Psychological Medicine (2009), **39**, 591–601. © Cambridge University Press 2008
doi:10.1017/S0033291708003875 Printed in the United Kingdom

Use of item response theory to understand differential functioning of DSM-IV major depression symptoms by race, ethnicity and gender

L. A. Uebelacker*, D. Strong, L. M. Weinstock and I. W. Miller

Warren Alpert Medical School of Brown University and Butler Hospital, Providence, RI, USA

Background. Psychological literature and clinical lore suggest that there may be systematic differences in how various demographic groups experience depressive symptoms, particularly somatic symptoms. The aim of the current study was to use methods based on item response theory (IRT) to examine whether, when equating for levels of depression symptom severity, there are demographic differences in the likelihood of reporting DSM-IV depression symptoms.

Method. We conducted a secondary analysis of a subset ($n = 13\,753$) of the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) dataset, which includes a large epidemiological sample of English-speaking Americans. We compared data from women and men, Hispanics and non-Hispanic Whites, African Americans and Whites, Asian Americans and Whites, and American Indians and Whites.

Results. There were few differences overall, although the differences that we did find were primarily limited to somatic symptoms, and particularly appetite and weight disturbance.

Conclusions. For the most part, individuals responded similarly to the criteria used to diagnose major depression across gender and across English-speaking racial and ethnic groups in the USA.

Received 26 September 2007; Revised 15 May 2008; Accepted 17 May 2008; First published online 30 June 2008

Key words: Major depression, item response theory.

Introduction

Major depression is a heterogeneous disorder, with different individuals exhibiting different symptom profiles. However, whether there are systematic differences in how various demographic groups experience depression remains an open question. If differences exist, understanding what they are will help to ensure that the construct of depression is appropriately understood and assessed in various groups. Furthermore, there is some evidence that antidepressant treatment response differs by racial group (Lesser *et al*. 2007). If there are differential symptom patterns by demographic group, these symptom patterns should be investigated as a possible explanation for the differential treatment response.

Both clinicians and researchers have focused on gender and racial or ethnic differences in the degree to which individuals report somatic symptoms of depression, namely appetite/weight disturbance, sleep disturbance, fatigue and psychomotor symptoms (agitation or retardation). Some have suggested that women are more likely than men to report somatic symptoms, possibly because of biological or hormonal differences (e.g. Wenzel *et al*. 2005) and/or differences in social roles or cultural norms (e.g. Silverstein & Lynch, 1998). Epidemiological data (Silverstein, 1999, 2002) suggest that, in comparison to men, women have a higher prevalence of 'somatic depression'. Among clinical samples with depression, women are more likely than men to endorse somatic symptoms (Young *et al*. 1990; Wenzel *et al*. 2005), However, not all studies detect gender differences in somatic symptom reports (Santor *et al*. 1994; Salokangas *et al*. 2002) and consistent patterns of gender differences in somatic symptoms have not emerged (e.g. Khan *et al*. 2002).

In advising clinicians to adopt a contextual view when diagnosing depression, DSM-IV-TR (APA, 2000) notes that 'culture can influence experience and communication of symptoms of depression' and that 'in some cultures, depression may be experienced largely in somatic terms, rather than with sadness or guilt' (p. 353). Recommendations for diagnosing depression in Hispanics (Lewis-Fernandez *et al*. 2005) state that

* Address for correspondence: Dr L. A. Uebelacker, Butler Hospital, 345 Blackstone Boulevard, Providence, RI 02906, USA.
(Email: Lisa_Uebelacker@brown.edu)

somatic presentations may be particularly common in this group. Similarly, a recent article suggested that 'when evaluating African Americans for depression, look for somatic and neurovegetative symptoms rather than mood or cognitive symptoms' (Das *et al.* 2006). Thus, clinicians are instructed to attend differentially to somatic symptoms when evaluating depression severity among African Americans and Hispanics *versus* Whites.

A literature review reveals only limited evidence that particular racial or ethnic groups are more likely than others to report somatic symptoms. Some reports suggest that African Americans (Brown *et al.* 1996; Ayalon & Young, 2003), Hispanics (Myers *et al.* 2002), Chinese Americans (Huang *et al.* 2006) and American Indians (Iwata & Buka, 2002) are more likely to endorse somatic symptoms than their White counterparts. However, these findings have not been robust, and other investigators have failed to find consistent differences in reports of somatic symptoms between depressed Whites and depressed African Americans (Blazer *et al.* 1998; Cole *et al.* 2000), Hispanics (Gallo *et al.* 1998; Iwata *et al.* 2002) or Chinese Americans (Yen *et al.* 2000).

In summary, the existing literature is mixed as to whether there are gender, racial or ethnic differences in the likelihood of reporting somatic symptoms of depression. Furthermore, there are several limitations to the existent literature, including the fact that some studies have relied upon samples of convenience (e.g. Brown *et al.* 1996; Wenzel *et al.* 2005), and a large number analyze self-report inventories such as the Center for Epidemiologic Studies Depression Scale (CES-D; Blazer *et al.* 1998; Cole *et al.* 2000; Yen *et al.* 2000; Iwata *et al.* 2002) or the original Beck Depression Inventory (BDI; Santor *et al.* 1994; Salokangas *et al.* 2002; Ayalon & Young, 2003). These self-report inventories do not assess all DSM-IV depressive symptoms, and therefore provide an incomplete evaluation of DSM diagnostic criteria. A prominent limitation of this literature is that, with some exceptions (Santor *et al.* 1994; Gallo *et al.* 1998; Cole *et al.* 2000; Iwata & Buka, 2002; Iwata *et al.* 2002; Ayalon & Young, 2003), researchers do not use techniques that adequately take into account level of depression when assessing frequency of somatic symptoms. It is often unclear whether one group is more likely to endorse a particular symptom simply because that group is more likely to be depressed.

