Assessing impact of differential symptom functioning on post-traumatic stress disorder (PTSD) diagnosis

QIWEI HE, CEES A. W. GLAS & BERNARD P. VELDKAMP

Department of Research Methodology, Measurement and Data Analysis, University of Twente, Enschede, The Netherlands

Key words

post-traumatic stress disorder, item response theory, differential item functioning, NCS-R

Correspondence

Qiwei He, Department of Research Methodology, Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500AE Enschede, The Netherlands. Telephone (+31) 53-4892829
Fax (+31) 53-4894239
Email: q.he@utwente.nl

Received 3 July 2012;
revised 20 December 2012;
accepted 4 February 2013

Abstract

This article explores the generalizability of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) diagnostic criteria for post-traumatic stress disorder (PTSD) to various subpopulations. Besides identifying the differential symptom functioning (also referred to as differential item functioning [DIF]) related to various background variables such as gender, marital status and educational level, this study emphasizes the importance of evaluating the impact of DIF on population inferences as made in health surveys and clinical trials, and on the diagnosis of individual patients. Using a sample from the National Comorbidity Study-Replication (NCS-R), four symptoms for gender, one symptom for marital status, and three symptoms for educational level were significantly flagged as DIF, but their impact on diagnosis was fairly small. We conclude that the DSM-IV diagnostic criteria for PTSD do not produce substantially biased results in the investigated subpopulations, and there should be few reservations regarding their use. Further, although the impact of DIF (i.e. the influence of differential symptom functioning on diagnostic results) was found to be quite small in the current study, we recommend that diagnosticians always perform a DIF analysis of various subpopulations using the methodology presented here to ensure the diagnostic criteria is valid in their own studies. Copyright © 2014 John Wiley & Sons, Ltd.

Introduction

The diagnostic criteria for post-traumatic stress disorder (PTSD) have been revised several times since the debut in the 1980s (Spitzer *et al.*, 2007). The current Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV; APA, 2000) summarizes the PTSD diagnostic criteria in six clusters: (A) exposure to a traumatic event (A1–A2); (B) re-experience such as flashbacks and nightmares (B1–B5); (C) avoidance of stimuli associated with the trauma and numbing (C1–C7); (D)

increased arousal such as anger and hypervigilance (D1–D5); (E) duration of symptoms (more than one month); (F) significant impairment in social life. A person is diagnosed with PTSD when he meets both conditions within criterion A, has a specified number of symptoms from each of the three symptom clusters (for criterion B at least one symptom, for criterion C at least three symptoms, and for criterion D at least two symptoms), and meets the conditions of criteria E and F (for the complete list of PTSD criteria see the DSM-IV, APA, 2000).

Numerous studies have been conducted to validate the DSM-IV PTSD diagnostic criteria using either factor analytic methods (e.g. Blake *et al.*, 1995; Asmundson *et al.*, 2000; Andrews *et al.*, 2006) or item response theory (IRT; e.g. Betemps *et al.*, 2003; Conrad *et al.*, 2004; Palm *et al.*, 2009). Evidence shows that demographic factors (e.g. gender, age, education) could influence the generalizability of the DSM-IV in such a way that patients with different backgrounds may have differential symptoms (Gavranidou and Rosner, 2003). For instance, Peters *et al.* (2006) found that men were less likely than women to have particular symptoms (e.g. easily startled) that are considered by society to be a sign of weakness. To ensure the accuracy of cross-group comparisons and the appropriateness of the PTSD diagnosis, differential symptom functioning needs to be identified and its potential influence needs to be estimated.

Examination of differential item functioning (DIF) based on IRT (Rasch, 1960; Lord, 1980) is a method to address these concerns. DIF examines the measurement equivalence across groups; that is, it explores the relationship between item response and a group variable defined by demographic attributes (e.g. gender, race), conditional on a measure of an underlying construct such as PTSD severity (Teresi and Fleishman, 2007). In the context of PTSD assessment, an item (i.e. a symptom) is identified to exhibit DIF if respondents from different subgroups with the same degree of PTSD do not have the same probability of endorsing each response category (e.g. yes/no) of that item. For example, Palm *et al.* (2009) examined differential PTSD diagnostic symptoms between genders based on a national sample from the National Comorbidity Study – Replication (NCS-R; Kessler *et al.*, 2004). Of the 17 symptoms comprising symptom clusters B through D, seven symptoms were identified as having significant gender bias (i.e. marked as DIF items). The results suggested that among PTSD sufferers, women were more likely than men to report feeling emotionally distant and easily startled, whereas men were more likely than women to report having a lack of a plan for the future, unwanted memories, unpleasant dreams, and a short temper.

However, statistical significance is not the only important aspect of a DIF analysis; the subsequent consequences in terms of diagnosis are also of interest. That is, in addition to identifying specific items exhibiting DIF, it is also important to assess whether DIF impacts the diagnosis of PTSD. Little investigation in this area has been done in earlier studies. Therefore, the objective of this study is to further explore the generalizability of DSM-IV PTSD diagnostic criteria in two ways: (a) to identify differential symptom functioning by the background variables of

gender, marital status, and educational level; and (b) to introduce two methods for estimating the impact of DIF on PTSD diagnosis.

