

Differential Item Functioning on the Five Facet Mindfulness Questionnaire Is Minimal in Demographically Matched Meditators and Nonmeditators

Ruth A. Baer, Douglas B. Samuel and Emily L. B. Lykins
Assessment 2011 18: 3 originally published online 30 December 2010
DOI: 10.1177/1073191110392498

The online version of this article can be found at:
<http://asm.sagepub.com/content/18/1/3>

Published by:



<http://www.sagepublications.com>

Additional services and information for *Assessment* can be found at:

Email Alerts: <http://asm.sagepub.com/cgi/alerts>


Subscriptions: <http://asm.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

Citations: <http://asm.sagepub.com/content/18/1/3.refs.html>

Differential Item Functioning on the Five Facet Mindfulness Questionnaire Is Minimal in Demographically Matched Meditators and Nonmeditators

Assessment
18(1) 3–10
© The Author(s) 2011
Reprints and permission: <http://www.sagepub.com/journalsPermissions.nav>
DOI: 10.1177/1073191110392498
<http://asm.sagepub.com>


Ruth A. Baer¹, Douglas B. Samuel², and Emily L. B. Lykins¹

Abstract

A recent study of the Five Facet Mindfulness Questionnaire reported high levels of differential item functioning (DIF) for 18 of its 39 items in meditating and nonmeditating samples that were not demographically matched. In particular, meditators were more likely to endorse positively worded items whereas nonmeditators were more likely to deny negatively worded (reverse-scored) items. The present study replicated these analyses in demographically matched samples of meditators and nonmeditators ($n = 115$ each) and found that evidence for DIF was minimal. There was little or no evidence for differential relationships between positively and negatively worded items for meditators and nonmeditators. Findings suggest that DIF based on items' scoring direction is not problematic when the Five Facet Mindfulness Questionnaire is used to compare demographically similar meditators and nonmeditators.

Keywords

Five Facet Mindfulness Questionnaire, differential item functioning, reverse-scored items, mindfulness meditation

The assessment of mindfulness is critically important for several reasons. Mindfulness is often conceptualized as a traitlike or dispositional variable (Brown & Ryan, 2003). People who are high in mindfulness tend to be aware of and attentive to the present moment experiences of daily life and to adopt an attitude of nonjudgmental acceptance toward these experiences (Kabat-Zinn, 1982). Without methods for assessing mindfulness, it is difficult to study relationships between the tendency to be mindful in daily life and other psychological variables. In addition, recently developed interventions that emphasize training in mindfulness skills have accrued substantial empirical support for their efficacy and are increasingly available in a wide variety of medical and mental health settings. Understanding the mechanisms through which these treatments produce beneficial effects requires assessment of whether participants are becoming more mindful as a result of participation in treatment and whether changes in mindfulness are functionally related to other outcomes (Dimidjian & Linehan, 2003).

Several measures of mindfulness have been developed in recent years. Most use self-report methods to assess the general tendency to be mindful in daily life. The Five Facet Mindfulness Questionnaire (FFMQ; Baer, Smith, Hopkins, Krietemeyer, & Toney, 2006) is based on factor analysis of the combined item pool from five independently developed mindfulness questionnaires. Findings

suggested that mindfulness can be conceptualized as a multifaceted construct consisting of several related skills. *Observing* is the tendency to notice or attend to internal and external experiences, such as sensations, emotions, cognitions, sounds, sights, and smells. *Describing* involves labeling observed experiences with words. *Acting with awareness* refers to paying attention to ongoing activity and is often contrasted with behaving mechanically while attention is focused elsewhere (often called *automatic pilot*). *Nonjudging of inner experience* involves taking a nonevaluative stance toward cognitions and emotions. *Nonreactivity to inner experience* is the tendency to allow feelings and thoughts to come and go, without getting carried away by or caught up in them.

Recent studies of the FFMQ (Baer et al., 2006; Baer et al., 2008) have shown that the five facets have adequate to very good internal consistency in several samples, including students, nonmeditating community members, and experienced meditators. Most alpha coefficients have been more than .80,

¹University of Kentucky, Lexington, KY, USA

²Yale School of Medicine, New Haven, CT, USA

Corresponding Author:

