Major depressive disorder in primary care: screening, diagnosis and treatment
Baas, K.D.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
CHAPTER 5

Measurement invariance with respect to ethnicity of the Patient Health Questionnaire-9 (PHQ-9)


Journal of Affective Disorders 2010; 129: 229-235
Abstract

Background The Patient Health Questionnaire-9 (PHQ-9) has been widely used in research and clinical settings. To be able to attribute differences in PHQ-9 scores between groups with different cultural backgrounds to differences in the level of depression, the instrument has to possess measurement invariance.

Methods Data from the Apollo-D study were used. We used two strongly contrasting cultural groups (n=1772). Measurement invariance was assessed by comparing four categorical single factor models with an increasing number of restrictions, representing an increasingly stronger measurement invariance assumption.

Results The PHQ-9 was measurement invariant for ethnicity in women and partially measurement invariant for ethnicity in men. The item ‘psychomotor problems’ seemed to be culturally biased in the Surinam Dutch males. It had a higher loading and threshold compared to Dutch males.

Limitations The sample is restricted to high risk primary care patients, we did not include a gold standard measure of depression and the analyses pertain to a single cross cultural comparison.

Conclusions The observed higher total depression score for females in the Surinam Dutch group can be attributed to a true difference in the latent trait depression. For Surinam Dutch and Dutch men some caution is warranted when comparing results obtained with the PHQ-9. In the former group the scores may be biased slightly downward. Future research is needed to examine how the item ‘psychomotor problems’ performs in different populations. These findings highlight the necessity of establishing measurement invariance before drawing conclusions based on observed scores.
Introduction

Sadness is a universal phenomenon\(^1\) and the broader concept, depression, is a common cause of poor health all over the world.\(^2\) Early studies that looked at depression across cultures indicated that in each culture depression consists of components of three symptom dimensions: affective, cognitive and somatic.\(^3\) The concept of depression thus seems to represent approximately the same cluster of symptoms in different cultural groups. However, each culture knows its own expressions of distress which makes diagnosing a challenging task. Also, a person’s cultural background may bias screening outcomes and diagnostic procedures.\(^4\) If this is the case, the screening instrument can be considered culturally sensitive.

Suppose we have two culturally different groups of participants who fill in a depression screener. The first group acquires a lower mean score than the second group. The question that arises is the following: does the difference in test scores reflect an actual difference in the severity of the depression or is there another, unobserved, difference between these participants that attributes to this difference in mean scores? This question addresses one of the main methodological issues in cross cultural research, namely measurement invariance.

A measurement invariant test across different groups implies that group differences in the test scores can be attributed to group differences on the underlying variable.\(^5\) If women have a higher score on a measurement invariant depression test than men, the group difference in test score is attributable to depression and not to some other, unobserved, difference between men and women.

In order to compare the true incidence and/or prevalence between ethnic groups, instruments must be measurement invariant for ethnicity (e.g. culturally insensitive). If this is not the case, biased items should be identified. Bias occurs when the probability of endorsing a test question differs for individuals who experience the same underlying level of a variable (for example “depression”), but belong to different (cultural) groups.\(^6\) So far, only a few studies have examined measurement invariance for ethnicity in depression screening instruments using techniques that control for the level of depression.\(^7,9\) The two-parameter logistic Item Response Theory model (2-PL IRT model)\(^10,11\) and categorical factor model\(^12\) both offer a powerful tool to separate measurement bias from true group differences.\(^13\)

The diagnosis and classification of depression is based on the presence of nine symptoms (affective, cognitive and somatic).\(^14\) These items have a long existence and withstood the test of time. The Patient Health Questionnaire (PHQ) is a short self-report version of the Primary Care Evaluation of Mental Disorders (PRIME-MD).\(^15\) The PHQ-9 is the 9 item sub-scale that screens for depressive disorders.\(^16\) Since its publication it has been widely used in research.\(^17\) Each item of the PHQ-9 evaluates the presence of one of the
nine DSM-IV criteria for depressive disorder during the last two weeks. This makes the instrument useful for cross cultural comparison of the cluster of depression symptoms. In this paper we want to address the question whether the PHQ-9 is measurement invariant (e.g. culturally insensitive) for ethnicity. For this purpose we will use data from the Apollo-D study and compare two strongly contrasting cultural groups using a categorical factor model.