Methods

Sample

The data used in the present study were from the PTSD module in the NCS-R. The NCS-R, conducted between February 2001 and April 2003 in the United States, is a nationally representative community household survey of the prevalence and correlates of mental disorders (Kessler *et al.*, 2004). A total of 9282 interviews were completed in the main survey (Part I of the NCS-R). Questions regarding PTSD symptomology were included in Part II of the NCS-R interview, which assessed disorders that were either considered to be secondary importance or that were very time-consuming to assess (Kessler *et al.*, 2004). Part II was administered to only 5692 of the total respondents, over-sampling those with clinically significant psychopathology (Kessler *et al.*, 2004).

In the PTSD module of the interview, participants were asked about exposure to 27 different traumatic events (e.g. domestic violence, rape, life-threatening automobile accident). A total of 4984 endorsed at least one traumatic event and 1197 met criterion A (i.e. had experienced, witnessed or been confronted with a traumatic event and involved intense fear, helplessness, and/or horror) regarding the worst event. Of the 1197 participants, 317 were excluded because they endorsed neither of two filter items: “did you have any problems after the event like upsetting memories or dreams, feeling emotionally distant or depressed, trouble sleeping or concentrating, or feeling jumpy or easily startled?” and “did any of these reactions ever last for 30 days or longer?” PTSD symptoms were assessed by dichotomous items (i.e. “yes” = 1, “no” = 0). Each item corresponded to one symptom criterion listed in the DSM-IV. The present study focused on the 17 items regarding PTSD symptoms corresponding to clusters B through D. PTSD criteria requiring the highest thresholds were assessed first in the NCS-R, that is, symptoms were assessed in the following order: criterion C, B and D. The NCS-R implemented a “skip-out” principle: If the number of experienced symptoms for any criterion domain was below the threshold, e.g. the number of experienced symptoms for criterion C was smaller than three, the administration of items stopped. The respondents who were skipped out in the survey were directly diagnosed as non-PTSD. A total of 880 respondents received all items in criterion C regarding the worst event. Forty-one respondents did not continue in the assessment

because they did not meet the threshold in criterion C, and 65 respondents did not receive any items regarding symptoms in criterion D because they did not meet threshold of criterion B. Fourteen respondents stopped the assessment after giving all the negative responses to items in criterion D, and 88 respondents did not go further because they reported that none of these reactions in criterion C, B and D lasted over one month. Consequently, a total of 672 participants who completed the entire set of PTSD diagnostic items in the NCS-R were included in the present study. Of the 672 participants, 472 (70.2%) were diagnosed as PTSD and 200 (29.8%) diagnosed as non-PTSD. Table 1 gives a number of characteristics of the sample. The current sample had a mean age of 42.2 years (standard deviation [SD] = 13.97), and the majority of respondents was female (500, 74.4%), 156 (23.2%) participants had higher education (i.e. over 16-year education), 120 (18.2%) participants reported that they had never been married.

Dimensionality and model fit

The DIF analysis consisted of two parts: identifying DIF items and assessing the impact of DIF on PTSD diagnosis. Three background variables – gender, marital status, and educational level – were used for the analysis. All the results were obtained using the software program MIRT (Glas, 2010).

Responses to the 17 items of the NCS-R were used as input for the statistical analysis. The item parameters in an IRT model were estimated by marginal maximum

likelihood (MML; Bock and Aitkin, 1981). Given that the three symptom domains defined by the DSM-IV are used to index a general level of PTSD severity, we first considered a unidimensional two-parameter logistic (2PL) model underlying responses to the 17 symptoms (i.e. all 17 items on a single dimension). Next, given that the 17 symptoms are placed *a priori* into three separate criterion domains, we also considered a three-dimensional IRT model where each domain was associated with a separate dimension.

In the unidimensional 2PL model, that is, the probability of a score in category “yes” ($X_{ni} = 1$) of item i is given by the item response function

$$P(X_{ni} = 1|\theta) = \frac{\exp[\alpha_i(\theta_n - \beta_i)]}{1 + \exp[\alpha_i(\theta_n - \beta_i)]} \quad (1)$$

where θ_n is the latent PTSD level of person n , β_i is an item location parameter representing the severity level of each diagnostic symptom, and α_i is an item discrimination parameter indicating the extent to which the item response is related to the latent scale. A special version of the model is the Rasch model or one-parameter logistic (1PL) model (Rasch, 1960) where the item discrimination parameter α_i is simply fixed as one. The Rasch model is often used in clinical applications as well (e.g. Wong *et al.*, 2007; Elhai *et al.*, 2011).

In the multidimensional version of the 2PL model, the probability of a positive response depends on M latent variables, say $\theta_{n1}, \dots, \theta_{nm}, \dots, \theta_{nM}$. In the multidimensional case, in Equation 1, the product $\alpha_i\theta_n$ is replaced by $\sum_m \alpha_{im}\theta_{nm}$. Note that the discrimination parameters $\alpha_{i1}, \dots, \alpha_{im}, \dots, \alpha_{iM}$ can be viewed as factor loadings. In the present study, we used a between-items multidimensional model, where each item loaded on one of the three dimensions only.