Methods based in item response theory (IRT; cf. Lord, 1980) provide significant improvements on previous techniques (e.g. simply comparing frequency counts of particular symptoms in groups of interest) as IRT approaches can be used to examine the likelihood that a particular symptom will be reported given a particular level of depression severity. Application of IRT methods is emerging in the evaluation of DSM-IV diagnostic criteria, including criteria for depression (Aggen *et al.* 2005; Simon & Von Korff, 2006), alcohol dependence (Kahler *et al.* 2003) and bulimia (Rowe *et al.* 2002). Only one previous study has examined DSM symptoms of depression to determine whether symptoms functioned differently between groups (Simon & Von Korff, 2006). These authors focused on depressed individuals with and without a co-morbid medical condition, and the few differences that they did find were clinically modest.

The aim of the current study was to use IRT methods to examine whether, when equating for levels of depression symptom severity, there are gender, race or ethnic differences in the likelihood of endorsing DSM-IV depression symptoms. Given previous research and clinical advice, we tested a 'somatic hypothesis': that women (*versus* men), Latinos or Hispanics (*versus* non-Hispanic Whites), Blacks or African Americans (*versus* Whites), Asian Americans (*versus* Whites) and American Indians or Alaskan Natives (*versus* Whites) would be more likely to report somatic symptoms. We examined other DSM-IV symptoms to contextualize our results regarding somatic symptoms. To conduct these analyses, we used a large, non-treatment-seeking, epidemiological sample of the US population.

Method

Data collection

The National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) is a survey of a nationally representative sample from the USA. Methods for obtaining the sample have been detailed elsewhere (Grant *et al.* 2004*a*, *b*). The NESARC involved interviews from 43 093 adults aged ⩾18 years. Only those respondents who reported either depressed mood ($n=12785$) or anhedonia ($n=10533$) in their lifetime completed the section of the NESARC survey assessing lifetime occurrence of all DSM-IV symptoms of major depressive disorder (MDD). The present analysis consists of those individuals ($n=13753$, or 32% of the total sample). This subsample had a mean age of 45.87 years (S.D. = 17.16, range = 18–98) and included 65.7% women ($n=9040$). Individuals chose their racial group from the following categories: American Indian or Alaska Native ($n=468$), Asian ($n=291$), Black or African American ($n=2396$), Native Hawaiian or other Pacific Islander ($n=101$) and White ($n=10958$). Because of the small sample size and lack of any specific hypotheses, we did not conduct analyses comparing the Native Hawaiian or other Pacific

https://doi.org/10.1017/S0033291708003875 Published online by Cambridge University Press

Table 1. *Initial symptom parameters in the entire subsample* ($n = 13\,753$)

DSM-IV MDD symptoms	Frequency (%)	Severity parameter	S.E.	Discrimination parameter	S.E.
3. Appetite/weight disturbance	65	−0.62	0.02	1.35	0.03
4. Sleep disturbance	74	−0.78	0.02	2.34	0.06
5. Psychomotor symptoms	49	0.03	0.01	1.88	0.05
6. Fatigue	62	−0.40	0.02	1.84	0.05
7. Worthlessness/guilt	55	−0.15	0.01	2.01	0.05
8. Concentration	69	−0.64	0.02	2.23	0.06
9. Suicide	43	0.31	0.02	1.22	0.03

Islander group to Whites. Individuals were also asked whether they were Latino or Hispanic, and 2258 responded affirmatively. The percentage of women in each racial group was as follows: Hispanic 65%; African American 71%; Asian 64%; American Indian or Alaskan Native 63%; and White 65%. The only group in which gender composition differed significantly from the White group was the African American group ($\chi^2 = 32.2$, $p < 0.001$). With regard to education, 83.2% of the sample had completed high school or its equivalent, and 55.3% had completed college.

Symptoms of MDD

The Alcohol Use Disorders and Associated Disabilities Interview Schedule (AUDADIS; Grant *et al*. 2001, 2003*a*) was used to assess DSM-IV MDD criteria. Experienced interviewers received extensive training in this fully structured interview and used computer-assisted software to decrease error in measurement (Grant *et al*. 2004*b*). Developers of the AUDADIS-IV also made considerable efforts to ensure that questions were comprehensible for lay persons (Grant *et al*. 2003*a*). NESARC estimates of lifetime and 12-month prevalence of MDD were 13.2% and 5.3% respectively (Hasin *et al*. 2005). These estimates are comparable to those found in the National Comorbidity Survey (Kessler *et al*. 2005*a*, *b*). Test–retest reliability for the MDD diagnosis was good in this sample (Grant *et al*. 2003*a*).

As described earlier, analyses for this study were by necessity limited to respondents who reported a 2-week period of depressed mood and/or anhedonia in their lifetime. Analyses focused on the seven MDD symptoms that could be present specifically within the context of a 2-week episode of depressed mood or anhedonia. We present these seven MDD symptoms in Table 1.

Analyses

Item response modeling allows us to establish whether symptoms of depression index levels of depression severity similarly across subgroups. A two-parameter model involves estimating the following for each symptom (or item): a severity parameter to describe the point on the latent continuum where a symptom becomes likely to be observed (e.g. >50%), and a discrimination parameter to describe how rapidly the probability of observing the symptom changes across increasing levels of the latent continuum [i.e. the slope of the item response function (IRF)]. In a one-parameter model, the severity parameters are estimated for each symptom (item), whereas the discrimination parameters are constrained to be equivalent for all items.

Unidimensionality assumption

The primary assumption of unidimensional item response models is that responses to symptom queries are a function of individual variation along a single underlying dimension. We tested this assumption in the sample as a whole and in subsamples with Mplus (Muthen & Muthen, 1998–2007) using confirmatory factor analyses of tetrachoric correlations with the robust weighted least squares method of parameter estimation.

Parametric item response model selection

Given the previous support for unidimensionality and the utility of fitting parametric models to the symptoms of MDD, we evaluated both one- and two-parameter parametric models using marginal maximum likelihood estimation methods. We compared the fit of these models using a likelihood ratio test (LRT) that involves subtracting the log-likelihood values (LL) for the models being compared. We used MULTILOG (Thissen, 1991) for model-fitting analyses.