**Methods**

**Setting**
The Apollo-D study was conducted from January 2005 through July 2007 in patients between 18 and 70 years old from general practices located in a wide area around two academic institutions in the Netherlands: the Academic Medical Center in Amsterdam and the University Medical Center St Radboud in Nijmegen. These general practices were connected to these institutions with regard to education and patient care. The patients from these general practices are comparable to other primary care patients. The study protocol was approved by the institutional ethics review committee of both centres (see also Baas et al18).

**Patients**
The APOLLO-D project aimed to screen and detect depression in primary care. We selected patients who had a high risk for depression, namely patients with mental health problems (MHP), patients frequently attending their general practitioner (FA) and patients with unexplained somatic complaints (USC). Previously we published specific details of the selection of these high risk groups.18

**Procedure**
An invitation letter signed by the patients’ own general practitioner was sent to the selected high risk patients. It described the study and contained an informed consent form and the PHQ.15 If a patient did not respond within two weeks, a reminder was sent.

**Selection of the two subgroups**
For the analysis we selected two subgroups, Surinam Dutch (N = 321: M/F = 89/232; Age, mean (SD) = 48.7 (12.2)) and Dutch participants (N = 1451: M/F = 556/895; Age, mean (SD) = 51.6 (12.9)). Both groups differ culturally, but speak the same native language, namely Dutch (so both could be administered the Dutch PHQ-9). Ethnic groups were defined by the country of birth of an individual and his/ her parents, a widely accepted
Measurement invariance of the PHQ-9

basis for the identification of ethnic groups. For the Surinam Dutch participants both parents had to be born in Surinam and for the Dutch participants both parents had to be born in the Netherlands. For Surinam ethnicity this is a more stringent criterion than the standard definition of the Dutch Central Bureau of Statistics (CBS) which requires only one of the parents to be born in Surinam. We used this more stringent criterion to maximize cultural contrast between the groups. Of the Surinam Dutch participant group 88% was born in Surinam and migrated to the Netherlands. These participants came from the former Dutch colony of Surinam (South America) and are mainly descendents of former slaves from West-Africa. So our two groups differ in ethnic origin and ethnic aspects as norms, values, traditions, customs and religion.

Measurement

The PHQ-9 screens for depression and has been shown to be a valid instrument to do so. Each item of the PHQ-9 evaluates the presence of one of the nine DSM-IV criteria for depressive disorder: (1) anhedonia, (2) depressed mood, (3) trouble sleeping, (4) feeling tired, (5) change in appetite, (6) guilt or worthlessness, (7) trouble concentrating, (8) feeling slowed down or restless, (9) suicidal thoughts. Each of the 9 items can be scored 0 (not at all), 1 (few days), 2 (more than half the days) and 3 (nearly every day). The PHQ-9 offers a categorical algorithm for the diagnosis of depressive disorder based on modified criteria of depressive disorder according to the DSM-IV. A positive outcome on the categorical algorithm requires that 5 or more of the 9 depressive symptom criteria are present “more than half the days” in the past 2 weeks (suicidal thoughts counts if present at all), and 1 of these symptoms has to be depressed mood or anhedonia. As a consequence, the rating is collapsed from an ordinal one into dichotomous one. Except for item 9, all item scores of 0 (i.e. not at all) and 1 (i.e. few days) were recoded 0 and all item scores of 2 (i.e. more than half the days) and 3 (i.e. nearly every day) were recoded 1. For item 9 the item score 0 was recoded 0 and the item scores of 1, 2 and 3 were recoded 1 (i.e. suicidal thoughts counts if present at all).