Two approaches were used to investigate which model fitted the data better: a likelihood ratio-statistic and an item-oriented Lagrange multiplier (LM) test. First, the likelihood-ratio test of the 2PL model against the Rasch model yielded a value of the test statistic of $\chi^2 = 78.53$, $df = 16$, $p < 0.001$, while the multidimensional model against the unidimensional 2PL model yielded a value of $\chi^2 = 37.41$, $df = 3$, $p < 0.001$. It was concluded that the multidimensional model fit the data best and the 2PL fit the data significantly better than the Rasch model. However, though using a more complex model will generally result in better model fit, using a more parsimonious model might still lead to an adequate data description. To investigate this, a second approach was used. Under each model, item fit was evaluated using an LM item fit statistic (Glas, 1998, 1999). These statistics can be used

Table 1. Sample characteristics

Characteristics	Total	PTSD	NONPTSD
<i>N</i>	672	472	200
Gender			
Female	500	365	135
Male	172	107	65
Age (years)			
Mean	42.2	41.55	43.69
(SD)	(13.97)	(13.62)	(14.71)
Marriage status			
Married	351	243	108
Separated/widowed	201	142	59
Never married	120	87	33
Education level			
0–11 years	122	98	24
12 years	192	124	68
13–15 years	202	141	61
Over 16 years	156	109	47

to evaluate the fit of the expected item response function given by Equation 1 to the observed item responses.

Item fit was tested with a significance level of 0.01. For the Rasch model, the test was significant for six items, while no tests were significant for either the 2PL model or the multidimensional model. Further, the LM test statistic is accompanied by an effect size which measures the difference in observed and expected average item responses. For the 2PL model and the multidimensional model, these differences had the same magnitude. Hence, although a multidimensional IRT model fit the data better than 2PL in terms of the likelihood ratio test, it was not clearly superior in item fit. Therefore, the simpler unidimensional 2PL model was preferred over the more complicated multidimensional one.

Searching for DIF

Also DIF can be evaluated using an item-oriented LM item fit statistic (Glas, 1998, 1999). To compute the LM statistic, the sample of respondents is divided into subgroups labeled $g = 1, 2, \dots, G$. For instance, to examine DIF items related to gender, we divided the sample into two groups (i.e. male and female), and $G = 2$. The statistic is based on the difference between average observed scores on every item i in the subgroups, namely, $S_{ig} = \frac{1}{N_g} \sum_{n|g} X_{ni}$ (where the summation is over the N_g respondents in subgroup g), and their expectations $E(S_{ig})$. The differences are squared and divided by their covariance matrix (for more details refer to Glas, 1998, 1999; Glas and Falcon, 2003). The LM statistic has an asymptotic chi-square distribution with $G - 1$ degree of freedom. The statistics are accompanied by effect size $d_{ig} = \max_g |S_{ig} - E(S_{ig})|$ showing the degree of model violation. Since the effect size d_{ig} is on a scale ranging from zero to the maximum score m_i , effect size $d_{ig} < 0.10$, although somewhat arbitrary, is commonly suggested as an indicator of minor and acceptable model violation because of its effectiveness to detect a sufficient number of DIF items (van Groen *et al.*, 2010). In the present study, an item was identified as DIF when the LM statistic was significant ($p < 0.05$) or the effect size was above the cutoff point ($d_{ig} < 0.10$).

When items are flagged as DIF, one way to ensure comparability across subpopulations is to remove these items. However, removing items may decrease the reliability and potentially threaten the validity of the test. An alternative is to assign group-specific item parameters to the DIF items. In this approach, it is assumed that the same construct is measured in all subgroups, but for some subgroups the item locations on the latent scale or the discrimination parameters vary. This procedure was, for

instance, successfully applied by Weisscher *et al.* (2010); van Groen *et al.* (2010), and de Jong *et al.* (2009) to establish measurement equivalence in such areas as quality of life, physical disabilities, and marketing scales. An iterative procedure was conducted in this study, where the item with the largest significant LM statistic was given group-specific item parameters in each step. The iteration steps were repeated until no items were left with significant LM statistics (i.e. $p < 0.05$) or when the effect size was smaller than 0.10 (for more on the iteration procedure, refer to Glas and Verhelst, 1995).

Assessing the impact of DIF

The impact of DIF was evaluated by introducing group-specific parameters and comparing the results via two approaches, one approach that measures the differences between group distributions and the other that measures the change in accuracy of diagnosis of individuals.

Measurement on group distributions

In MML estimation, all subgroups (e.g. genders) have their own normal distribution of latent trait values. In the gender-related case, the latent scale is identified by setting the mean and variance of the latent distribution of female respondents equal to zero and one, respectively. A common method is evaluating whether the means of the distributions of male and female respondents are different. This boils down to testing the hypothesis whether the mean of the males differs from zero. The impact of DIF can be evaluated by estimating the parameters of the population distributions, with and without group-specific item parameters.

Measurement on diagnosis accuracy on individuals

The second method focuses on examining whether the introduction of group-specific item parameters can significantly enhance the accuracy of the diagnosis of individuals. Based on the diagnostic results from the NCS-R survey (i.e. diagnoses on the presence or absence of PTSD concerning DSM-IV criteria), the distribution of latent scores of PTSD and non-PTSD groups were obtained. A cutoff point on the latent scale was estimated to distinguish PTSD from non-PTSD.