Differential item functioning (*DIF*)

We used model-based assessment of DIF for these analyses (Thissen *et al*. 1993). Following Thissen *et al*. (1993), we used an LRT statistic to provide a significance test for the null hypothesis that the item

https://doi.org/10.1017/S0033291708003875 Published online by Cambridge University Press

parameters do not differ between two groups (e.g. women and men).

We used version 2.0 of IRTLRDIF (Thissen, 2001) to complete DIF analyses. IRTLRDIF automatically accommodates group differences with respect to the latent trait. IRTLRDIF sets the scale of item parameters using the population distribution for the reference group. With the reference group mean set to zero and standard deviation set to 1, the estimated focal group mean reflects a standardized difference from the reference group and the standard deviation reflects the ratio of the focal and reference group standard deviations (Thissen, 2001).

The IRTLRDIF approach was implemented without a set of anchor items. Although it is possible to use iterative analyses to isolate a set of 'DIF-free' items for use as an anchor, with seven examined items we were concerned about a potentially small number of anchor items and the complexity of describing analyses that may have different sets of anchor items across the planned comparisons in this study. The LRT applied to each item is conditional on all equal item parameters, or no DIF, for all of the other items in the test. Analyses proceeded by initially constraining both the discrimination and the severity estimates to be equal for the two subgroups across all seven symptoms (Model A). For each of the seven symptoms, a model was then fit that constrained all of the remaining symptoms' discrimination and severity estimates to be equal (i.e. all remaining items were used as an anchor) but allowed these estimates for one symptom to differ across the two groups (Model B). The difference in the log-likelihoods of Model A and Model B [$G^2 = -2(\mathrm{LL}_{\mathrm{Model\ A}} - \mathrm{LL}_{\mathrm{Model\ B}})$] provided an omnibus test (df = 2) of whether there was DIF for the discrimination and/or severity estimate for this particular symptom. If significant, follow-up tests (1 df) were conducted to identify whether DIF was present in discrimination or severity estimates by further constraining models. In conducting DIF analyses that involved 1 df tests, we controlled for the statistical risk of making false conclusions by using the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995; Thissen *et al*. 2002).

Given the large sample sizes involved, relatively small differences in severity between groups could be significant statistically. *A priori*, we decided that a difference $\geqslant 0.25$ in item severity was a clinically meaningful difference. According to Steinberg & Thissen (2006), a difference of 0.25 can be interpreted as one-quarter of the 'standard unit difference between the values of the [underlying] trait necessary to have a 50–50 chance of responding positively in one group compared to another' (pp. 405–406). This may be considered to be a small effect size (Cohen, 1988). For example, a DIF of 0.25 for a given item severity would mean that, depending on the values of the discrimination parameters as well as how close the actual group severity parameters are to 0, differences in group proportions endorsing a given item could range from 2% to 8% (for discrimination parameters ranging from 0.50 to 2.00).

With respect to discrimination parameters, Steinberg & Thissen (2006) suggest that the best way to determine whether a statistically significant discrimination parameter is also clinically significant is by visual inspection of the respective IRFs.

Results

Unidimensionality assumption

We conducted confirmatory factor analyses to test the assumption of the unidimensionality of depression symptoms. Fit statistics for the sample as a whole indicated a reasonable fit to the data [$\chi^2 = 642.0$, Comparative Fit Index (CFI) = 0.978, Tucker–Lewis Index (TLI) = 0.979, root mean square error of approximation (RMSEA) = 0.059]. Fit statistics for the subgroups were also adequate, and were in range 0.969–0.992 for the CFI, 0.971–0.992 for the TLI, and 0.043–0.078 for the RMSEA. We determined that these fit statistics were sufficient to proceed to fitting IRT models.

Parametric item response model selection

We fit a one- and a two-parameter model to the seven symptoms. Evaluation of the log likelihoods suggested that allowing the discrimination parameter to vary in the two-parameter model provided better fit to the data ($G^2 = 2240$, df = 9, $p < 0.001$). Table 1 lists the severity and discrimination parameter estimates for each of the seven symptoms.

DIF

Tables 2–4 list the severity and discrimination parameter estimates for each symptom across all group comparisons.

Women and men

Two of the seven symptoms, appetite/weight disturbance and fatigue, exceeded our criteria for both clinical and statistical significance in DIF for the severity parameter (see Table 2). The DIF in appetite/weight disturbance was non-uniform; that is, we also found statistically significant DIF in the discrimination parameter for this item. However, inspection of the graphs revealed that, given the same level of depression severity, women consistently tended to be

https://doi.org/10.1017/S0033291708003875 Published online by Cambridge University Press

Table 2. *Differential item functioning of DSM-IV MDD symptoms for women* ($n=9040$) *and men* ($n=4713$)

	G^2	Severity parameter (b)			Discrimination parameter (a)			Women	
DSM-IV MDD symptoms	(df=2)	Men	Women	Difference	Men	Women	Difference	Mean	s.d.
3. Appetite/weight disturbance	166.8*	−0.18	−0.56	**−0.38***	1.27	1.46	0.19*	0.24	0.94
4. Sleep disturbance	4.4*	−0.55	−0.61	−0.06*	2.41	2.39	−0.02	0.25	0.94
5. Psychomotor symptoms	79.0*	0.06	0.28	0.22*	2.03	1.94	−0.09	0.27	0.95
6. Fatigue	94.5*	−0.06	−0.31	**−0.25***	1.88	1.94	0.06	0.24	0.94
7. Worthlessness/guilt	24.6*	0.35	0.47	0.12*	2.01	2.11	0.10	0.26	0.94
8. Concentration	26.5*	−0.54	−0.39	0.15*	2.03	2.38	0.35*	0.26	0.93
9. Suicide	42.1*	0.35	0.51	0.16*	1.19	1.40	0.21*	0.26	0.94

MDD, Major depressive disorder; df, degrees of freedom; s.d., standard deviation.