The categorical factor model

Since the items of the PHQ-9 were all categorical in nature, we employed the categorical single factor model. This model hypothesizes that answers to categorical items are sampled from an underlying continuous distribution. In case of dichotomous items, the item threshold is related to the amount of the underlying characteristic needed to give the “yes” response. For example, it is likely that “depressed mood” is a continuous variable yet the item is dichotomous. If the threshold is high, then people need to experience more severe depression in order to respond “yes” to that item. Furthermore, this model assumes that one latent trait (“depression” in this case) causally influences the continuous variables that underlie the categorical items with a strength (loading)
that is separately estimated for each item: the higher the loading, the higher the causal influence of the latent trait. For example, it is reasonable to assume that the item “depressed mood” is more strongly influenced by the latent trait “depression” than the item “sleep disturbances”.\(^1\) Finally, the model assumes that not all observed score variance can be explained by the latent trait but, instead, some of it is due to measurement error. Such error variance is represented in the final parameter of the model, the residual variance for each item: the higher the residual variance of a particular item, the more measurement error is associated with that particular item.

**Statistical analyses**

*Unidimensionality assumption*

We first verified the unidimensionality assumption by fitting a confirmatory categorical single factor model to the entire sample in Mplus.\(^2\) By verifying unidimensionality, one checks the pivotal assumption of the PHQ-9 that, regardless of group membership, the items measure the symptoms of one disorder, depression.

*Statistical model and fit comparisons*

Subsequently, we assessed measurement invariance for ethnicity of the entire sample. Because women are more prevalent in the Surinam Dutch subgroup and women are more inclined to endorse items on an affective screener than men, we also conducted measurement invariance tests for gender. We did this by comparing the fit of four categorical multiple group single factor models ranging from a non-invariant model to a fully invariant model (see table 1). Except for the baseline model, each model was compared to the previous model.\(^1\)

Measurement invariance is demonstrated when the second model does not fit the data significantly poorer than the baseline model and when the third model does not fit the data significantly poorer than the second model. Full invariance is demonstrated if, in addition, the fourth full invariance model does not fit the data significantly poorer than the third model. Partial measurement invariance of the screener is demonstrated when some but not all items are measurement invariant.\(^2\)

We used Mplus 4.2 for the estimation of all models with a Weighted Least Squares (WLS) estimator for the entire sample. Because the WLS estimator requires large samples for adequate performance we also used a Weighted Least Squares Mean and Variance adjusted (WLSMV) estimator for subsamples.\(^2\) Because the chi square \(\chi^2\) statistic is overly sensitive to minor deviations from a good fit in very large samples, the fit of our models is not only assessed with the \(\chi^2\) statistic (with \(0 \leq \chi^2 \leq 2df\) indicating good

---

\(^1\) Please note that the interpretation of the threshold parameter in a categorical factor model is equivalent to that of the difficulty parameter in Item Response Theory (IRT) approaches. Also, the interpretation of the loading parameter is equivalent to that of the discrimination parameter in IRT approaches.
Measurement invariance of the PHQ-9

Chapter 5

fit and $2df \leq \chi^2 \leq 3df$ indicating acceptable fit) but also with the Root Mean Square of Approximation (RMSEA $\leq 0.06$ indicating good fit\(^2\)) and the Comparative Fit Index (CFI $\geq 0.95$ indicating good fit\(^{29,30}\)). For comparisons of model fit in the total sample we used the $\chi^2$ difference test with the WLS estimator and we used the Mplus DIFFTEST option with the WLSMV estimator\(^{2b}\) for comparisons in the gender stratified samples. Whenever a difference test indicated poorer fit of a more constrained model compared to a less constrained model, we used the modification indices provided by Mplus to free one or more parameters.

Results

Descriptive statistics

The two subgroups comprised 1772 respondents. The Surinam Dutch group ($n = 321$) had a significantly (Mann-Whitney test: $z = -5.79$, $p < 0.001$) higher total depression score ($Mdn = 2.00$, $IQR = 5.00$) compared to the Dutch group ($n = 1451$; $Mdn = 1.00$, $IQR = 3.00$).

The proportion of females was substantially higher in the Surinam Dutch group (72%) compared to the Dutch group (62%). Therefore, we had to reject the null hypothesis that in our sample gender and ethnicity are independent (continuity corrected $\chi^2 = 12.29$, $df = 1$, $p < 0.001$).