We used three approaches to set the standard (i.e. obtain a cutoff point on the latent scale). The first approach entailed finding the midpoint between the medians of the two distributions (Cizek and Bunch, 2007). The medians of the distribution of PTSD and non-PTSD groups were 0.191 and -0.874 , respectively, and the midpoint was

−0.341. The second approach was the contrasting-groups method (Brandon, 2002), which uses logistic regression to determine the latent score point at which the probability of category membership is 50%. Setting the respondent status as a dichotomous variable coded 0 = non-PTSD and 1 = PTSD, we entered the latent scores of all the respondents into a general logistic regression equation; that is, $y^* = a + bx$, where y^* is the predicted value of the outcome variable (respondent status) for a respondent and x is the respondent's observed score. The intercept and slope parameters were estimated as $a = 2.113$ and $b = 3.200$, respectively. Given $y^* = 0.5$, the classification cutoff point for PTSD and non-PTSD groups was obtained at $x = -0.504$. The third approach used the Bayesian discrimination function, which minimizes expected risk. Using the zero-one loss function, the decision boundary becomes

$$g_i(x) = P(C_i|x) = \frac{P(C_i)p(x|C_i)}{p(x)} \quad (2)$$

where $P(C_i)$ is the prior probability (i.e. the prevalence of PTSD or non-PTSD in the total sample); $p(x|C_i)$ represents the class likelihood (we assumed the latent trait scores have a normal distribution); and $p(x)$ indicates the marginal probability of observation x . Given the assumption of normal distribution in both PTSD and non-PTSD groups, denoted as $PTSD \sim N(0.276, 0.731)$ and $NONPTSD \sim N(-0.874, 0.521)$ respectively, using Equation 2, we derived the cutoff point at -0.577 . Finally, we calculated the average of these three cutoff points, $[(-0.341) + (-0.504) + (-0.577)]/3 = -0.474$ as the cutoff point on the latent scale.

The cutoff diagnostic classification results on the new latent scale are analogous to the original inferences. The impact of DIF can be evaluated by estimating the diagnosis agreement, e.g. Cohen's Kappa (Cohen, 1960), between classification with and without group-specific item parameters. Kappa represents the level of agreement between two tests (raters) beyond that accounted for by chance alone. The Kappa coefficient is defined as

$$k = \frac{\Pr(o) - \Pr(e)}{1 - \Pr(e)} \quad (3)$$

where $\Pr(o)$ is the relative observed agreement among tests, and $\Pr(e)$ is the hypothetical probability of chance agreement. The Kappa coefficient is always less than or equal to one (i.e. 1 = perfect agreement). The universal guideline for using Kappa is that 0–0.20 is slight,

0.21–0.40 is fair, 0.41–0.60 is moderate, 0.61–0.80 is substantial, and 0.81–1 is almost perfect agreement (Landis and Koch, 1977).

Results

DIF related to gender

The DIF detection results for gender are shown in Table 2. The third and fourth columns give the item discrimination (α) and difficulty (β) parameters and the standard errors in estimation. The columns labeled LM and Prob give the value of the LM statistics and the significance probabilities of the LM test, respectively. The last four columns give the observed and expected average item scores for male and female respondents, respectively. Finally, the column labeled Group Dif gives the effect size averaged across the two gender groups. Four items/symptom criteria, corresponding to criteria B1, C3, C5, and C7, were identified as the DIF items because of significant LM statistics ($p < 0.05$). No items had an effect size larger than 0.10. The maximum effect size (0.07) was found in criterion C7 (i.e. sense of foreshortened future). It was noted that the majority of differential symptom functioning by gender occurred within symptom cluster C; that is, the greatest difference between the male and female PTSD respondents was in the area of avoidance and numbing.

Table 3 presents the estimates of gender-specific item parameters for DIF. The differences between group parameters suggested that female respondents were more likely to feel emotionally distant (C5) whereas male respondents were more likely to have a sense of a foreshortened future (C7) as well as repeated unwanted memories (B1). These findings were consistent with those of Palm *et al.* (2009). DIF was also found in item C3 ($\beta_{male} = 0.562, \beta_{female} = 0.177, DIF = 0.385, p < 0.05$); that is, female respondents were more likely to be unable to remember important aspects of the stressful event. It was also noticeable that the male group had a much higher discriminate parameter in C5 than the female group ($\alpha_{male} = 3.271, \alpha_{female} = 1.304, DIF = 1.967, p < 0.05$), which suggests that the symptom of feeling emotionally distant is more discriminative to distinguish PTSD and non-PTSD for males than females.

The results showing the impact of DIF regarding the population distributions are presented in Table 4. This table gives the estimates of mean and variance of the distribution of male respondents, without group-specific item parameters (row P0) and with group-specific item parameters (row P1) for the DIF items. Note that in row P0, the two-sided hypothesis that the estimated mean of male