The G^2 test with 2 df evaluates differences between groups in both severity and discrimination parameters. Differences between groups on either parameter are evaluated using 1 df tests. p values for 1 df tests were adjusted using the Benjamini–Hochberg procedure. Severity parameters that (a) represent a statistically significant difference between groups and (b) exceed our effect size criteria (0.25) are shown in bold.

* $p<0.05$.

Table 3. *Differential item functioning of DSM-IV MDD symptoms for Hispanics* ($n=2258$) versus *non-Hispanic Whites* ($n=8885$)

	G^2	Severity parameter (b)			Discrimination parameter (a)			Hispanic	
DSM-IV MDD symptoms	(df=2)	White	Hispanic	Difference	White	Hispanic	Difference	Mean	s.d.
3. Appetite/weight disturbance	46.0*	−0.56	−0.89	**−0.33***	1.33	1.28	−0.05	−0.05	1.11
4. Sleep disturbance	9.5*	−0.82	−0.95	−0.13	2.34	1.96	−0.38*	−0.04	1.13
5. Psychomotor symptoms	23.6*	0.07	−0.08	−0.15*	1.69	2.05	0.36*	−0.06	1.07
6. Fatigue	41.7*	−0.51	−0.29	0.22*	1.83	1.73	−0.10	0.00	1.10
7. Worthlessness/guilt	2.4	0.23	0.24	0.01	1.89	2.09	0.20	−0.03	1.10
8. Concentration	20.1*	−0.74	−0.60	0.14*	2.11	2.00	−0.11	−0.01	1.10
9. Suicide	0.6	0.28	0.27	−0.01	1.22	1.17	−0.05	−0.03	1.11

MDD, Major depressive disorder; df, degrees of freedom; s.d., standard deviation.

The G^2 test with 2 df evaluates differences between groups in both severity and discrimination parameters. Differences between groups on either parameter are evaluated using 1 df tests. p values for 1 df tests were adjusted using the Benjamini–Hochberg procedure. Severity parameters that (a) represent a statistically significant difference between groups and (b) exceed our effect size criteria (0.25) are shown in bold.

* $p<0.05$.

more likely to endorse appetite/weight disturbance than men (Fig. 1). Given equivalent levels of depression severity, women also tended to be more likely to endorse fatigue. There were two other statistically significant differences in discrimination parameters: concentration difficulties and suicide were more discriminating among women than among men. The clinical significance of the discrimination parameters is discussed later.

Hispanics and non-Hispanic Whites

Only one symptom, appetite/weight disturbance, was identified as having DIF in the severity parameter according to criteria for clinical and statistical significance. Hispanic respondents tended to be more likely to endorse appetite/weight disturbance than non-Hispanic respondents given the same level of depression severity. There were two statistically significant differences in discrimination parameters: sleep disturbance was more discriminating among non-Hispanic than Hispanic respondents, and fatigue was more discriminating among Hispanic respondents.

African Americans and Whites

One of the seven symptoms, namely appetite/weight disturbance, met criteria for statistically and clinically

Table 4. *Differential item functioning of DSM-IV MDD symptoms by racial group*

		Severity parameter (*b*)			Discrimination parameter (*a*)			AA, Asian or AI/AN	
DSM-IV MDD symptoms	G^2 (df = 2)	White	AA, Asian, AI/AN	Difference	White	AA, Asian, AI/AN	Difference	Mean	S.D.
African American (AA) (*n* = 2396) and White (*n* = 10 895)									
3. Appetite/weight disturbance	37.4*	−0.60	−0.85	**−0.25*[a]**	1.33	1.39	0.06	−0.14	1.09
4. Sleep disturbance	2.5	−0.81	−0.87	−0.06	2.33	2.14	−0.19	−0.13	1.1
5. Psychomotor symptoms	55.2*	0.05	−0.18	−0.23*[b]	1.78	2.23	0.45*[c]	−0.16	1.04
6. Fatigue	32.3*	−0.47	−0.28	0.19*[a]	1.85	1.72	−0.13	−0.10	1.09
7. Worthlessness/guilt	12.7*	0.23	0.36	0.13*	1.96	1.91	−0.05[c]	−0.10	1.10
8. Concentration	6.5*	−0.69	−0.61	0.08*	2.12	2.10	−0.02	−0.11	1.08
9. Suicide	1.2	0.28	0.32	0.04	1.23	1.27	0.04	−0.12	1.08
Asian (*n* = 291) and White (*n* = 10 920)									
3. Appetite/weight disturbance	0.2	−0.61	−0.66	−0.05	1.33	1.32	−0.01	−0.18	1.12
4. Sleep disturbance	5.6*	−0.81	−0.67	0.14*	2.34	1.97	−0.37	−0.14	1.11
5. Psychomotor symptoms	1.4	0.05	0.02	−0.03	1.78	2.19	0.41	−0.18	1.09
6. Fatigue	2.8	−0.46	−0.36	0.10	1.86	2.37	0.51	−0.17	1.09
7. Worthlessness/guilt	2.3	0.23	0.25	0.02	1.95	1.51	−0.44	−0.17	1.15
8. Concentration	7.4*	−0.69	−0.85	−0.16	2.13	1.36	−0.77*	−0.18	1.19
9. Suicide	7.5*	0.28	−0.04	**−0.32***	1.23	1.41	0.18	−0.2	1.09
American Indian/Alaskan Native (AI/AN) (*n* = 468) and White (*n* = 10 703)									
3. Appetite/weight disturbance	3.6	−0.60	−0.69	−0.09	1.33	1.51	0.18	0.18	1.04
4. Sleep disturbance	0.2	−0.81	−0.79	0.02	2.33	2.27	−0.06	0.2	1.04
5. Psychomotor symptoms	1.7	0.06	−0.04	−0.10	1.78	1.78	0	0.18	1.04
6. Fatigue	9.9*	−0.45	−0.43	0.02	1.87	1.31	−0.56*	0.23	1.09
7. Worthlessness/guilt	2.4	0.24	0.25	0.01	1.94	2.45	0.51	0.19	1.02
8. Concentration	6.4*	−0.69	−0.67	0.02	2.14	1.63	−0.51	0.22	1.07
9. Suicide	4.3*	0.29	0.12	−0.17	1.23	1.37	0.14	0.17	1.02

MDD, Major depressive disorder; df, degrees of freedom; S.D., standard deviation.

