Personality and Individual Differences 47 (2009) 516–521

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Differential item function across meditators and non-meditators on the Five Facet Mindfulness Questionnaire

Nicholas T. Van Dam *, Mitch Earleywine, Sharon Danoff-Burg

University at Albany, SUNY, Department of Psychology, Social Sciences 369, 1400 Washington Avenue, Albany, NY 12222, USA

A R T I C L E   I N F O

Article history:
Received 13 January 2009
Received in revised form 3 May 2009
Accepted 7 May 2009
Available online 30 May 2009

Keywords:
Mindfulness
Self-report
Differential item functioning
Meta-consciousness

A B S T R A C T

Mindfulness has been defined as a state of consciousness involving intentional attention and awareness of the present moment. Reporting on past conscious experience is inherently tricky and presents unique challenges to the assessment of mindfulness. Mindfulness-present and mindfulness-absent items may represent different aspects of the construct to different populations resulting from differential skill in assessing sustained or lapsed conscious attention. The current study shows that an online sample of meditators and non-meditators with similar overall levels of mindfulness differentially endorse response options for positively and negatively worded items. While meditators endorse mindfulness-present and mindfulness-absent items with nearly equivalent frequency, student non-meditators are much more likely to reject mindfulness-absent items than to accept mindfulness-present items. The differential item functioning between these two groups represents a potential problem regarding construct validity when comparing meditators to non-meditators and when assessing mindfulness as a pre-post measure with meditation practice.

1. Introduction

Common definitions suggest that mindfulness is a state of consciousness involving present-centered attention or awareness (e.g. Brown & Ryan, 2003). Assessing this aspect of consciousness, however, proves tricky (see Grossman, 2008; Schooler, 2002). Assessing the frequency of one's own present-centered awareness may require complicated recall as well as metacognitive awareness of awareness (see Schooler, 2002). In addition, some reverse-scored items from mindfulness questionnaires assess lapses in attention rather than mindful moments. The number of lapses an individual might notice depends upon recognizing a lapse has occurred, which also requires meta-consciousness. Thus, people who are more *mindful* might actually be more adept at recognizing attentional lapses, creating odd relations between these items and the construct of interest. Trying to re-represent an experience one was potentially unaware of in the first place likely increases error and bias (Schooler, 2002). Recent work on negatively worded items assessing self-esteem reveals that impulsive, potentially less-mindful participants are less likely to endorse negatively worded items simply because of the way they are worded (DiStefano & Motl, 2009). Thus, negatively worded items may prove less accurate in assessing mindfulness than other items.

Evidence from meditators provides modest support for the validity of mindfulness questionnaires. Meditators score higher than non-mediators on relevant scales (e.g., Baer, Smith, Hopkins, Krietemeyer, & Toney et al., 2006) – a result potentially consistent with genuine assessment of mindfulness. Alternatively, higher mindfulness scores in meditators might stem from a different understanding of the items. Interpreting group differences on a scale requires that the scale has equivalent meaning across the groups.

Important conclusions are often reached via comparisons between long-term meditating samples and samples of convenience (Grossman, 2008). These comparisons and the notion of mindfulness as dispositional (e.g., Brown & Ryan, 2003) seem to suggest measurement invariance. Unfortunately, data suggest that relationships between aspects of mindfulness may be different for meditators. For example, factor analyses identify an Observe subscale of the Five Facet Mindfulness Questionnaire (FFMQ; Baer et al., 2006) in meditators that does not appear in non-meditators. Similarly, the FFMQ subscales differentially relate to psychological well-being in comparisons of meditators and non-meditators (Baer, Smith, Lykins, Button, & Krietemeyer et al., 2008). While assessing factorial invariance can identify different response properties by group, differential item functioning (DIF) better assesses differential response bias or demand (Teresi, 2006), an issue of central concern for comparisons of meditators and non-meditators (Grossman, 2008).

DIF occurs when individuals with the same overall 'amount' of a given construct (as assessed by the relevant scale) have a different probability of selecting a given response option on an individual item. Under item invariance, overall scale score should be the

* Corresponding author. Tel.: +1 518 445 5533; fax: +1 518 442 4867.
E-mail address: NV122142@albany.edu (N.T. Van Dam).