Table 2. DIF analysis in gender using 2PL model

Item	Question in NCS-R	Item Parameters		LM	Prob	Group Dif	Male		Female	
		α (SE)	β (SE)				Obs	Exp	Obs	Exp
B1	Did you ever have repeated unwanted memories of the event, that is, you kept remembering it even when you didn't want to?	1.12 (0.21)	-2.32 (0.19)	5.94	0.01	0.04	0.82	0.80	0.70	0.76
B2	Did you ever have repeated unpleasant dreams about the event?	0.88 (0.15)	-0.91 (0.11)	2.84	0.09	0.03	0.69	0.64	0.67	0.68
B3	Did you have flashbacks, that is, suddenly act or feel as if the event were happening over again?	1.18 (0.16)	-0.74 (0.12)	0.78	0.38	0.02	0.56	0.58	0.65	0.64
B4	Did you get very upset when you were reminded of the event?	1.13 (0.20)	-1.69 (0.16)	3.19	0.07	0.02	0.88	0.85	0.86	0.87
B5	When you were reminded of the event, did you ever have physical reactions like sweating, your heart racing, or feeling shaky?	1.30 (0.18)	-0.89 (0.13)	1.01	0.31	0.02	0.57	0.60	0.67	0.66
C1	After the event, did you try not to think about it?	0.67 (0.17)	-1.59 (0.13)	0.01	0.91	0.00	0.79	0.79	0.81	0.81
C2	After the event, did you purposely stay away from places, people or activities that reminded you of it?	1.17 (0.16)	-0.88 (0.13)	0.41	0.52	0.01	0.63	0.61	0.66	0.66
C3	After the event, were you ever unable to remember some important parts of what happened?	0.70 (0.11)	0.30 (0.10)	7.65	0.01	0.05	0.31	0.39	0.46	0.43
C4	After the event, did you lose interest in doing things you used to enjoy?	1.41 (0.19)	-0.87 (0.14)	0.18	0.68	0.01	0.60	0.59	0.64	0.65
C5	After the event, did you feel emotionally distant or cutoff from other people?	1.57 (0.22)	-1.45 (0.17)	4.90	0.03	0.03	0.63	0.68	0.75	0.73
C6	After the event, did you have trouble feeling normal feelings like love, happiness, or warmth toward other people?	1.68 (0.22)	-1.16 (0.16)	0.16	0.69	0.01	0.63	0.62	0.68	0.68
C7	After the event, did you feel you had no reason to plan for the future because you thought it would be cut short?	1.31 (0.16)	0.79 (0.12)	16.63	0.00	0.07	0.39	0.29	0.32	0.35
D1	During the time this event affected you most, did you have trouble falling or staying asleep?	0.79 (0.19)	-1.86 (0.15)	1.58	0.21	0.02	0.85	0.84	0.79	0.82
D2	During the time this event affected you most, were you more irritable or short-tempered than you usually are?	0.91 (0.13)	-0.49 (0.11)	4.03	0.05	0.04	0.58	0.60	0.61	0.55
D3	During the time this event affected you most, did you have more trouble concentrating or keeping your mind on what you were doing?	1.13 (0.19)	-1.53 (0.15)	0.02	0.90	0.00	0.78	0.77	0.73	0.74
D4	During the time this event affected you most, were you much more alert or watchful, even when there was no real need to be?	0.90 (0.16)	-1.18 (0.13)	1.43	0.23	0.02	0.72	0.73	0.73	0.69
D5	During the time this event affected you most, were you more jumpy or easily startled by ordinary noises?	1.22 (0.19)	-0.99 (0.14)	2.31	0.13	0.03	0.70	0.68	0.59	0.63

Note. The item parameters were estimated from 2PL model without taking DIF into consideration. The columns LM and Prob give the value of the LM statistics and the significance probabilities of the LM test, respectively. Group Dif gives the absolute value of effect size averaged across the two gender groups. Obs and Exp are the observed and expected scores by the model, respectively.

Table 3. Estimation of item parameters for DIF items in gender

Item	Male				Female			
	α	SE (α)	β	SE (β)	α	SE (α)	β	SE (β)
B1	0.938	0.357	-2.337	0.260	1.248	0.255	-1.914	0.208
C3	1.301	0.369	0.562	0.226	0.550	0.121	0.177	0.102
C5	3.271	1.177	-0.998	0.736	1.304	0.214	-1.444	0.167
C7	1.612	0.398	0.102	0.232	1.474	0.197	1.050	0.149

Note. α indicates the item discrimination parameter, SE (α) is the standard error of α estimation. β indicates the item difficulty parameter, SE (β) is the standard error of β estimation.

Table 4. Estimation of gender (male) population parameters

Pattern	$\hat{\mu}$	SE($\hat{\mu}$)	σ	SE(σ)	Specified DIF items
P0	-0.261	0.107	0.891	0.094	
P1	-0.280	0.100	0.763	0.096	B1, C3, C5, C7

Note: The table shows the population parameters of male group. The female group is set as the reference group with mean and variance equal to zero and one, respectively. P0 is the baseline before the introduction of gender specific item parameters. P1 is the pattern with gender specific parameters for the DIF items defined in this study.

Table 5. Diagnostic confusion matrix between P0 and P1

		P0		Total
		PTSD	NONPTSD	
P1	PTSD	422	8	430
	NONPTSD	10	232	242
Total		432	240	672

Kappa = 0.942.

Note. P0 is the pattern before the introduction of gender specific item parameters. P1 is the pattern with gender specific parameters for the DIF items defined in this study.

respondents equals zero has to be rejected at 5% (the critical values are $\pm (0.107 \times 1.96) = \pm 0.210$. In row P1, the hypothesis also must be rejected; that is, the estimated mean of male respondents is not equal to the mean of female respondents (at zero). So the introduction of gender-specific item parameters demonstrates that DIF did not have an impact in this sample.