In both gender groups, participants in the Surinam Dutch group had a higher total depression score (male: $Mdn = 1.00$, $IQR = 4.00$; female: $Mdn = 2.00$, $IQR = 5.00$) compared to the participants in the Dutch group (male: $Mdn = 0.00$, $IQR = 2.00$; female: $Mdn = 1.00$, $IQR = 3.00$). Mann-Whitney tests revealed that in both groups, this difference was significant (male: $z = -2.86$, $p = 0.004$; female: $z = -4.83$, $p < 0.001$).

\(^{2b}\) With the WLSMV estimator, the estimation of the degrees of freedom is based on the empirical data instead of on model specification. As a result, a normal $\chi^2$ difference test (i.e., subtracting df of the more restrictive model from the df of the less restrictive model) is not possible because the degrees of freedom of both models are incomparable. The DIFFTEST option uses another method to calculate a $\chi^2$ difference and degrees of freedom but only the resulting $p$-value is interpretable.\(^2\)

Table 1. The four categorical multiple group single factor models.

<table>
<thead>
<tr>
<th>Models</th>
<th>Thresholds, loadings, and residual variances are freely estimated in both groups*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Baseline (non-invariance)</td>
<td>All loadings and thresholds are constrained to be equal across groups.</td>
</tr>
<tr>
<td>2. Invariant $\Lambda$</td>
<td>All loadings and thresholds are constrained to be equal across groups.</td>
</tr>
<tr>
<td>3. Invariant $\Lambda, \tau$</td>
<td>Loadings, thresholds, and residual variances are constrained to be equal across groups</td>
</tr>
</tbody>
</table>

*Except for a small subset that - for identification purposes - is constrained to be equal across groups.
Chapter 5

Unidimensionality assumption

Table 2 presents the tetrachoric correlations (i.e. the correlations between the hypothesized continuous variables underlying the dichotomous items) between the depression items. Those correlations were, in general, quite high and roughly of the same magnitude for all 9 items and resulted in a dominant first eigenvalue (6.52, second eigenvalue: 0.56), and was thus evidence for a single factor model structure. Consequently, based on a RMSEA of 0.041 and a CFI of 0.989, a categorical single factor model provided a good fit. However, based on the $\chi^2$, 105.85 with 27 degrees of freedom, a categorical single factor model provided an acceptable fit to the data of the entire sample. Combining the above results, it was justified to maintain the assumption that the data was unidimensional.

<table>
<thead>
<tr>
<th></th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
<th>d4</th>
<th>d5</th>
<th>d6</th>
<th>d7</th>
<th>d8</th>
<th>d9</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>loss of interest</td>
<td>1.00</td>
<td>0.89</td>
<td>0.68</td>
<td>0.85</td>
<td>0.71</td>
<td>0.75</td>
<td>0.75</td>
<td>0.67</td>
</tr>
<tr>
<td>d2</td>
<td>depressed mood</td>
<td>1.00</td>
<td>0.69</td>
<td>0.79</td>
<td>0.65</td>
<td>0.80</td>
<td>0.70</td>
<td>0.74</td>
<td>0.73</td>
</tr>
<tr>
<td>d3</td>
<td>sleep disturbances</td>
<td>1.00</td>
<td>0.71</td>
<td>0.64</td>
<td>0.62</td>
<td>0.58</td>
<td>0.59</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td>d4</td>
<td>fatigue</td>
<td>1.00</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
<td>0.70</td>
<td>0.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d5</td>
<td>appetite</td>
<td>1.00</td>
<td>0.68</td>
<td>0.62</td>
<td>0.58</td>
<td>0.59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d6</td>
<td>self-reproach</td>
<td>1.00</td>
<td>0.69</td>
<td>0.71</td>
<td>0.71</td>
<td>0.70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d7</td>
<td>concentration problems</td>
<td>1.00</td>
<td>0.76</td>
<td>0.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d8</td>
<td>psychomotor problems</td>
<td>1.00</td>
<td>0.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d9</td>
<td>(thoughts) of suicide</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Measurement invariance for ethnicity in the entire sample

We compared the fit of the four models in order to assess measurement invariance with respect to ethnicity in the entire sample. Table 3 presents the goodness-of-fit statistics and chi square difference tests. First, based on RMSEA and CFI, all models provided a good fit: all RMSEAs were well below 0.06 and all CFIs were well above 0.95. The $\chi^2$ values were between 2df and 3df and thus indicated an acceptable fit of all models. Second, and most important, the assumption of measurement invariance could not be rejected: all $\chi^2$ difference tests indicated that imposing more constraints - and thus assuming more measurement invariance - did not result in a poorer fit.