0191-8869/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.paid.2009.05.005

primary predictor of which response option a participant selects, not group membership. Individuals with high levels of the construct (as assessed by the scale) should be more likely to choose response options that represent higher levels of the construct on a given item, regardless of any group affiliation. When an item lacks invariance (exhibits DIF), individuals with similar overall levels of the construct in question have different probabilities for choosing given response options, resulting from group membership. DIF has important implications for identifying out-group bias (e.g., Mazor, Clauser, & Hambleton, 1992) and is one means of testing item invariance across groups.

There are several ways to assess item invariance, including covariance modeling (factor analytic approaches), Item Response Theory (IRT), or nonparametric statistics (see Teresi, 2006). Covariance modeling and IRT approaches require large samples to satisfy assumptions and achieve convergence (e.g., Embretson & Reise, 2000; French & Finch, 2006). Additionally, these parametric models result in inflated Type I error given larger samples and model misspecifications (Teresi, 2006). In contrast, nonparametric statistics (e.g., the Mantel-Haenszel statistic; MH) have fewer assumptions; although they are "...generally less powerful...they may identify the most problematic [items] with large effect sizes ..." (Teresi, 2006, p. S164). Accordingly, nonparametric statistics like the MH are ideal in early stage exploration of DIF, particularly when access to out-group members is limited.

Long-term meditators prove difficult to recruit, yielding relatively small samples even for privileged contacts with meditation centers (119 – Baer et al., 2008; 50 – Brown & Ryan, 2003). Recruitment difficulties have led some researchers to use the internet to examine properties of mindfulness scales (e.g., Kohls, Sauer, & Walach, 2009). Difficulties in recruitment necessitate means of exploring potential DIF that are capable of handling discrepancies in sample distributions and sizes while maintaining a careful balance between Type I and Type II error.

To explore the potential for group bias with consideration of these limitations, we used nonparameteric DIF analyses to examine online responses on a popular self-report mindfulness scale, the FFMQ, in samples of non-meditating students and non-student meditators.

2. Methods

2.1. Procedure

Undergraduates at a state university in the northeastern United States participated via an online survey for course credit. Additionally, an email was sent to meditation and Buddhist list servers (see Kohls et al., 2009). Participants who were willing forwarded the email to others with meditation experience (the 'snowballing' technique; e.g., Van Dam, Earleywine, & DiGiacomo, 2008). Responses were not associated with individually identifying information. Procedures were approved by the local investigational review board.

2.2. Data screening

The Internet survey model provides advantages and disadvantages relative to other methods of assessment (see Van Dam et al., 2008). Careful data screening is extremely important; fraudulent data and duplicate responses can threaten experimental integrity. Forty-four participants in the student sample (10.6%) indicated that they were either impaired or did not take the study seriously; these subjects were removed. Eighteen Interpersonal (IP) addresses appeared more than once. Incomplete data sets from a repeated IP address were deleted. Potential duplicate data sets were carefully screened; there were no discernable duplicates.

Thirty-five (12.4%) students reported meditation experience. Twenty of those provided evidence that their 'meditation' practice may have involved mindfulness components. Data from these individuals were removed; they were not representative of a non-meditating student sample and the style of meditation reported was too disparate from the meditator population. The remaining ten provided descriptions of 'meditation' dissimilar to mindfulness-based practice; their data were included in the student non-meditator population. The final number of student non-meditators who completed the questionnaire was 283, though only 263 were included in the analyses for the above reasons.

Sixty-four individuals responded from meditation list servers. Data from six individuals were deleted because they reported no regular meditation practice or they did not engage in mindfulness meditation. The final number of meditator participants was 58.

2.3. Participants

The student non-meditator sample was 51.7% male, with an average age of 18.9 years (SD = 1.4). The sample was largely Caucasian (76.8%), with Other (8.4%), Asian (6.4%), Hispanic (4.9%), and African American (3.4%). Over 95% had never read any books related to Buddhism, meditation, and/or mindfulness (here termed "Dharma" books), 3.8% reporting having read a few, 0.8% reading them as a low priority, and 0.4% reading them as a moderate priority. The majority of the sample was unfamiliar with the concept of mindfulness (61.2%).