The G^2 test with 2 df evaluates differences between groups in both severity and discrimination parameters. Differences between groups on either parameter are evaluated using 1 df tests. *p* values for these 1 df tests were adjusted using the Benjamini–Hochberg procedure. Severity parameters that (*a*) represent a statistically significant difference between groups and (*b*) exceed our effect size criteria (0.25) are shown in bold.

[a] Subgroup analyses demonstrated a clinically and statistically significant difference for African American men *versus* White men, but only a statistically significant difference for African American women *versus* White women.

[b] Subgroup analyses demonstrated a clinically and statistically significant difference for African American women *versus* White women, but only a statistically significant difference for African American men *versus* White men.

[c] Subgroup analyses demonstrated a statistically significant difference for African American women *versus* White women, and failed to find a statistically significant difference for African American men *versus* White men.

* $p < 0.05$.

significant DIF for the severity parameter. Given equivalent levels of depression severity, African Americans tended to be more likely to endorse appetite/weight disturbance than White respondents. One symptom met criteria for statistically significant DIF on the discrimination parameter. Psychomotor symptoms was more discriminating for African Americans than Whites.

Because there were different proportions of women in the African American and White groups, we next compared African American to White women, and African American to White men. This allowed us to examine whether the different gender proportions in these samples might account for the racial differences. For the most part, findings were similar to the larger group analyses, with no clinically and statistically significant differences in sleep disturbance, worthlessness/guilt, concentration, or suicide severity parameters for either men or women. With only two exceptions, there were no differences in any

https://doi.org/10.1017/S0033291708003875 Published online by Cambridge University Press

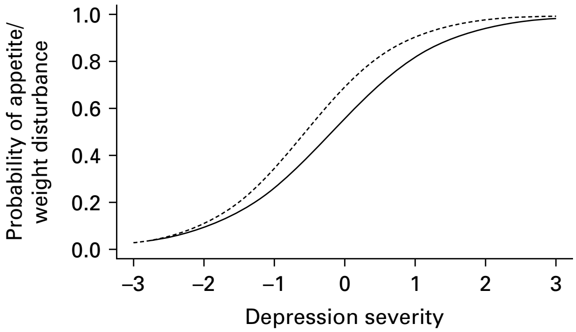

Fig. 1. Differences between men (——) and women (- - -) in the probability of endorsing appetite/weight disturbance across levels of depression severity.

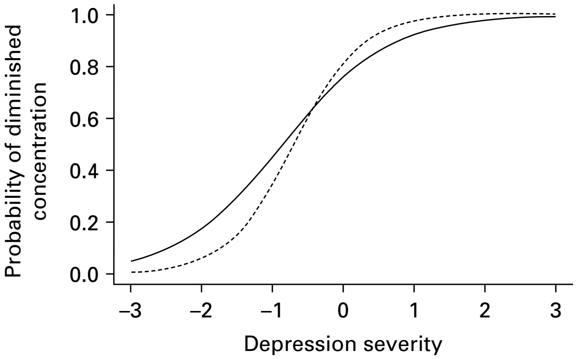

Fig. 2. Differences between Asians (——) and Whites (- - -) in the probability of endorsing concentration difficulties across levels of depression severity.

discrimination parameters, for either men or women. The two exceptions were worthlessness/guilt and psychomotor symptoms for the comparison between African American and White women.

Three symptoms requiring further explanation. First, when we examined appetite/weight disturbance, we found statistically and clinically significant differences in the severity parameter for men ($b_{\text{white}} = -0.17$, $b_{\text{AA}} = -0.51$, $b_{\text{dif}} = -0.34$, $p < 0.05$) but not for women ($b_{\text{white}} = -0.85$, $b_{\text{AA}} = -1.05$, $b_{\text{dif}} = -0.20$, $p < 0.05$). That is, African American men tended to be more likely to endorse appetite/weight disturbance than White men, given equivalent depression severity. This suggests that the greater preponderance of women in the African American sample may partially (but only partially) account for African American and White differences in severity of the appetite/weight disturbance. Second, we found statistically and clinically significant severity differences in psychomotor symptoms for women ($b_{\text{white}} = 0.04$, $b_{\text{AA}} = -0.22$, $b_{\text{dif}} = -0.26$, $p < 0.05$) but not for men ($b_{\text{white}} = 0.07$, $b_{\text{AA}} = -0.13$, $b_{\text{dif}} = -0.20$, $p < 0.05$). Although DIF was non-uniform for the analyses in women (i.e. there was also statistically significant DIF in the discrimination parameter), inspection of the graphs revealed that, given equivalent levels of depression severity, this symptom consistently tended to be more frequently endorsed by African American than White women. Finally, there were statistically and clinically significant differences in the severity parameter for fatigue for men ($b_{\text{white}} = -0.15$, $b_{\text{AA}} = 0.17$, $b_{\text{dif}} = 0.32$, $p < 0.05$) but not for women ($b_{\text{white}} = -0.66$, $b_{\text{AA}} = -0.49$, $b_{\text{dif}} = 0.17$, $p < 0.05$). Given equivalent levels of depression severity, White men were more likely to endorse this symptom than African American men. This suggests that the greater preponderance of women among African Americans may serve to decrease racial group differences in fatigue observed in DIF comparison that included both genders (as the gender analysis revealed that women experience more fatigue at a lower level of depression severity).

Asians and Whites

One of the seven symptoms, namely suicide, met criteria for statistically and clinically significant DIF for the severity parameter. Given equivalent levels of depression severity, Asian Americans were more likely to endorse suicidal ideation than White respondents. Only one symptom met criteria for statistically significant DIF in discrimination: concentration was more discriminating for Whites than Asian Americans.

American Indians/Alaska Natives and Whites

No items met statistical and clinical criteria for DIF in severity parameter. Only one item, namely fatigue, met criteria for statistically significant DIF in discrimination. Fatigue was more discriminating for White respondents than American Indian respondents.