The implications of gender-related DIF on diagnostic results for individuals are presented in Table 5. Analogous to the previous analysis, we labeled the pattern without and with group-specific item parameters as P0 and P1, respectively. The Kappa coefficient of agreement was nearly perfect (0.942) between P0 and P1, suggesting that the introduction of group-specific item parameters had little impact on the diagnostic results.

DIF related to marital status

Only one item, B1 (i.e. repeated unwanted memories), was marked DIF in the analysis of marital status on the basis of the significant results of the LM test ($p < 0.05$). The item

characteristic curves (ICC) of item B1 with and without group-specific parameters are exhibited in Fig. 1. The horizontal axis indicates the latent score of individuals' PTSD level, while the vertical axis indicates the probability of individuals giving a positive response to the item. The solid line indicates the ICC without using the group-specific parameters, whereas the dashed line (with marriage experience) and dotted line (without marriage experience) indicate the ICCs with the group-specific parameters. Note that the item difficulty parameter differed moderately between those with and without marriage experience ($\beta_{mar} = -2.609, \beta_{nevermar} = -1.546, DIF = 1.063, p < 0.05$) when using the group-specific item parameters. It suggested that those who have been married (line veers to left of solid line) were more likely to have unwanted memories than those who have never been married (line veers to right of solid line).

Analogous to the process of estimating the impact of gender-related DIF, we identified the latent scale by setting the mean and variance of the latent distributions of the never-married group equal to zero and one, respectively. The two-sided hypothesis that the estimated means of

Figure 1. Item characteristic curves (ICCs) of item B1 with and without marital status group-specific parameters.

the married group ($\hat{\mu}_{Mar} = -0.156$, $SE(\hat{\mu}_{Mar}) = 0.133$) and separated/widowed group ($\hat{\mu}_{Sep} = -0.044$, $SE(\hat{\mu}_{Sep}) = 0.134$) are equal to zero was not rejected at 5%, regardless of whether or not the group-specific item parameters were used. Further, the Kappa coefficient for diagnosis agreement between tests with and without group-specific item parameters was 0.942, a nearly perfect agreement. Therefore, although we did detect the DIF criteria B1 by evaluating for marital status, the impact of DIF for this item on PTSD diagnosis was extremely small.

DIF related to educational level

Three items, C5 (i.e. feeling of emotionally distant), D2 (i.e. irritability) and D5 (i.e. exaggerated startle response) were detected as DIF related to educational background on the basis of the significant results of the LM tests ($p < 0.05$). The ICCs of these three items with and without group-specific parameters are plotted in Fig. 2. The results suggested that the lower-educated respondents identified as PTSD were more easily startled (D5) than the higher-educated respondents, whereas the higher-educated respondents were more likely to feel emotionally distant from others (C5) than the lower-educated patients. The respondents with 13–15 years of education were more likely to be irritable (D2) than the other groups.

The impact of DIF related to educational level was also assessed via two methods. With the implication regarding population distributions, we identified the latent scale by setting the mean and variance of the latent distributions of the group with over 16 year of education equal to zero and one, respectively. The two-sided tests of the hypothesis that the means of the other groups were equal to zero

were not rejected at 5%, regardless of whether or not the group-specific item parameters were used. Further, with respect to the implications for individuals, the Kappa coefficient of diagnosis agreement was more than 0.90 between tests with and without group-specific item parameters. Thus, we concluded that the impact of DIF related to educational background was negligible.

Discussion

This article explored DIF analyses for the background variables of gender, marital status, and educational level with respect to differential symptom functioning in persons with PTSD as determined by DSM-IV criteria. Besides detecting DIF items, this article emphasized the importance of evaluating the impact of DIF, and introduced two methods for estimating the impact of DIF on inferences regarding population distributions and diagnosis of individuals. The results showed that four symptoms for gender (B1, C3, C5, and C7), one symptom for marital status (B1), and three symptoms for educational level (C5, D2, and D5) were significantly identified as DIF. But the impact of DIF was marginal and had a negligible impact on diagnosis.

Current findings partly reconfirm the conclusions from the previous studies by Palm *et al.* (2009). In their study, only the comparison across gender was taken into account and seven symptoms including B1, B2, B3, C5, C7, D2, and D5 were detected as DIF. Three symptoms (B1, C5 and C7) were repetitively identified as DIF in our study. Taking a further look at Table 2, it was noticeable that although the item B2 and D2 were not significantly labeled as DIF in the present study, they were pretty close to the significance level 0.05. Consistent with the findings of

Figure 2. Item characteristic curves (ICCs) of C5, D2, and D5 with and without educational level group-specific parameters. Note: The educational background is divided into four levels: 1-11 years, 12 years, 13-15 years and 16 years. These four corresponding ICCs in each plot are exhibited with group-specific parameters. The red ICC in each plot is exhibited without using group-specific parameters.