The sample of meditators was 63.8% female, with an average age of 47.5 years (SD = 14.2). The sample was 73.2% Caucasian, 12.5% Other, 7.1% Hispanic, and 7.1% Asian. The majority of the sample had exposure to Dharma books, 34.5% reporting reading them as a high priority, 39.7% as a moderate priority, 10.3% as a low priority, 13.8% reporting having read a few, and 1.7% reporting never having read any. The entire sample was familiar with the concept of mindfulness and the majority (67.2%) belonged to a meditation group. Most of the sample (61.4%) had been meditating for more than 5 years, with 28.1% reporting a meditation history between 1 and 5 years, 5.3% between 6 and 12 months, and 5.3% between 1 and 6 months. Nearly the entire sample had a regular personal meditation practice, 46.6% reporting a daily practice, 27.6% reporting practice 3–5 times a week, 15.5% reporting practice 1–2 times a week, 6.9% reporting weekly practice, and 3.4% reporting practice monthly or less.

2.4. Measures

2.4.1. Mindfulness

Participants completed the 39-item version of the Five Facet Mindfulness Questionnaire (FFMQ; Baer et al., 2006). The FFMQ has good psychometric properties in students, community members, and meditators (Baer et al., 2006, 2008). Internal consistency of the FFMQ among non-meditators (Cronbach's $\alpha = 0.86$) and meditators was high (Cronbach's $\alpha = 0.95$).

2.5. Statistical methods

2.5.1. Differential item functioning

Several statistical approaches can identify DIF in polytomous items, but no one statistic is best in all situations (Mazor et al., 1992; Teresi, 2006). One of the most popular nonparametric methods is an extension of the Mantel-Haenszel (MH) statistic, the Mantel chi-square (Mazor et al., 1992). The Mantel chi-square is based on a group (2) x response options (5) contingency table for each item. The larger the chi-square, the more disparate the probability of response options across meditation groups at the same level of overall mindfulness (see Penfield, 2007b).

A potentially more robust measure of DIF, able to handle extreme deviations in proportions of responses better than alternatives (Penfield, 2007a), is the Liu-Agresti common log odds ratio (L-A LOR; Liu & Agresti, 1996). The L-A LOR relies on the log odds ratio of one group selecting a particular response option relative to the other group, stratified by overall level of the construct. Ratios typically range from −1 to 1; values can go beyond this range if DIF is large. Negative values indicate bias against the reference group (in this case, meditators) and positive values indicate bias against the focal group (students).

Cox's noncentrality parameter (Cox's *B*), an alternative measure of DIF, parallels the MH approach, but relies upon the hypergeometric mean (Penfield, 2007b). In the absence of DIF, the odds ratio of a given response option across groups for each level of the construct will approximate zero, suggesting no group difference. The value of Cox's *B* is the sum of odds ratios for all responses across all levels of the construct and ranges from −1 to 1. Negative values indicate bias against the reference group (meditators); positive values indicate bias against the focal group (students). See Camilli and Congdon, (1999) for details.

2.6. Current analytic approach

We examined group differences with chi-squares (categorical variables) and *t*-tests (continuous variables) corrected for unequal variance where applicable. Differences in overall FFMQ score as well as positively and negatively worded items were compared both within and between groups.

An item that displays DIF can lead researchers to believe incorrectly that groups differ on a construct (e.g., Van Dam, Earleywine, & Forsyth, 2009). Thus, power for detecting DIF is very important, but Type I errors that incorrectly flag items as problematic must be avoided. DIF analysis is more commonly associated with inflation of Type II than Type I error (see Fidalgo, Hashimoto, Bartram, & Muñiz, 2007; Mazor et al., 1992), particularly when the proportion of DIF exceeds 10% (Teresi, 2006). However, limited overlap of ability distributions (see Embretson & Reise, 2000) can lead to inaccurate item DIF identification with small samples, necessitating cautious item flagging rules and interpretations (see Zwick, 1990). In simulations with MH values set to $p < .05$, for unequal distributions on the construct and unequal sample sizes similar to the current study (200:50), power reaches .35 (for items exhibiting large DIF) and Type I error does not exceed .03 (Fidalgo et al., 2007). Despite inflated Type II error rates (Teresi, 2006), the Mantel Chi-Square, L-A LOR, and Cox's *B* statistics were simultaneously Bonferroni corrected to minimize the possibility of Type I error. Given inflated Type II error and stringent significance criteria, the current statistical approach should only identify items showing large DIF (Fidalgo et al., 2007; Mazor et al., 1992; Teresi, 2006).The impact of multiple DIF items can be assessed with Differential Test Functioning (DTF). DTF concerns the additive effect of item DIF across all items on a test. Tests may exhibit biased functioning if more than 25% of items show moderate to large DIF (Penfield & Algina, 2006). The weighted v^2 statistic indicates DIF effects across a set of items (Penfield, 2007b), and is less susceptible than alternatives to violations of unequal variance (Penfield & Algina, 2006). All DIF and DTF statistics were computed using DIFAS 4.0 (Penfield, 2007b). All other statistics were computed using SPSS 16.0.