Discrimination parameters

To determine the clinical significance of differences in discrimination parameters, we inspected IRFs in all cases in which there were statistically significant differences in discrimination. Visual inspection suggested that it was unlikely that any of the differences would have a large clinical impact. For illustration purposes, we present two IRFs. Fig. 1 shows the IRFs for appetite/weight disturbance by gender; this was the only example of non-uniform DIF in a statistically and clinically significant severity parameter. In Fig. 2, we illustrate the DIF analysis that yielded the largest difference in discrimination parameters: the comparison

https://doi.org/10.1017/S0033291708003875 Published online by Cambridge University Press

of Asians and Whites on the symptom of concentration.

Discussion

The aim of the current study was to evaluate gender, race and ethnic differences in the likelihood of reporting DSM-IV major depression symptoms using an IRT-based methodology. The benefit of the IRT-based approach is that it accounts for the potential confounding effect of depression severity in evaluating group differences. Additional strengths of the current study include the use of a large representative community sample of non-treatment-seeking individuals, and the use of an *a priori* defined threshold of clinical significance so that emphasis was placed on differences that were both statistically significant and substantively meaningful.

We examined three symptoms considered to be somatic symptoms by all authors (appetite/weight disturbance, sleep disturbance, and fatigue), one symptom considered to be somatic by some authors (psychomotor symptoms), and three symptoms that are not somatic symptoms (concentration, worthlessness, and suicidality). Most of the clinically and statistically differences in item severity parameters were found among somatic symptoms. There were differences in appetite/weight disturbance in the expected direction, across three groups (women *versus* men, Hispanics *versus* non-Hispanics, and African American *versus* White men). We found two group differences in fatigue, one in the expected direction (in the gender comparison) and one in a direction opposite from what a 'somatic' hypothesis would predict (in the comparison of African American *versus* White men). We failed to find group differences in the likelihood of endorsing sleep disturbance, given equivalent levels of depression severity. We found one difference in the likelihood of endorsing psychomotor symptoms (in the comparison of African American and White women), in the expected direction. Finally, we found one difference in suicide severity, with Asian Americans being more likely to endorse suicide than Whites at equivalent depression severity.

Overall, the results failed to find support for the idea that one group, and particularly one racial or ethnic group, is more likely to 'somatize' than another group. Of 28 possible group comparisons for four somatic symptoms (sleep disturbance, appetite disturbance, fatigue, and psychomotor symptoms), we found only five differences supporting a somatic hypothesis and one difference in the opposite direction. The group comparison that yielded the most differences in likelihood of endorsing somatic symptoms was the gender comparison, in which two of four comparisons were clinically and statistically significant (i.e. appetite/weight disturbance and fatigue). However, another difference in somatic symptoms that was statistically reliable (but not clinically significant), the difference in psychomotor symptoms, in fact showed that men were more likely to endorse this symptom. Therefore, we do not believe it informative or useful to make generalizations about 'somatization' as a unitary construct.

The only symptom for which we consistently noted differences in the expected direction was appetite/weight disturbance. Although we can only speculate, gender differences could relate to differential social pressures and expectations around weight and body image, perhaps best reflected in the striking gender difference in prevalence of eating disorders in the USA (APA, 2000). For example, endorsement of appetite/weight disturbance may occur at lower levels of depression severity for women *versus* men because women may pay more attention or may be more sensitive to appetite symptoms. Alternatively, there may be biological or hormonal explanations for these gender differences. Given the heterogeneity inherent in racial and ethnic categories, reasons for differences in appetite symptoms are also speculative. It is possible that, to the extent that food is a symbol of one's racial or cultural affiliation (Airhihenbuwa *et al*. 1996) and represents a way of expressing interpersonal connections for particular racial or ethnic groups (Ahye *et al*. 2005), a higher prevalence of decreased appetite may occur among African Americans and Hispanics *versus* Whites because this symptom is reflective of other depression-related problems, such as a loss of cultural identity or isolation.

Despite this discussion of differences that emerged from current study analyses, it is important to emphasize again that, overall, our study results fail to provide support for a broad somatic hypothesis. Rather, as Kirmayer *et al*. (1993) have argued, it may be that somatization is relevant in *many* groups. For example, in a Canadian sample, it has been shown that up to 80% of depressed primary care patients initially present with somatic symptoms (Kirmayer *et al*. 1993). We note that *beliefs* about the tendency to somaticize typically apply to lower-status groups, that is women or racial/ethnic minorities. In commenting on a similar belief that it is 'non-Western' cultures that tend to somaticize distress, Kirmayer (2001) suggested that we should instead ask whether Westerners are prone to 'psychologization', that is the tendency to express distress in cognitive or affective terms.

The Kirmayer *et al*. (1993) study described above raised another key issue in understanding differences between clinical impressions and the results from the current study. In the Kirmayer study, most patients

https://doi.org/10.1017/S0033291708003875 Published online by Cambridge University Press

presented for treatment with somatic symptoms. However, when queried directly, patients endorsed other depression symptoms and often cited a psychosocial cause for the symptoms. Therefore, clinical impressions of group differences in symptom reporting may reflect how depressed individuals initially present to clinicians, but not their actual experience of depression.

In evaluating the study findings, it is important to consider the limitations of this research. First, racial groups represent broad categorizations that reflect substantial within-group heterogeneity. We were unable to evaluate, and indeed may have missed, differential patterns of symptom functioning that relate to specific cultural groups. These classification systems also leave no clear way to categorize people who are multi-racial. Second, because the AUDADIS was conducted only in English, we cannot generalize our findings to individuals who primarily speak other languages. Third, to the extent that we detected DIF between groups, we cannot determine whether differences were due to differential experience of a symptom itself, or to measurement-related differences in symptom reporting. That is, how the question was worded may have a differential impact on how various groups respond. We do note that the rankings of item severity that were found in our entire subsample ($n=13\,753$) were largely similar to those reported by Aggen *et al*. (2005), even though different interviews were used. However, our results need to be replicated in other large samples using different assessment instruments.