Ruth A. Baer, Department of Psychology, 115 Kastle Hall, University of Kentucky, Lexington, KY 40506-0044, USA
Email: rbaer@email.uky.edu

except for the *nonreactivity* scale in student samples, where alpha has been somewhat lower (.67-.72). The five facets are moderately correlated with each other, and with a few exceptions are correlated in the expected directions with a wide variety of constructs that should be related to mindfulness, such as emotional intelligence, thought suppression, and experiential avoidance (Baer et al., 2006). Factor analyses with experienced meditators support a hierarchical model in which the five facets are elements of an overarching mindfulness construct. In nonmeditating samples, four of the five facets (all but *observing*) fit this model (Baer et al., 2006; Baer et al., 2008). Carmody and Baer (2008) reported that FFMQ scores increased with participation in mindfulness-based stress reduction (Kabat-Zinn, 1982, 1990), an 8-week group intervention based on intensive training in mindfulness meditation practices. Carmody and Baer (2008) also found that increases in FFMQ scores partially mediated the relationship between home practice times and improved psychological functioning. FFMQ scores have been shown to correlate significantly with the extent of meditation experience in long-term practitioners of mindfulness meditation and to account for significant variance in the relationship between meditation experience and psychological well-being (Baer et al., 2008).

This emerging literature provides promising evidence for the construct validity of interpretations based on FFMQ scores. However, another potentially informative step in the validation process is to examine the measure for differential item functioning (DIF), in which groups of respondents with the same level of the construct being measured (i.e., comparable total scores on a particular instrument) have significantly different responses to individual items. This can be problematic if group differences on such an item are determined to be the result of bias (characteristics extraneous to the test) rather than item impact (true differences in the ability measured by the test; see Ackerman, 1992). Tests of DIF are typically conducted for demographic variables such as gender, age, or race to detect items that may favor one group over another. For example, analyses of DIF are used extensively in tests of academic achievement to identify items that differ across racial groups. The fundamental logic of DIF is that when groups are equated on the trait under investigation, then any significant differences that emerge for individual items might be related to something other than the trait assessed by the measure.

In a recent examination of the FFMQ, Van Dam, Earleywine, and Danoff-Burg (2009) offered a novel extension of this logic by investigating DIF between groups with differing levels of meditation experience. Van Dam et al. recruited a sample of experienced meditators from meditation listservs ($n = 58$) and a nonmeditating sample of undergraduate students ($n = 263$) and asked them to complete the FFMQ online. As expected, meditators scored higher on the FFMQ than nonmeditators and extent of meditation history

was significantly correlated with FFMQ total score. Because the sample of meditators was relatively small, Van Dam et al. used three nonparametric procedures to investigate DIF, including the Mantel-Haenszel Statistic, the Liu-Agresti common log odds ratio, and Cox's noncentrality parameter. All three statistics were calculated using DIFAS 4.0 (Penfield, 2007b). These analyses showed DIF in 18 of the FFMQ's 39 items. For these 18 items, the apparent direction of the DIF was influenced by the items' scoring direction. Positively worded items, which describe an element of mindfulness, favored the meditators, whereas negatively worded items, which describe an element of *mindlessness* and are reverse-scored, favored the nonmeditators.

Van Dam et al. (2009) also conducted *t* tests to examine the mean endorsement of items by the meditators and nonmeditators and found differences between the groups for positively and negatively worded items. Specifically, meditators were more likely to endorse the positively worded items whereas nonmeditators were more likely to deny the negatively worded items. Thus, meditators were more likely to endorse mindfulness whereas nonmeditators were more likely to deny mindlessness.

From these results, Van Dam et al. (2009) concluded that "despite good classical psychometric properties, the FFMQ functions differently in meditators and non-meditators" (p. 520). They further suggested that the use of the FFMQ for comparing groups of meditators and nonmeditators, as well as for assessing changes in mindfulness over the course of an intervention, would prove problematic. Nonetheless, Van Dam et al. also noted that a potential weakness of their study was that the meditating and nonmeditating samples were not matched on age, gender, or education. Thus, differences in item functioning could have been related to demographic differences between the samples, rather than meditating status (e.g., Finch & French, 2008). The purpose of the present study, therefore, was to examine DIF in FFMQ items using samples of meditators and nonmeditators that were matched on several demographic variables.

Method

Participants and Procedures

Participants were 115 meditators and 115 nonmeditators. FFMQ responses for all participants were taken from existing data sets. Although other findings for these participants have been reported by Baer et al. (2008) and Lykins and Baer (2009), DIF analyses have not previously been reported. Meditators were recruited in several ways. Some had attended a conference on mindfulness at the University of Massachusetts Medical School in 2005 and were subsequently mailed a packet of questionnaires including the FFMQ and numerous other measures. Others were recruited through announcements posted to Internet-based

groups focused on mindfulness or meditation. In addition, flyers describing the study were distributed in meditation and yoga centers and posted in the local community. Interested persons contacted the experimenter to request a questionnaire packet, which was mailed along with a postage-paid return envelope. Many of the experienced meditators held graduate degrees and some worked in the mental health field. Demographically similar nonmeditators therefore were recruited through mailings and flyers sent to faculty and staff in several departments at local colleges and universities and to mental health professionals in local clinics, hospitals, and private practices. More detailed description of these samples is provided by Baer et al. (2008).