Martin *et al.* (2007), the symptom B1 was detected bias between the patients with and without marriage experience, that is, patients with marriage experience (married or divorced) were more likely to repeat the unwanted memories than those who never got married. The probable reason was interpreted as patients who have ever married might have more chance of exposure to traumatic events, such as marital rape and domestic violence (Martin *et al.*, 2007). Dekel and Solomon (2006) further found that the PTSD patients with marriage experience exhibited decreased marital satisfaction, increased verbal aggression and heightened sexual dissatisfaction than those in single status. Educational background is also an important factor to be included in the study of PTSD. For instance, Shih *et al.* (2010) examined prevalence rates of PTSD in a sample of 677 individuals experiencing different types of trauma and found that people who reported having less than a high school education were more likely to develop PTSD than their more educated counterparts. However, the differential endorsement of PTSD symptoms across education groups was seldom investigated in the existing literatures. Additionally, we noted that symptom B1 was marked DIF for both gender and marital status and that symptom C5 was marked DIF for both gender and educational level. Further studies could be conducted to determine whether there are interaction effects for symptoms B1 and C5 in these corresponding demographic variables.

Whether the DIF makes a difference to diagnosis is very important to clinical practice. Teresi and Fleishman (2007) emphasized that besides identifying specific items exhibiting DIF, additional research was also needed to examine the impact of DIF in health-related measures. The methods for detecting DIF and assessing DIF impact presented in this article can be generally applied to a broader scope in psychiatric and clinical fields for checking the validity of diagnostic criteria and instruments. For instance, the methods could be used in the functional disability measurement, where differences of symptoms across demographic groups (e.g. age, gender) could be identified and the impact of DIF could be estimated. The results from the DIF analysis can play an important role in estimating the anticipating demands for services (Fleishman *et al.*, 2002).

References

- American Psychiatric Association (APA). (2000) Diagnostic and Statistical Manual of Mental Disorders: DSM-IV, Washington: DC, APA.
- Andrews L., Joseph S., Shevlin M., Troop N. (2006) Confirmatory factor analysis of posttraumatic stress symptoms in emergency personnel: an examination of seven alternative models. *Personality and Individual Differences*, **41**(2), 213–224, DOI: 10.1016/j.paid.2005.11.034
- Asmundson G.J.G., Frombach I., McQuaid J., Pedrelli P., Lenox R., Stein M.B. (2000) Dimensionality of posttraumatic stress symptoms: a confirmatory factor analysis of DSM-IV symptom clusters and other symptom

The major difference between the two methods for assessing DIF impact is the measurement level, one focuses on a group level and the other on an individual level. The first approach to assessing impact is in terms of effect size. That is, an analysis is conducted to examine how much mean group differences in total score distributions change with and without using specific group item parameters. This method is simple and straightforward to see whether the group means have significant changes in different patterns, but rather difficult to describe the DIF impact on individuals. The second approach focuses on an individual level, that is, to compare the accuracy rate with and without using group specific item parameters. It is a bit more complicated than the first one, since the optimal cutoff points need to be estimated on a latent scale. In the daily practices, we recommend to use both of these two methods to make a thorough investigation.

Another issue that is worth concern is the prerequisites of using the methods. The LM test statistics are closely related to the analysis of cross tabulations (Glas and Falcon, 2003). For instance, the LM statistic targetted at DIF splits up the sample of respondents into G groups, and the count in each cell, that is, for each group, should be at least five. However, such low frequencies do not result in an acceptable power to detect item misfit. Glas and Falcon (2003) suggested that a sample size of 500 respondents was required to obtain the acceptable power.

In summary, on the one hand we conclude that the DSM-IV PTSD diagnostic criteria do not produce substantially biased results in the three investigated subpopulations, and therefore there should be few reservations regarding their use. On the other hand, we recommend that diagnosticians always perform a DIF analysis of various subpopulations using the methodology presented here to ensure that the diagnostic criteria is valid to use for the target population.

Acknowledgments

The authors thank Inter-university Consortium for Political and Social Research for providing the dataset on the NCS platform. The PTSD diagnosis research project was supported and sponsored by Stichting Achmea Slachtofferhulp Samenleving, the Netherlands.

Declaration of interest statement

The authors have no competing interests.