Fourth, we acknowledge that the size of the Asian American group was small relative to the other groups in this study. However, because there are relatively few studies examining symptom patterns in Asian Americans empirically, we made the decision to include this group *a priori* despite potential limitations in identifying group differences due to an underrepresentation of a full range of depression severity. Fifth, although we did not weight our model estimates, the incorporation of complex sampling information is becoming increasingly available in software that is also capable of fitting models based in IRT. By allowing the model to incorporate information from the sampling methods of the survey design, it has been suggested that standard error estimates and tests of model fit can be improved (Asparouhov, 2005).

Finally, we note that, to be included in the data analysis, individuals had to report either sad mood or anhedonia. If a respondent did not endorse either, the remaining symptom item questions were not asked. It could argued that there are depressed individuals who would not endorse either sad mood or anhedonia, and that it is exactly these individuals who are more likely to endorse somatic symptoms. However, if a respondent does not endorse one of the two 'core' symptoms of DSM-IV depression, it cannot be argued that other symptoms that they may endorse (e.g. insomnia, concentration difficulties) are indicators of a mild depression as opposed to being related to another problem [e.g. a health problem or attention deficit hyperactivity disorder (ADHD)]. A limitation of this paper is that a certain level of depressed mood or anhedonia must be endorsed (namely, depressed mood most of the time for at least 2 weeks or anhedonia for at least 2 weeks) to be included in the data analyses. Therefore, we may miss some participants with lower levels of depressed mood who should be considered to be part of the depressive spectrum. We acknowledge that the results could be different in a sample that included individuals with lower levels of depression severity in the analyses. However, we note that 32% of the total NESARC sample endorsed depressed mood or anhedonia sometime in their life and were therefore included in our analyses. We were also reassured by the similarity of our results to those of Aggen *et al*. (2005). In their community sample, all participants reported on all symptoms, regardless of whether they endorsed sad mood or anhedonia.

These results have both clinical and research implications. The only symptom that consistently showed differences in prevalence between groups is appetite/weight disturbance. Clinicians may want to differentially weigh endorsement of this particular symptom when determining whether an individual from a particular group is clinically depressed. For example, women who tend to endorse this symptom may not be as depressed as men who tend to endorse this symptom. In general, however, we failed to find support for a broad somatization hypothesis. This line of research suggests that we need to evaluate critically the circumstances in which we advise clinicians to expect and look for increased levels of all somatization symptoms in particular gender, racial or ethnic groups.

Acknowledgments

The NESARC was funded by the National Institute on Alcohol Abuse and Alcoholism with supplemental funding from the National Institute on Drug Abuse. Work on this manuscript was supported by grant MH067779 from the National Institute of Mental Health to L.A.U.

Declaration of Interest

None.

https://doi.org/10.1017/S0033291708003875 Published online by Cambridge University Press

References

Aggen SH, Neale MC, Kendler KS (2005). DSM criteria for major depression: evaluating symptom patterns using latent trait item response models. *Psychological Medicine* **35**, 475–487.

Ahye BA, Devome CM, Odoms-Young AM (2005). Values expressed through intergenerational family food and nutrition management systems among African American women. *Family and Community Health* **29**, 5–16.

Airhihenbuwa CO, Kumanyika S, Agurs TD, Lowe A, Saunders D, Morssink CB (1996). Cultural aspects of African American eating patterns. *Ethnicity and Health* **1**, 245–260.

APA (2000). *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn, text revision. American Psychiatric Association: Washington, DC.

Asparouhov T (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling* **12**, 411–434.

Ayalon L, Young MA (2003). A comparison of depressive symptoms in African Americans and Caucasian Americans. *Journal of Cross-Cultural Psychology* **34**, 111–124.

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

Blazer DG, Landerman LR, Hays JC, Simonsick EM, Saunders WB (1998). Symptoms of depression among community-dwelling elderly African-American and white older adults. *Psychological Medicine* **28**, 1311–1320.

Brown C, Schulberg HC, Madonia MJ (1996). Clinical presentations of major depression by African Americans and whites in primary medical care practice. *Journal of Affective Disorders* **41**, 181–191.

Cohen J (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates: Hillsdale, NJ.

Cole SR, Kawachi I, Maller SJ, Berkman LF (2000). Test of item-response bias in the CES-D scale: experience from the New Haven EPESE study. *Journal of Clinical Epidemiology* **53**, 285–289.

Das AK, Olfson M, McCurtis HL, Weissman MM (2006). Depression in African Americans: breaking barriers to detection and treatment. *Journal of Family Practice* **55**, 30–39.

Gallo JJ, Cooper-Patrick L, Lesikar S (1998). Depressive symptoms of whites and African Americans aged 60 years and older. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* **53**, P277–P286.

Grant BF, Dawson DA, Hasin DS (2001). *The Alcohol Use Disorders and Associated Disabilities Interview Schedule – DSM-IV Version (AUDADIS-IV)*. National Institute on Alcohol Abuse and Alcoholism: Bethesda, MD.

Grant BF, Dawson DA, Stinson FS, Chou PS, Kay W, Pickering R (2003*a*). The Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV (AUDADIS-IV): reliability of alcohol consumption, tobacco use, family history of depression and psychiatric diagnostic modules in a general population sample. *Drug and Alcohol Dependence* **71**, 7–16.

Grant BF, Dawson DA, Stinson FS, Chou SP, Dufour MC, Pickering RP (2004*a*). The 12-month prevalence and trends in DSM-IV alcohol abuse and dependence: United States, 1991–1992 and 2001–2002. *Drug and Alcohol Dependence* **74**, 223–234.

Grant BF, Moore TC, Shepard J, Kaplan K (2003*b*). *Source and Accuracy Statement: Wave 1 National Epidemiologic Survey on Alcohol and Related Conditions (NESARC)*. National Institute on Alcohol Abuse and Alcoholism: Bethesda, MD.