Data Analytic Approach

Defined most broadly, DIF is the degree to which groups show differential probabilities of endorsing an item *after* the groups have been equated on the ability the item is supposed to measure. There are many methods for detecting whether DIF is present. For example, item response theory methodologies can be used to equate groups in terms of the latent trait. Although item response theory techniques can be quite useful for detecting DIF, they require very large sample sizes and assume unidimensionality within the instrument. Another approach from classical test theory is to equate groups using the total score on an instrument. DIF can then be detected using logistic regression (Zumbo, 1999), to determine whether group membership has incremental validity over the total score in predicting a response to an individual item.

As the intention of the current study was to replicate the work of Van Dam and colleagues, we elected to use the same statistical procedures and software (DIFAS 4.0; Penfield, 2007b). The three indicators of DIF were the Mantel–Haenszel chi square (Mazor, Clauser, & Hambleton, 1992), the Liu–Agresti common log odds ratio (L-A LOR; Liu & Agresti, 1996), and Cox’s noncentrality parameter (Cox’s B; Penfield, 2007a). All three of these approaches first divide individuals into 10 “bands” based on their total scores. In the current study, these bands were 10 units wide. All individuals within the same band are considered equated and thus any differences between individuals based on meditator status within a given band are considered indicative of DIF. The Mantel–Haenszel is a chi-square statistic with one degree of freedom and is calculated by arraying the five response options for each item against group membership (i.e., meditators vs. nonmeditators) and then determining whether the probability of each response differs based on group status. The L-A LOR is an extension that considers the log odds ratio of one group endorsing a response option relative to another. Positive values indicate DIF in favor of the reference group (e.g., meditators), and negative values indicate DIF in favor of the focal group

Table 1. Demographic Characteristics of Meditating and Nonmeditating Samples

	Meditators	Nonmeditators	F or χ^2	p
Age in years	45.69 (11.55)	43.57 (11.75)	F = 1.89	.17
Years of education	18.77 (2.12)	18.68 (2.02)	F = 0.12	.73
Sex (% male)	27%	37%	$\chi^2 = 2.87$.09
Race (% White)	95%	91%	$\chi^2 = 1.07$.30
Percentage of MH professionals	54%	50%	$\chi^2 = .28$.60
Years of experience in MH field	14.37 (9.28)	15.17 (10.27)	F = 0.18	.67

Note. MH = mental health.

(e.g., nonmeditators). Finally, Cox’s B is similar to the Mantel–Haenszel statistic except that it uses the hypergeometric mean. It is distributed similarly to L-A LOR such that positive values favor the meditators, whereas negative values favor the nonmeditators.

Results

Demographic characteristics of the meditating and nonmeditating samples are shown in Table 1. Differences between groups were tested with one-way analyses of variance for continuous variables and chi-square analyses for categorical variables. Group differences for age, years of education, sex, race, status as a mental health professional, and years of experience in the mental health field were not significant.

As expected, total scores for the FFMQ differed significantly between groups, with a moderate effect size ($d = .57$). For meditators, mean (M) = 148.97 ($SD = 17.46$), whereas for nonmeditators, $M = 138.87$ ($SD = 18.04$), $F = 18.24$, $p < .0001$. DIF indicators for all FFMQ items are shown in Table 2. We used the same strict Bonferroni correction suggested by Van Dam et al. (2009). That is, because we tested 39 items with 3 tests each (117 comparisons), we adopted a p value of .00043 (.05 divided by 117) as the criterion for statistical significance. As noted Van Dam et al. (2009), conservative rules for identifying DIF are important when the two groups have different ability distributions on the measure in question. Only one item (Item 11) showed significant DIF using this strict criterion, and only for two of the three indicators. Van Dam et al. (2009) also reported significant DIF for this item, which favored the meditators. We also examined a somewhat less stringent p value of .001, which is consistent with a Bonferroni correction for only one test (.05 divided by 39 comparisons). By this less stringent standard, Items 11 and 18 showed significant DIF