- models. *Behaviour Research and Therapy*, **38**(2), 203–214, DOI: 10.1016/S0005-7967(99)00061-3
- Betemps E.J., Smith R.M., Baker D.G., Rounds-Kugler B. A. (2003) Measurement precision of the clinician administered PTSD scale (CAPS): a RASCH model analysis. *Journal of Applied Measurement*, **4**(1), 59–69.
- Blake D.D., Weathers F.W., Nagy L.M., Kaloupek D.G., Gusman F.D., Charney D.S., Keane T. M. (1995) The development of a clinician-administered PTSD scale. *Journal of Traumatic Stress*, **8**(1), 75–90, DOI: 10.1007/BF02105408
- Bock R.D., Aitkin M. (1981) Marginal maximum likelihood estimation of item parameters: an application of an EM-algorithm. *Psychometrika*, **46**(4), 443–459, DOI: 10.1007/BF02293801
- Brandon P.R. (2002) Two versions of the contrasting-groups standard-setting method: a review. *Measurement and Evaluation in Counseling and Development*, **35**(3), 167–181.
- Cizek G.J., Bunch M.B. (2007) Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests, Thousand Oaks: CA, Sage.
- Cohen J. (1960) A coefficient of agreement for nominal scale. *Educational and Psychological Measurement*, **20**, 37–46, DOI: 10.1177/001316446002000104
- Conrad K.J., Wright B.D., McKnight P., McFall M., Fontana A., Rosenheck R. (2004) Comparing traditional and Rasch analyses of the Mississippi PTSD scale: revealing limitations of reverse-scored items. *Journal of Applied Measurement*, **5**(1), 15–30.
- Dekel R., Solomon Z. (2006) Marital relations among former prisoners of war: contribution of posttraumatic stress disorder, aggression, and sexual satisfaction. *Journal of Family Psychology*, **20**(4), 709–712, DOI: 10.1037/0893-3200.20.4.709
- Elhai J.D., de Francisco Carvalho L., Miguel F.K., Palmieri P.A., Primi R., Frueh B.C. (2011) Testing whether posttraumatic stress disorder and major depressive disorder are similar or unique constructs. *Journal of Anxiety Disorders*, **25**, 404–410, DOI: 10.1016/j.janxdis.2010.11.003
- Fleishman J.A., Spector W.D., Altman B.M. (2002) Impact of differential item functioning on age and gender differences in functional disability. *Journal of Gerontology: Social Sciences*, **57B**(5), 275–284, DOI: 10.1093/geronb/57.5.S275
- Gavranidou M., Rosner R. (2003) The weaker sex? Gender and post-traumatic stress disorder. *Depression and Anxiety*, **17**(3), 130–139, DOI: 10.1002/da.10103
- Glas C.A.W. (1998) Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica*, **8**(3), 647–667.
- Glas C.A.W. (1999) Modification indices for the 2-PL and the nominal response model. *Psychometrika*, **64**(3), 273–294, DOI: 10.1007/BF02294296
- Glas C.A.W. (2010) MIRT: Multidimensional Item Response Theory (computer software). <http://www.utwente.nl/gw/omd/afdeling/Glas>
- Glas C.A.W., Falcon J.C.S. (2003) A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, **27**(2), 87–106, DOI: 10.1177/0146621602250530
- Glas C.A.W., Verhelst N.D. (1995) Testing the Rasch model. In Fischer G.H., Molenaar I.W. (eds) *Rasch Models: Their Foundations, Recent Developments and Applications*, pp. 69–96, New York: Springer.
- van Groen M.M., ten Klooster P.M., Taal E., van de Laar M.A.F.J., Glas C.A.W. (2010) Application of the health assessment questionnaire disability index to various rheumatic diseases. *Quality of Life Research*, **19**(9), 1255–1263, DOI: 10.1007/s11136-010-9690-9
- de Jong M.G., Steenkamp J.B.E.M., Veldkamp B. P. (2009) A model for the construction of country-specific yet internationally comparable short-form marketing scales. *Marketing Science*, **28**, 674–689, DOI: 10.1287/mksc.1080.0439
- Kessler R.C., Berglund P., Chiu W.T., Demler O., Heeringa S., Hiripi E., Jin R., Pennell B.E., Walters E.E., Zaslavsky A., Zheng H. (2004) The US National Comorbidity Survey Replication (NCS-R) design and field procedures. *International Journal of Methods in Psychiatric Research*, **13**(2), 69–92, DOI: 10.1002/mpr.167
- Landis J.R., Koch G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174, DOI: 10.2307/2529310
- Lord FM (1980) *Applications of Item Response Theory to Practical Testing Problems*, NJ, Erlbaum: Hillsdale.
- Martin E.K., Taft C.T., Resick P.A. (2007) A review of marital rape. *Aggression and Violent Behavior*, **12**(3), 329–347, DOI: 10.1016/j.avb.2006.10.003
- Palm K.M., Strong D.R., MacPherson L. (2009) Evaluating symptom expression as a function of a posttraumatic stress disorder severity. *Journal of Anxiety Disorders*, **23**(1), 27–37, DOI: 10.1016/j.janxdis.2008.03.012
- Peters L., Issakidis C., Slade T., Andrews G. (2006) Gender differences in the prevalence of DSM-IV and ICD-10 PTSD. *Psychological Medicine*, **36**(1), 81–89, DOI: 10.1017/S003329170500591X
- Rasch G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhagen: Danish Institute for Educational Research.
- Shih R.A., Schell T.L., Hambarsoomian K., Belzberg H., Marshall G.N. (2010) Prevalence of posttraumatic stress disorder and major depression after trauma center hospitalization. *Journal of Trauma: Injury, Infection and Critical Care*, **69**(6), 1560–1566, DOI: 10.1097/TA.0b013e3181e59c05
- Spitzer R.L., First M.B., Wakefield J.C. (2007) Saving PTSD from itself in DSM-V. *Journal of Anxiety Disorders*, **21**(2), 233–241, DOI: 10.1016/j.janxdis.2006.09.006
- Teresi J., Fleishman J. (2007) Differential item functioning and health assessment. *Quality of Life Research*, **16**(1), 33–42, DOI: 10.1007/s11136-007-9184-6
- Weisscher N., Glas C.A.W., Vermeulen M., de Haan R.J. (2010) The use of an item response theory-based disability item bank across diseases: accounting for differential item functioning. *Journal of Clinical Epidemiology*, **63**(5), 543–549, DOI: 10.1016/j.jclinepi.2009.07.016
- Wong E., Ungvari G.S., Leung S.K., Tang W.K. (2007) Rating catatonia in patients with chronic schizophrenia: Rasch analysis of the Bush–Francis Catatonia Rating Scale. *International Journal of Methods in Psychiatric Research*, **16**(3), 161–170, DOI: 10.1002/mpr.224