Grant BF, Stinson FS, Dawson DA, Chou SP, Dufour MC, Compton W, Pickering RP, Kaplan K (2004*b*). Prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders: results from the National Epidemiologic Survey on Alcohol and Related Conditions. *Archives of General Psychiatry* **61**, 807–816.

Hasin DS, Goodwin RD, Stinson FS, Grant BF (2005). Epidemiology of major depressive disorder: results from the National Epidemiologic Survey on Alcoholism and Related Conditions. *Archives of General Psychiatry* **62**, 1097–1106.

Huang FY, Chung H, Kroenke K, Delucchi KL, Spitzer RL (2006). Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. *Journal of General Internal Medicine* **21**, 547–552.

Iwata N, Buka S (2002). Race/ethnicity and depressive symptoms: a cross-cultural/ethnic comparison among university students in East Asia, North and South America. *Social Science and Medicine* **55**, 2243–2252.

Iwata N, Turner RJ, Lloyd DA (2002). Race/ethnicity and depressive symptoms in community-dwelling young adults: a differential item functioning analysis. *Psychiatry Research* **110**, 281–289.

Kahler CW, Strong DR, Stuart GL, Moore TM, Ramsey SE (2003). Item functioning of the alcohol dependence scale in a high-risk sample. *Drug and Alcohol Dependence* **72**, 183–192.

Kessler RC, Berglund P, Demler O, Jin R, Walters EE (2005*a*). Lifetime prevalence and age-of-onset distribution of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry* **62**, 593–602.

Kessler RC, Chiu WT, Demler O, Walters EE (2005*b*). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry* **62**, 617–627.

Khan AA, Gardner CO, Prescott CA, Kendler KS (2002). Gender differences in the symptoms of major depression in opposite-sex dizygotic twin pairs. *American Journal of Psychiatry* **159**, 1427–1429.

Kirmayer LJ (2001). Cultural variations in the clinical presentation of depression and anxiety: implications for diagnosis and treatment. *Journal of Clinical Psychiatry* **62** (Suppl. 13), 22–28.

Kirmayer LJ, Robbins JM, Dworkind M, Yaffe MJ (1993). Somatization and the recognition of depression and anxiety in primary care. *American Journal of Psychiatry* **150**, 734–741.

Lesser IM, Castro DB, Gaynes BN, Gonzalez J, Rush AJ, Alpert JE, Trivedi M, Luther JF, Wisniewski SR (2007). Ethnicity/race and outcome in the treatment of

https://doi.org/10.1017/S0033291708003875 Published online by Cambridge University Press

depression: results from STAR*D. *Medical Care* **45**, 1043–1051.

Lewis-Fernandez R, Das AK, Alfonso C, Weissman MM, Olfson M (2005). Depression in US Hispanics: diagnostic and management considerations in family practice. *Journal of the American Board of Family Practice* **18**, 282–296.

Lord FM (1980). *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum: Hillsdale, NJ.

Muthen LK, Muthen BO (1998–2007). *Mplus User's Guide*, 5th edn. Muthen & Muthen: Los Angeles, CA.

Myers HF, Lesser I, Rodriguez N, Mira CB, Hwang WC, Camp C, Anderson D, Erickson L, Wohl M (2002). Ethnic differences in clinical presentation of depression in adult women. *Cultural Diversity and Ethnic Minority Psychology* **8**, 138–156.

Rowe R, Pickles A, Simonoff E, Bulik CM, Silberg JL (2002). Bulimic symptoms in the Virginia Twin Study of Adolescent Behavioral Development: correlates, comorbidity, and genetics. *Biological Psychiatry* **51**, 172–182.

Salokangas RK, Vaahtera K, Pacriev S, Sohlman B, Lehtinen V (2002). Gender differences in depressive symptoms. An artifact caused by measurement instruments? *Journal of Affective Disorders* **68**, 215–220.

Santor DA, Ramsay JO, Zuroff DC (1994). Nonparametric item analyses of the Beck Depression Inventory: evaluating gender item bias and response option weights. *Psychological Assessment* **6**, 255–270.

Silverstein B (1999). Gender difference in the prevalence of clinical depression: the role played by depression associated with somatic symptoms. *American Journal of Psychiatry* **156**, 480–482.

Silverstein B (2002). Gender differences in the prevalence of somatic versus pure depression: a replication. *American Journal of Psychiatry* **159**, 1051–1052.

Silverstein B, Lynch AD (1998). Gender differences in depression: the role played by paternal attitudes of male superiority and maternal modeling of gender-related limitations. *Sex Roles* **38**, 539–555.

Simon GE, Von Korff M (2006). Medical comorbidity and validity of DSM-IV depression criteria. *Psychological Medicine* **36**, 27–36.

Steinberg L, Thissen D (2006). Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychological Methods* **11**, 402–415.

Thissen D (1991). *MULTILOG User's Guide: Version 6*. Scientific Software, Inc.: Chicago, IL.

Thissen D (2001). IRTLRDIF v.2.0b: Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning (www.unc.edu/~dthissen/dl.html).

Thissen D, Steinberg L, Kuang D (2002). Quick and easy implementation of the Benjamini–Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics* **27**, 77–83.

Thissen D, Steinberg L, Wainer H (1993). Detection of differential item functioning using the parameters of item response models. In *Differential Item Functioning* (ed. P. W. Holland and H. Wainer), pp. 67–113. Lawrence Erlbaum Associates: Hillsdale, NJ.

Wenzel A, Steer RA, Beck AT (2005). Are there any gender differences in frequency of self-reported somatic symptoms of depression? *Journal of Affective Disorders* **89**, 177–181.

Yen S, Robins CJ, Lin N (2000). A cross-cultural comparison of depressive symptom manifestation: China and the United States. *Journal of Consulting and Clinical Psychology* **68**, 993–999.

Young MA, Scheftner WA, Fawcett J, Klerman GL (1990). Gender differences in the clinical features of unipolar major depressive disorder. *Journal of Nervous and Mental Disease* **178**, 200–203.

https://doi.org/10.1017/S0033291708003875 Published online by Cambridge University Press