Table 2. Tests of Differential Item Functioning for All FFMQ Items

FFMQ Item	Subscale	Mantel χ^2	L-A LOR	SE	Cox's B	SE
1	Observe	11.31**	1.07	.33	.60**	.18
2	Describe	0.11	0.12	.35	.08	.23
3	Nonjudge (r)	2.07	-0.46	.33	-.27	.19
4	Nonreact	0.02	-0.05	.32	-.04	.27
5	Act aware (r)	5.54	-0.72	.30	-.52	.22
6	Observe	0.07	0.08	.31	.05	.19
7	Describe	0.22	-0.15	.32	-.11	.24
8	Act aware (r)	6.08	-0.79	.31	-.68	.27
9	Nonreact	2.42	-0.56	.35	-.40	.26
10	Nonjudge (r)	0.52	0.21	.31	.15	.21
11	Observe	13.15***	1.13**	.34	.60***	.17
12	Describe (r)	0.19	-0.14	.32	-.11	.25
13	Act aware (r)	5.80	-0.74	.30	-.54	.22
14	Nonjudge (r)	0.03	0.06	.37	.04	.22
15	Observe	6.07	0.76	.33	.53	.22
16	Describe (r)	0.02	-0.05	.33	-.04	.25
17	Nonjudge (r)	0.00	0.02	.33	.01	.18
18	Act aware (r)	11.08**	-1.16**	.35	-.87**	.26
19	Nonreact	0.42	0.22	.32	.15	.24
20	Observe	1.37	0.37	.31	.25	.21
21	Nonreact	1.18	-0.35	.32	-.24	.22
22	Describe (r)	3.30	-0.54	.30	-.44	.24
23	Act aware (r)	9.94	-1.07**	.32	-.70	.22
24	Nonreact	7.49	1.00	.35	.57	.21
25	Nonjudge (r)	0.00	0.01	.33	.01	.25
26	Observe	0.28	0.16	.33	.13	.25
27	Describe	0.20	0.16	.34	.10	.22
28	Act aware (r)	1.89	-0.43	.30	-.35	.25
29	Nonreact	2.94	0.56	.32	.42	.24
30	Nonjudge (r)	0.19	-0.14	.34	-0.10	.24
31	Observe	1.98	0.44	.32	.28	.20
32	Describe	0.03	-0.05	.32	-0.03	.19
33	Nonreact	7.36	0.96	.34	.65	.24
34	Act aware (r)	3.63	-0.63	.30	-.40	.21
35	Nonreact (r)	1.96	-0.47	.34	-.33	.23
36	Observe	3.87	0.66	.32	.54	.27
37	Describe	2.11	0.48	.30	.32	.22
38	Act aware (r)	8.15	-0.97	.33	-.64	.22
39	Nonjudge (r)	0.00	0.02	.35	.01	.21

Note. FFMQ = Five Facet Mindfulness Questionnaire; r = reverse-scored item; Mantel = Mantel-Haenszel chi-square (1 *df*, one-tailed); L-A LOR = Liu-Agresti log odds ratio (two-tailed); SE = standard error; Cox's B = Cox's noncentrality parameter (two-tailed).

** $p < .001$ (two-tailed; Bonferroni correction for one test). *** $p < .00043$ (two-tailed; Bonferroni correction for three tests).

according to all three indicators, Item 1 according to two indicators, and Item 23 according to only one indicator. Items 1 and 11 favored the meditators, whereas Items 18 and 23 showed DIF that favored the nonmeditators. Items 1

Table 3. Tests of Differential Item Functioning for 18 FFMQ Items

FFMQ Item	Subscale	Mantel χ^2	L-A LOR	SE	Cox's B	SE
1	Observe	19.64*	1.48*	.35	.81*	.18
2	Describe	.07	.08	.31	.06	.23
5	Act aware (r)	8.91	-.94	.33	-.66	.22
10	Nonjudge (r)	.03	-.05	.31	-.04	.23
11	Observe	17.90*	1.32*	.31	.67*	.16
12	Describe (r)	1.50	-.44	.35	-.30	.24
13	Act aware (r)	11.48*	-1.00*	.31	-.72	.21
14	Nonjudge (r)	.91	-.36	.36	-.23	.24
15	Observe	7.99	.96	.36	.66	.23
17	Nonjudge (r)	2.24	-.45	.31	-.30	.20
25	Nonjudge (r)	2.77	-.59	.36	-.45	.27
29	Nonreact	.27	.15	.31	.14	.27
31	Observe	1.07	.32	.32	.21	.20
33	Nonreact	3.62	.59	.32	.44	.23
34	Act aware (r)	7.26	-.87	.32	-.60	.22
35	Nonjudge (r)	12.85*	-1.28*	.37	-.85*	.24
36	Observe	3.30	.64	.36	.44	.24
39	Nonjudge(r)	2.84	-.59	.35	-.43	.25

Note. FFMQ = Five Facet Mindfulness Questionnaire; r = reverse-scored item; Mantel = Mantel-Haenszel chi-square (1 *df*, one-tailed); L-A LOR = Liu-Agresti log odds ratio (two-tailed); SE = standard error; Cox's B = Cox's noncentrality parameter (two-tailed).

