lt;.0001$</td>
<td>0.026</td>
<td>0.023–0.028$^a$</td>
<td>0.933</td>
</tr>
<tr>
<td>Norm B</td>
<td>819.129</td>
<td>457</td>
<td>$&lt;.0001$</td>
<td>0.025</td>
<td>0.022–0.028$^a$</td>
<td>0.942</td>
</tr>
<tr>
<td>TBI</td>
<td>505.529</td>
<td>457</td>
<td>0.578</td>
<td>0.021</td>
<td>0.000–0.031$^a$</td>
<td>0.979</td>
</tr>
</tbody>
</table>

Notes: $^a$The 90% CI reported from Mplus.
practical impact analysis, sensitivity and specificity values are required for the TBI sample in the strict and partial measurement invariance condition for any factor that failed the test of strict invariance. In this case, items that failed the test of strict invariance were from Factors 3 and 4. Observed scores and factor scores are used as respective proxy measures of test-related diagnosis (observed scores) and the criterion variable of psychopathology (factor scores). These measures are necessary to calculate the sensitivity and specificity values, the values of which are then compared under assumptions of strict invariance versus partial measurement invariance, as explained below.

The Millsap and Kwok (2004) procedure requires selection of a cutoff score to differentiate those with and without psychopathology. An observed score and factor score that reflects the 93.32 percentile rank will be used to define cutoff points. This percentile rank is 1.5 $SD$ above the mean which corresponds to an approximate $T$-score of 65. An MMPI-2 $T$-score of 65 or greater is often used to reflect a score of clinical interest. A reference sample is required to define the cut-point scores (Millsap & Kwok, 2004) which is the Community sample in this study.
Therefore, each person in a sample will have an observed score and factor score. Each factor will have a cut-score to reflect the presence or absence of psychopathology and each observed score will have a cut-point to represent diagnosis. In traditional notation for assessing the sensitivity and specificity of a diagnostic instrument, the factor scores represent the presence or absence of the latent estimate of psychopathology, while the observed scores represent the diagnosis of psychopathology (present or absent). See Fig. 1 which represents a hypothetical distribution of factor scores and observed scores into each of the four categories described with a hypothetical cut-point at the 90th percentile.

To undertake the sensitivity and specificity analysis, a count of the number of participants that are located within each category, or quadrant, in Fig. 1, is calculated. These tabulated data will be used to calculate sensitivity and specificity values for each factor in each MMPI-2 Scale that fails the test of strict invariance. This procedure is repeated for the partial measurement invariance condition and the strict invariance condition. Difference in values across the invariance conditions may occur as changes in parameter values of the underlying measurement model potentially alter the calculated factor scores. Changes in factor scores then may alter the sensitivity and specificity values. The practical impact of non-invariant items is then evaluated by comparing the sensitivity and specificity values in the partial measurement invariance condition (the clinical application of the MMPI-2) with the strict invariance condition (the desired condition of measurement model equivalence across groups). See Fig. 1 which represents a hypothetical distribution of factor scores and observed scores into each of the four categories described with a cut-point at the 90th percentile, in both a partial measurement invariance and strict invariance condition.

The TBI population is the focus for this study therefore only the sensitivity and specificity values for the TBI sample are required for review. An important practical impact is suggested if, for the TBI sample the value of sensitivity or specificity in the partial measurement invariance condition is below the lower bound of the 95% CI in the strict invariance condition.

Table 4 shows the sensitivity and specificity for the TBI sample on the factors of interest. All values generated in Table 4 were calculated using the VassarStats calculator (Lowry, 2013). This review is necessary for both Factors 3 and 4 because these factors showed partial measurement invariance above.

A comparison of the sensitivity and specificity values for the TBI sample found no instances where the value in the partial measurement invariance condition was below the lower bound of the 95% CI value in the strict invariance condition (Table 4). The practical impact analysis supports retaining all items from Factors 3 and 4 when assessing persons with a TBI because there is

![Fig. 1. Distribution of factor scores and observed scores into quadrants assuming cut-point at 90th percentile for the partial measurement invariance (continuous line) and strict invariance (broken line) conditions.](image)

| Table 4. Sensitivity and specificity analysis (with 95% CI) for Factors 3 and 4 across invariance conditions in the TBI sample |
|---|---|---|---|
| | Factor 3 | Factor 4 |
| | Sensitivity 95% CI | Specificity 95% CI | Sensitivity 95% CI | Specificity 95% CI |
| Partial measurement invariance | 0.877 | 0.949 | 0.789 | 0.987 |
| | 0.822–0.917 | 0.814–0.991 | 0.718–0.847 | 0.919–0.999 |
| | 0.833 | 1.000 | 0.799 | 0.987 |
| | 0.775–0.879 | 0.840–1.000 | 0.728–0.856 | 0.919–0.999 |
no significant loss of sensitivity or specificity. In other words, there was no significant impact on the sensitivity or specificity values from the failure to find strict invariance. Therefore, retaining the invariant items and assuming full invariance does not compromise criterion-related validity. Further support for this conclusion was the significant and large correlation between factor scores in the partial measurement invariance condition, and the factor scores in the strict invariance condition observed on both Factor 3 (Spearman’s rho = .988, p < .01) and Factor 4 (Spearman’s rho = .999, p < .01).

The invariance of items on Factors 1 and 2 combined with the finding of no practical impact from the non-invariance of items on Factors 3 and 4 supports retaining all MMPI-2 Hs items when assessing patients who sustained a TBI.