Measurement invariance for gender in the entire sample

Next, we compared the fit of the four models in order to assess measurement invariance with respect to gender in the entire sample. We found partial measurement invariance
for gender. The model with invariant loadings fit the data more poorly compared to the baseline model. Based on the modification indices, freeing the loadings for two items, ‘loss of interest’ and ‘sleep disturbances’, was sufficient to obtain restrictive models that did not fit the data more poorly than less restrictive models. In these models, the female group had a lower loading for ‘loss of interest’ and a higher loading for ‘sleep disturbances’ compared to the male group.

In order to disentangle the gender and ethnicity effect, we subsequently conducted the measurement invariance tests in both gender groups separately.

**Measurement invariance assessed in males and females separately**

Table 4 presents the goodness-of-fit statistics and $p$-values for the DIFFTEST for the four models testing measurement invariance with respect to ethnicity in the entire sample.

<table>
<thead>
<tr>
<th></th>
<th>$df$</th>
<th>$\chi^2$</th>
<th>RMSEA</th>
<th>CFI</th>
<th>$\chi^2$ difference test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Baseline</td>
<td>54</td>
<td>127.30</td>
<td>0.039</td>
<td>0.991</td>
<td></td>
</tr>
<tr>
<td>2. Invariant $\Lambda$</td>
<td>62</td>
<td>137.33</td>
<td>0.037</td>
<td>0.990</td>
<td>$\chi^2=10.03$, df=8, $p=0.2629$</td>
</tr>
<tr>
<td>3. Invariant $\Lambda$, $\tau$</td>
<td>67</td>
<td>138.97</td>
<td>0.035</td>
<td>0.991</td>
<td>$\chi^2=1.64$, df=5, $p=0.8964$</td>
</tr>
<tr>
<td>4. Invariant $\Lambda$, $\tau$, $\Theta$</td>
<td>70</td>
<td>142.59</td>
<td>0.034</td>
<td>0.991</td>
<td>$\chi^2=3.62$, df=3, $p=0.3055$</td>
</tr>
</tbody>
</table>

Baseline, non-invariance model; CFI, comparative fit index; df, degrees of freedom; RMSEA, root mean square error of approximation; $\chi^2$, chi-square statistic; $\chi^2$ difference test, chi-square for comparisons of model fit; $\Lambda$, loading; $\tau$, threshold; $\Theta$, residual variances.

Table 3. Goodness-of-fit statistics and chi square difference tests for four models testing measurement invariance with respect to ethnicity in the entire sample.

For males, the situation was somewhat different: model 2a, the invariant $\Lambda$ model provided a poorer fit to the data compared to the baseline model ($p$-value DIFFTEST < 0.05). Based on the modification indices, we freed the factor loading of item d8 (psychomotor problems). This resulted in model 2b that did not fit the data more poorly than the baseline model ($p$-value DIFFTEST = 0.0947). However, model 3a, the invariant $\Lambda$, $\tau$ model provided a poorer fit to the data compared to model 2b ($p$-value DIFFTEST < 0.05). Based on the modification indices, we freed the threshold of item d8. This resulted in model 3b that did not fit the data more poorly than model 2b ($p$-value DIFFTEST = 0.4442). Finally, the most restrictive model (except for the loading and threshold of item d8) did not fit the data more poorly than model 3b and thus, no
### Table 4. Goodness-of-fit statistics and p-values of the DIFFTEST for four models testing measurement invariance with respect to ethnicity in males (top panel) and females (bottom panel) separately