* $p < .0028$.

and 11 also were identified by Van Dam et al. (2009) as showing DIF whereas Items 18 and 23 were not. Overall, 4 of the 39 items (10.26%) showed some evidence of significant DIF when the less stringent standard was used. This is considerably less than the threshold of 25% of items suggested by Penfield and Algina (2006) for indicating that the instrument as a whole may yield biased results. Of these four items, two are from the *observing* subscale and two are from the *acting with awareness* subscale. We also investigated differential test functioning (DTF) and found that the weighted v^2 was .23 ($SE = .08$), which was nonsignificant.

To provide an even more conservative test and replication we also conducted a separate DIF analysis focusing only on those 18 items that were identified by Van Dam et al. (2009) as showing significant DIF. This approach is particularly rigorous because, as described above, the DIF procedure first equates the meditators and nonmeditators based on their summed score. Thus, this summed score ignores the 21 items that were not identified by Van Dam as having significant DIF. As these were secondary analyses, we also chose to use the rather liberal α level of .0028 (i.e., .05/18), which is consistent with a Bonferroni correction for only one test and does not take into account the previous analysis. As can be seen in Table 3, this test also yielded similar results, as only four items reached statistical significance.

Table 4. Comparisons of Means for Positively and Negatively Worded FFMQ Items

	Positively Worded Items		Negatively Worded Items		<i>t</i>	<i>d</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Analysis 1: All FFMQ items						
Meditators	3.83	0.46	3.81	0.51	.79 <i>ns</i>	.04
Nonmeditators	3.43	0.53	3.69	0.56	4.82**	.48
Analysis 2: Describe items only						
Meditators	3.93	0.75	4.00	0.72	1.28 <i>ns</i>	.10
Nonmeditators	3.64	0.78	3.88	0.73	4.53**	.30
Analysis 3: Nonreact and nonjudge items only						
Meditators	3.61	0.53	4.02	0.67	8.02**	.68
Nonmeditators	3.32	0.61	3.77	0.73	7.36**	.67

Note. FFMQ = Five Facet Mindfulness Questionnaire. Possible range for all means is 1 to 5.

** $p < .001$.

Items 1 and 11, which favored the meditators, were again identified as having DIF. However, items 13 and 35, which favored the nonmeditators, emerged as significant in this analysis despite nonsignificant results in the previous analysis. The overall test for DTF was also nonsignificant, suggesting that even this subset of items, when treated as a scale, did not evidence differential functioning (weighted $v^2 = .48$, $SE = .20$).

We also replicated analyses conducted by Van Dam et al. (2009) for examining differences between positively worded and negatively worded (reverse-scored) items. First, we correlated the total score for positively worded items with the total score for negatively worded items in each sample separately. For the meditators, $r = .73$ ($p < .001$). For the nonmeditators, $r = .46$ ($p < .001$). These correlations are significantly different ($z = 3.19$, $p < .001$), showing that the two types of items are less strongly associated in nonmeditators than in meditators. We also compared means for positively and negatively worded items in the two groups. As shown in the upper portion of Table 4 (Analysis 1), within the meditating sample, the means for positively and negatively worded items were nearly identical. In contrast, within the nonmeditators the difference between them was statistically significant. This pattern is also consistent with the findings of Van Dam et al. (2009) in showing a greater difference between positive and negative items in nonmeditators than in meditators.