**Discussion**

Exploratory and CFA showed a four-factor model best represented the 32 items from Hs. The strong fit statistics generated across all samples provides confidence in the final model selected, despite the poor fit expected from a model based on items rather than parcels (Table 2). Extensive testing in various subsets of the MMPI-2 normative sample was unable to identify a better-fitting model. Measurement invariance tests across the TBI and Community samples found this four-factor model met the requirements for partial measurement invariance when loadings and thresholds were freed on items 3, 10, 39, and 47. With the WLSMV estimator, this partial measurement invariance result is equivalent to demonstrating differential item function in a multifactorial item–response theory approach (Muthén & Muthén, 1998–2010). The correction procedure proposed by Gass (1991) includes items 101, 149, 175, 179, and 247 from the Hs scale. If these items were to falsely inflate MMPI-2 profiles in people experiencing a TBI, then these items should have been identified as a source of measurement variance in our study. However, none of the items were found to fail the test of strict invariance. Furthermore, the practical impact analysis supports retaining all Hs items when assessing patients who sustained a TBI.

Central to the argument for correction procedures is the proposition that endorsement of neurologic symptoms generates over-prediction of psychopathology in patients, for example, with TBI. Specificity is the probability that a test of psychopathology will correctly diagnose a person as healthy when no psychological disorder exists. Therefore, the hypothesis of overprediction can be assessed by reviewing the specificity values in the partial measurement invariance condition. Specificity values were 0.949 and 0.987, for third and fourth factors, respectively (Table 4). These results support the contention that concerns with overprediction on the Hs scale are unjustified.

These findings support some previous research that also concludes that removing the MMPI-2 items suggested by Gass’s (1991) procedure is not appropriate for clinical diagnosis (Arbisi & Ben-Porath, 1999; Brulot et al., 1997; Edwards et al., 2003). However, none of the previous research used measurement invariance analysis which provides a direct test of the implicit assumption of failure of invariance. Indeed, apart from item–response theory analysis in trait-homogeneous item sets, measurement invariance provides the only direct and strong test of the hypothesis of differential item function across populations. The results of the current study suggest that while items in the MMPI-2 that may refer to common neurological symptoms after injury, these items do not diminish the validity of an assessment in a person who has suffered a TBI. While invariance of remaining Gass items needs to be examined with other MMPI-2 scales, these findings imply that measurement invariance may hold for all the Gass items. However, to clearly resolve the question, measurement invariance analysis need to be undertaken with all relevant items from other MMPI-2 scales some of which are now incorporated into the MMPI-2-RF scales.

The baseline CFA model for Hs used in this study described four factors sequentially named: ‘Sleep quality/energy levels’, ‘Orofacial symptomology’, ‘General health,’ and ‘Gastrointestinal complaints’. Notwithstanding the value of independent replication studies, confidence in the four-factor model identified is enhanced by the observation that the four-factor model selected was tested across a variety of samples in this study. Additionally, the nonsignificant χ² result for the factor model in the TBI sample is notable, since a statistically nonsignificant CFA model is unusual even with parcelled indicators (Bandolos, 2008; Little et al., 2002). The successful replication four-factor model across diverse samples provides confidence in its generalizability.

Applying the practical impact analyses was crucial to clarify the interpretation from the finding of partial measurement invariance in the Hs factors. A conservative interpretation has been that if a test fails a test of strict invariance, then the test may be inappropriate for the target population (Meredith, 1993). However, other researchers have noted that it is unlikely that a test will meet the requirements for strict invariance, especially a test including large numbers of items (Borsboom, 2006; Byrne & van de Vijver, 2010). Therefore, it is becoming increasingly important to determine whether the non-invariant items from a partial measurement invariance model can be retained. Applying the Millsap and Kwok (2004) practical impact analysis showed that the clinical validity of the diagnostic scale was not reduced by inclusion of a small number of items that failed the test of invariance.

In this study, only Hs was subjected to invariance testing. Which leaves items 31, 106, 147, 165, 170, 172, 180, 295, and 325 from the Gass (1991) correction procedure requiring analysis. Examination of clinical scales Hy and Sc would provide invariance testing for all the items in the Gass correction model, which will be the subject of future study. As noted, scales Hy and Sc were not included in this analysis because of the limitations on the number of categorical items that can be modelled in one analysis.
While the study did employ a personal injury sample, as noted, the sample was screened for invalid MMPI-2 profiles. Separate cognitive performance validity tests were not used to exclude participants from this study sample. Failure of cognitive performance validity tests does not necessarily result in exaggerating on psychopathology symptom report measures such as the MMPI-2. However, the lack of performance validity testing as a screening device for participants is a potential limitation of the study.

In summary, the four-factor model of MMPI-2 Hs defined in this study was found to satisfy the criteria for partial measurement invariance across a TBI sample and a gender-matched subset of the MMPI-2 normative sample. None of the items from the Gass correction model that are included in the Hs scale were found to fail the test of invariance. In addition, practical impact analysis of the four-factor model supports retaining all items of the Hs scale when assessing patients with a TBI. Furthermore, while this study assessed the factor model with a TBI sample, additional groups from the spectrum of neurological impairments require evaluation because patients experiencing diverse illnesses and injuries may also endorse physical and neurological symptoms potentially complicating interpretation of the MMPI-2.

Acknowledgement

MMPI®-2 (Minnesota Multiphasic Personality Inventory®-2) normative data used with permission of the University of Minnesota Press.

Conflict of Interest

None declared.

References