<table>
<thead>
<tr>
<th>Females of Dutch and Surinam Dutch ethnicity</th>
<th>df</th>
<th>$\chi^2$</th>
<th>RMSEA</th>
<th>CFI</th>
<th>p-value DIFFTEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Baseline</td>
<td>42</td>
<td>104.69</td>
<td>0.052</td>
<td>0.989</td>
<td></td>
</tr>
<tr>
<td>2. Invariant $\Lambda$</td>
<td>39</td>
<td>78.83</td>
<td>0.043</td>
<td>0.993</td>
<td>0.6273</td>
</tr>
<tr>
<td>3. Invariant $\Lambda$, $\tau$</td>
<td>42</td>
<td>80.83</td>
<td>0.041</td>
<td>0.993</td>
<td>0.6926</td>
</tr>
<tr>
<td>4. Invariant $\Lambda$, $\tau$, $\Theta$</td>
<td>43</td>
<td>76.67</td>
<td>0.038</td>
<td>0.994</td>
<td>0.8122</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Males of Dutch and Surinam Dutch ethnicity</th>
<th>df</th>
<th>$\chi^2$</th>
<th>RMSEA</th>
<th>CFI</th>
<th>p-value DIFFTEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Baseline</td>
<td>30</td>
<td>48.51</td>
<td>0.044</td>
<td>0.993</td>
<td></td>
</tr>
<tr>
<td>2a. Invariant $\Lambda$</td>
<td>27</td>
<td>56.53</td>
<td>0.059</td>
<td>0.989</td>
<td>0.0052</td>
</tr>
<tr>
<td>2b. Invariant $\Lambda$, $\lambda_{d8}$</td>
<td>27</td>
<td>45.33</td>
<td>0.046</td>
<td>0.993</td>
<td>0.0947</td>
</tr>
<tr>
<td>3a. Invariant $\Lambda$, $\lambda_{d8}$, $\tau$</td>
<td>29</td>
<td>53.66</td>
<td>0.052</td>
<td>0.991</td>
<td>0.0040</td>
</tr>
<tr>
<td>3b. Invariant $\Lambda$, $\lambda_{d8}$, $\tau$, $\tau_{d8}$</td>
<td>30</td>
<td>48.93</td>
<td>0.045</td>
<td>0.993</td>
<td>0.4442</td>
</tr>
<tr>
<td>4. Invariant $\Lambda$, $\lambda_{d8}$, $\tau$, $\tau_{d8}$, $\Theta$</td>
<td>29</td>
<td>45.65</td>
<td>0.043</td>
<td>0.994</td>
<td>0.5059</td>
</tr>
</tbody>
</table>

Baseline, non-invariance model; CFI, comparative fit index; df, degrees of freedom; RMSEA, root mean square error of approximation; $\chi^2$, chi-square statistic; $\Lambda$, loading; $\lambda_{d8}$, loading of item d8 (psychomotor problems) freely estimated in both groups; $\tau$, threshold; $\tau_{d8}$, threshold of item d8 (psychomotor problems) freely estimated in both groups; $\Theta$, residual variances.

modifications were needed. In this final model, the males in the Dutch group had a lower loading and threshold for item d8 ($\lambda_{d8} = 0.447$ and $\tau_{d8} = 2.072$) compared to the males in the Surinam Dutch group ($\lambda_{d8} = 2.468$ and $\tau_{d8} = 9.382$).\(^{3c}\)

In sum, the measurement invariance assumption with respect to ethnicity did not hold in males. There was partial measurement invariance since freeing the loading and threshold of only one item (d8) was sufficient to obtain restrictive models that did not fit the data more poorly than less restrictive models.

\(^{3c}\) Loadings non-invariant item in baseline model: $\lambda_{d8} = 0.467$ (Dutch) and $\lambda_{d8} = 2.075$ (Surinam Dutch). Thresholds non-invariant item in baseline model: $\tau_{d8} = 2.077$ (Dutch) and $\tau_{d8} = 7.310$ (Surinam Dutch).
Discussion

Our study addressed the question whether the PHQ-9 is measurement invariant for ethnicity (e.g. culturally insensitive) by comparing two contrasting cultural groups using a categorical single factor model.