However, conducting the analyses in this way is problematic because it confounds the items' scoring direction with their content. The FFMQ's five subscales have been shown to assess distinct content and to be only moderately intercorrelated (Baer et al., 2006; Baer et al., 2008). This pattern holds in the current data set, as shown in Table 5. In addition, reverse-scored items are not evenly distributed

Table 5. Intercorrelations Between Five Facet Mindfulness Questionnaire Subscales for Meditators and Nonmeditators

	Describe	Act Aware	Nonjudge	Nonreact
Observe	.41* (.36*)	.42* (.16)	.54* (.11)	.53* (.26*)
Describe	—	28* (.33*)	.44* (.26*)	.38* (.41*)
Act aware	—	—	.44* (.41*)	.47* (.35*)
Nonjudge	—	—	—	.60* (.54*)

Note. Values for nonmeditators are in parentheses. Boldface indicates that the two correlations in the pair are significantly different at $p < .05$. * $p < .01$.

across the FFMQ's subscales. Instead, the *observing* and *nonreactivity* subscales are entirely positively worded, whereas the *acting with awareness* and *nonjudging* subscales are entirely negatively worded. Only the *describing* subscale includes a mix of positively and negatively worded items. This means that, to a very large extent, items with different scoring directions also have distinct content. Avoiding this confound when looking for wording effects would require examining items with similar content but opposite scoring directions. The FFMQ is not well suited to such analyses because most of the subscales are only moderately intercorrelated and do not contain both types of items.

Nevertheless, we explored two ways of analyzing FFMQ items with similar content but opposite scoring directions. First, we considered the *describing* scale alone, which contains three reverse-scored items and five positively worded items. Correlations between the scores for the two types of items were .69 for meditators and .73 for nonmeditators. These two correlations are not significantly different ($z = .60$, $p > .05$), suggesting that the positively and negatively worded *describing* items are equally strongly associated in meditators and nonmeditators. We also examined means for the positively and negatively worded items from the *describing* scale. These are shown in the middle section of Table 4 (Analysis 2). For meditators, the difference between positively and negatively worded items was nonsignificant. However, for nonmeditators, this difference was statistically significant, though small ($d = .30$).

Conducting these analyses with the *describing* items alone is less than ideal because the number of items available for analysis is quite small. Analyzing a larger pool of items requires combining subscales. To preserve the greatest possible similarity of item content, we repeated the analyses just described using only the 15 items that appear on the *nonjudging* and *nonreactivity* subscales. We chose these two subscales because they have the largest intercorrelation of any pair of subscales (see Table 5) in both meditators and nonmeditators. Thus, their content is reasonably similar (more similar than any other pair of subscales). When combined, they create a pool with eight positively worded and

seven negatively worded items. Internal consistencies (alpha coefficients) for this 15-item subset of the FFMQ were .91 in meditators and .90 in nonmeditators. Within this subset of items, correlations between positive and negative items were .60 for meditators and .54 for nonmeditators. These correlations are not significantly different ($z = .66, p > .05$), suggesting no differential relationship between meditators and nonmeditators for positive and negative items. Means for the two types of items are shown in the third section of Table 4 (Analysis 3). Although both groups scored significantly higher on the *nonjudging* than the *nonreactivity* items, the effect sizes for these differences were nearly identical for meditators and nonmeditators ($d_s = .68$ and $.67$, respectively). These findings also suggest that when the items have reasonably similar content, there is no differential relationship between meditators and nonmeditators for positively versus negatively worded items.

Discussion

The purpose of this study was to investigate DIF on the FFMQ in demographically matched samples of meditators and nonmeditators. Four FFMQ items showed DIF but only when using a less stringent standard for statistical significance than has previously been recommended (Van Dam et al., 2009). These findings differ from those reported by Van Dam et al. (2009), who used samples of meditators and nonmeditators that were not demographically matched. Thus, the current findings suggest that when groups are demographically similar, DIF in the FFMQ is minimal. As DIF is a necessary, but not sufficient, condition for item bias, this further indicates that problematic bias within the FFMQ's items is unlikely.

Although significant DIF was not shown for most of the FFMQ items, it was detected for a few. In particular, Items 1, 11, 13, and 35 showed significant DIF within our secondary analysis confined to only those items identified by Van Dam et al. (2009). The latter of these (Items 13 and 35) were not identified in the primary analyses of the current study, which included all 39 FFMQ items. Their significance in our secondary analysis, which used a more liberal alpha level, may have capitalized on chance. However, it must be noted that item 11 evinced significant DIF with Van Dam's analyses and across all the analyses and detection methods used in the current study. This convergence indicates that this particular item has notable DIF across groups and warrants further investigation.

It is worth noting that the presence of DIF, in and of itself, is not problematic but raises the possibility that the item is biased (i.e., that group differences are due to something other than the construct of mindfulness; Ackerman, 1992). DIF can also be indicative of an item with high impact, which is defined by Ackerman (1992) as a between-group difference caused by a true difference in the ability

being measured. The content of Item 11 ("I notice how foods and drinks affect my thoughts, bodily sensations, and emotions") does not suggest any obvious bias but instead appears quite face valid and consistent with the definition of mindfulness as the tendency to be aware of and attentive to the present-moment experiences of daily life (Brown & Ryan, 2003). In fact, this item obtained the largest effect size difference ($d = .83$) between the two groups in the current sample, suggesting that it is the most powerful item for discriminating among demographically similar meditators and nonmeditators. Although future research investigating whether Item 11 (and perhaps a few others with significant DIF) contains bias is warranted, the present analysis suggests that rather than bias, the DIF may be indicative of this item's impact and strength.

Even in our demographically matched samples, differential relationships between positively and negatively worded items in meditators and nonmeditators were observed when we replicated the analyses of Van Dam et al. (2009), who used the entire FFMQ item pool. However, we argue that these analyses confound distinct item content with scoring direction. When we repeated these analyses on subsets of FFMQ items with more similar item content but opposite scoring directions, evidence for differential relationships was minimal or absent.

The merits of reverse-scored items on self-report instruments have been widely discussed. Some experts argue that well designed questionnaires should have equal numbers of positive and negative items on each subscale to control for response biases (Nunnally, 1967; Paulhus, 1991). However, others suggest that reverse-scored items introduce method biases that complicate interpretation (DiStefano & Motl, 2009; Marsh, 1996). Empirical analyses provide conflicting findings. Some studies report that reverse-scored items introduce method effects but do not compromise the validity of the instrument's total score and can therefore be retained (e.g., Hazlett-Stevens, Ullman, & Craske, 2004). Others conclude that reverse-scored items should be eliminated because they cause confusion and reduce reliability (e.g., Conrad et al., 2004) or because they measure a somewhat different construct from the positively worded items (e.g., Rodebaugh, Woods, & Heimberg, 2007). Unfortunately, these studies are not readily applicable to the FFMQ because they test whether instruments that were designed to be unifactorial actually have two factors based on the items' scoring direction. We found no studies of the utility of reverse-scored items that fall on separate subscales within multidimensional instruments.

Brown and Ryan (2003) reported that, in the development of the Mindful Attention Awareness Scale, reverse-scored items were *more* psychometrically sound than positively worded items. As a result, the Mindful Attention Awareness Scale is entirely negatively worded. Similarly, Baer, Smith, and Allen (2004), in the development of the

Kentucky Inventory of Mindfulness Skills, reported that positively worded items for the *accept-without-judgment* subscale had to be eliminated because of poor item–total correlations. Only reverse-scored items were retained for this subscale. Thus, it is possible that some elements of mindfulness are more reliably assessed with reverse-scored items.

Most mindfulness questionnaires include both positively and negatively worded items. Only future research can clarify whether a multidimensional mindfulness instrument with either balanced scoring directions on each subscale, or items scored in only one direction throughout the instrument, would show stronger psychometric properties than the FFMQ in its current form. As noted earlier, FFMQ subscales show numerous expected relationships with other variables regardless of their scoring direction. Factor analyses are largely consistent with a hierarchical structure in which the subscales show high loadings on an overarching mindfulness construct. The only exception to this pattern involves the *observing* scale, which is entirely positively worded and which appears to function differently in meditating and nonmeditating samples (Baer et al., 2006). Baum et al. (IN PRESS) reported a similar pattern for the Kentucky Inventory of Mindfulness Skills, a four-factor mindfulness instrument similar to the FFMQ. The present study focused specifically on whether positively and negatively worded FFMQ items show different response patterns in meditators and nonmeditators. Results suggest that when item content is similar and groups are demographically matched, such differences are small or absent. Thus, we conclude that DIF based on scoring direction is not a significant problem when the FFMQ is used to compare demographically similar meditators and nonmeditators. Our findings also imply that the FFMQ is likely to be suitable for pre–post intervention data from mindfulness-based treatments such as mindfulness-based stress reduction. In these studies, most participants are nonmeditators at pretreatment but have been meditators for 8 weeks at posttreatment.

Limitations in the present study must be acknowledged. The samples, although demographically matched, have an unusually high level of education and proportion of mental health professionals. Our experience in recruiting regular meditators suggests that most have higher than average levels of education and that mental health professionals are commonly drawn to meditation. Even among the nonmeditating sample, those who were mental health professionals may have had some knowledge of mindfulness. Familiarity with mindfulness was not assessed. It is therefore important to examine DIF on the FFMQ (and perhaps on other mindfulness measures) in matched samples that are more representative of the general population. In addition, although it is difficult to recruit very large groups of meditators, larger sample sizes would allow for analyses using potentially more sophisticated DIF detection strategies based on item

response theory (Embretson & Reise, 2000). Finally, although DIF was minimal in the present study, a few items showed evidence of it. Two of these items were from the *observing* scale, which has previously been shown to have different relationships with other variables in meditating versus nonmeditating samples (Baer et al., 2006; Baer et al., 2008). Thus, it would be helpful for future research to clarify whether the DIF is because of potential bias within the items or simply indicates their strength and impact. On balance, however, the present findings are useful in suggesting that DIF in the FFMQ is probably not a significant issue when meditators and nonmeditators are demographically similar.