Measurement invariance for ethnicity in the entire sample

Overall the PHQ-9 is measurement invariant for ethnicity across Dutch and Surinam Dutch participants. Hence, it appears that the higher total depression score of Surinam Dutch group is not the result of cultural bias of the instrument, but can be attributed to a true difference in the latent trait depression.

Measurement invariance assessed in males and females separately

In contrast with a recently published study, we did find a gender bias. Therefore we also conducted the measurement invariance tests for ethnicity in both gender groups separately. With regard to females, the PHQ-9 is measurement invariant with respect to ethnicity. Hence, the higher total depression score of females in the Surinam Dutch group can be attributed to a true difference in the latent trait depression and is not the result of cultural bias of the instrument.

With regard to males, however, the measurement invariance assumption did not hold, although there was evidence for partial measurement invariance. Males in the Surinam Dutch group had both a higher loading and a higher threshold for the item ‘psychomotor problems’ compared to males in the Dutch group. This means that for males in the Surinam Dutch group the item psychomotor problems was more causally influenced by the trait (depression). It also means that, given a certain level of depression, Surinam Dutch males are less likely to endorse this item than Dutch males (i.e. more of the continuous underlying trait is needed in order to endorse the item).

The fact that measurement invariance could not be confirmed for males cannot easily be attributed to a methodological artefact such as the smaller sample size for males, because in general, smaller sample sizes decrease the power of the tests, i.e. make it more difficult to find statistical significant differences. In other words it facilitates finding measurement invariance. This is caused by the fact that a smaller sample size results in less reliable estimates. Because the estimates are less reliable, the confidence intervals are necessarily larger; and as such, parameter estimates of one group more easily fall within such an interval of the other group, thereby facilitating finding measurement invariance.

Finding cultural bias of the item ‘psychomotor problems’ is in line with an earlier finding on the psychometric properties of the PHQ-9. Further, the item ‘psychomotor problems’ seems to be related to the presence of depression. It has been shown that this item
is rarely endorsed. However, when this item, in particular psychomotor retardation, is endorsed, there is a 90% chance that MDD would be present.\textsuperscript{32} In our study this phenomenon is observed in the Surinam Dutch males rather than in the Dutch males. Interestingly, in the reported study the majority of the participants was Caucasian. Future research is needed to further examine these findings.

**Strengths and limitations**

The strengths of our study lie in the statistical method used and the relatively large sample size. However, the study also has some limitations. We did not include a gold standard measure of depression, therefore we were unable to assess and compare the psychometric properties (sensitivity and specificity) of the PHQ-9 separately in the two groups. Addition of a gold standard would have resulted in a better understanding of the implications of variance on the diagnostic accuracy of the PHQ-9. Also, our sample is restricted to high risk primary care patients and the analyses pertain to a single cross cultural comparison. It remains unclear whether the results of this study can be generalized to clinical samples and unselected patient groups.

**Conclusion**

Our findings imply that, in relation to the two cultures assessed in this study, the PHQ-9 is measurement invariant for ethnicity in women and partially measurement invariant for ethnicity in men. The item ‘psychomotor problems’ seems to be culturally biased in the Surinam Dutch males. The observed higher total depression score for females in the Surinam Dutch group could be attributed to a true difference in the latent trait depression. For Surinam Dutch and Dutch men some caution is warranted when comparing results obtained with the PHQ-9. In the former group the scores may be biased slightly downward.

Future research is needed to examine how the item ‘psychomotor problems’ performs in different populations. These findings seem to support the idea that the concept of depression (e.g. the DSM-IV criteria for major depression) represents approximately the same in different cultural groups.

**Funding**

Funding for this study was provided by a grant from the Netherlands Organization for Health Research and Development (ZonMw), program Mental Health (# 100.003.005 and # 100.002.021).
REFERENCES


13. Waller NG, Thompson JS, Wenk E. Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: an illustration with the MMPI. Psychol Methods 2000; 5: 125-146.


Chapter 5