Declaration of Conflicting Interests

The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

Data analysis for this manuscript was supported by the Office of Academic Affiliations, Advanced Fellowship Program in Mental Illness Research and Treatment, Department of Veterans Affairs.

References

- Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.
- Baer, R. A., Smith, G. T., & Allen, K. B. (2004). Assessment of mindfulness by self-report: The Kentucky Inventory of Mindfulness Skills. *Assessment, 11*, 191-206.
- Baer, R. A., Smith, G. T., Hopkins, J., Krietemeyer, J., & Toney, L. (2006). Using self-report assessment methods to explore facets of mindfulness. *Assessment, 13*, 27-45.
- Baer, R. A., Smith, G. T., Lykins, E., Button, D., Krietemeyer, J., Sauer, S., . . . Williams, M. (2008). Construct validity of the five facet mindfulness questionnaire in meditating and nonmeditating samples. *Assessment, 15*, 329-342.
- Baum, C., Kuyken, W., Bohus, M., Heidenreich, T., Michalak, J., & Steil, R. (IN PRESS). The psychometric properties of the Kentucky Inventory of Mindfulness Skills in clinical populations. *Assessment*.
- Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology, 84*, 822-848.
- Carmody, J., & Baer, R. A. (2008). Relationships between mindfulness practice and levels of mindfulness, medical and psychological symptoms and well-being in a mindfulness-based stress reduction program. *Journal of Behavioral Medicine, 31*, 23-33.
- 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*, 15-30.
- Dimidjian, S., & Linehan, M. (2003). Defining an agenda for future research on the clinical application of mindfulness practice. *Clinical Psychology: Science and Practice, 10*, 166-171.

- DiStefano, C., & Motl, R. (2009). Personality correlates of method effects due to negatively worded items on the Rosenberg Self-Esteem Scale. *Personality and Individual Differences, 46*, 309-313.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Finch, W. H., & French, B. F. (2008). Anomalous type I error rates for identifying one type of differential item functioning in the presence of the other. *Educational and Psychological Measurement, 68*, 742-759.
- Hazlett-Stevens, H., Ullman, J. B., & Craske, M. G. (2004). Factor structure of the Penn State Worry Questionnaire: Examination of a method factor. *Assessment, 11*, 361-370.
- Kabat-Zinn, J. (1982). An outpatient program in behavioral medicine for chronic pain patients based on the practice of mindfulness meditation: Theoretical considerations and preliminary results. *General Hospital Psychiatry, 4*, 33-47.
- Kabat-Zinn, J. (1990). *Full catastrophe living: Using the wisdom of your body and mind to face stress, pain, and illness*. New York, NY: Delacorte.
- Liu, I., & Agresti, A. (1996). Mantel-Haenszel type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics, 52*, 1223-1234.
- Lykins, E. L., & Baer, R. A. (2009). Psychological functioning in a sample of long-term practitioners of mindfulness meditation. *Journal of Cognitive Psychotherapy, 23*, 226-241.
- Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifacts? *Journal of Personality and Social Psychology, 70*, 810-819.
- Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*, 443-451.
- Nunnally, J. C. (1967). *Psychometric theory*. New York, NY: McGraw Hill.
- Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press.
- Penfield, R. D. (2007a). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education, 20*, 335-355.
- Penfield, R. D. (2007b). *DIFAS 4.0 Differential item functioning analysis system: User's manual*. Unpublished manuscript.
- Penfield, R. D., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement, 43*, 295-312.
- Rodebaugh, T. L., Woods, C. M., & Heimberg, R. G. (2007). The reverse of social anxiety is not always the opposite: The reverse-scored items of the Social Interaction Anxiety Scale do not belong. *Behavior Therapy, 38*, 192-206.
- Van Dam, N. T., Earleywine, M., & Danoff-Burg, S. (2009). Differential item functioning across meditators and nonmeditators on the Five Facet Mindfulness Questionnaire. *Personality and Individual Differences, 47*, 516-521.
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item score*. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.