3. Results

3.1. Demographics

Meditators and non-mediators did not differ in ethnicity, p's > 0.5, but did differ in gender and age. Significantly more meditators were female (63.8%) relative to non-meditators (48.3%), χ^2 (1, 321) = 4.57, $p < 0.05$. Meditators (M = 47.48, SD = 14.2) were also significantly older than non-meditators (M = 18.9, SD = 1.4), $t(57.2) = 15.3$, $p < 0.001$. The correlation between age and group membership was extremely high, $r = .874$, $p < .01$.

3.2. FFMQ and demographics

Meditators exhibited higher scores on the FFMQ (M = 144.2, SD = 20.9) than students (M = 126.3, SD = 13.8), $t(68.4) = 6.22$, $p < 0.001$, $d = 1.01$. The only demographic variable related to the overall FFMQ score in meditators was meditation history, r = 0.52, $p < 0.01$. Although not significant, students familiar with mindfulness (M = 128.4, SD = 14.3) had higher scores on the total FFMQ than students not familiar with mindfulness (M = 125.05, SD = 13.4), $t(260) = 1.93$, $p = 0.055$, $d = 0.25$.

3.3. Differences in positive and negative items

In meditators, the relation between positively worded and negatively worded items ($r = .645$, $p < .01$) was significantly higher than in students ($r = .264$, $p < .01$; $Z\Delta r = 3.34$, $p < .001$). In meditators, meditation history was correlated with both negatively worded ($r = .453$, $p < .01$) and positively worded items ($r = .481$, $p < .01$). Age correlated significantly with endorsement of negatively worded items ($r = .332$, $p < .05$), but not positively worded items ($r = .245$, $p = .064$). Age was also significantly correlated in meditators with meditation history, $r = .483$, $p < .01$. Age did not covary with negatively worded ($r = .040$) or positively worded items ($r = .090$) in non-meditators.

There was a significant main effect of wording; across groups, negatively worded items (M = 3.42, SD = 0.55) had a higher item mean than positively worded items (M = 3.25, SD = 0.51), $F(1, 319) = 5.14$, $p < 0.05$, $d = 0.32$. More importantly, group and wording interacted significantly, $F(1, 319) = 16.95$, $p < 0.001$. Paired sample *t*-tests with Bonferroni corrections indicated no significant difference between negative items and positive items on the FFMQ in meditators. In students, however, negative items (reverse-scored) had a significantly higher mean than positive items (see Table 1). Meditators endorsed responses indicating that positive and negative items were equally true of them. In contrast, student non-meditators gave significantly lower ratings to negative items than positive ones.

To examine the contributions of group to the main effect of wording, independent sample *t*-tests with appropriate Bonferroni corrections were conducted between groups. The majority of the effect was represented by group differences in positively worded items; meditators had significantly higher mean item endorsement of positively worded items than students, $t(69.8) = 7.90$, $p < 0.05$, $d = 1.25$. Similarly, but with half the effect size, meditators had significantly higher mean item endorsement of negatively worded items than students, $t(73.9) = 3.37$, $p < 0.05$, $d = 0.52$. See Table 1 for descriptive statistics.

3.4. Differential item and test functioning

Despite simultaneous Bonferroni correction across all 39 FFMQ items by 3 test statistics ($p < .05$ corrected to $p < .00043$), 18 items showed converging evidence of DIF across all three tests (see Table 2). Regarding the items that displayed DIF, all negatively worded and reverse-scored items showed bias against meditators, while all positively worded items showed bias against students (based on the directionality of the L-A LOR and Cox's *B* statistics). Additionally, 6 other items from the FFMQ met Penfield's (2007a) criteria for large DIF (|L−A LOR| > 0.64) despite not meeting criteria at

Table 1
Comparison of positively and negatively worded mindfulness items.

Group	Positively worded items	Negatively worded items	Within group statistics	
	M (SD)	*M* (SD)	*t* (df)	*d*
Meditators ($n = 58$)	3.73 (0.56)	3.66 (0.62)	1.10 (57)	0.12
Students ($n = 263$)	3.12 (0.40)	3.37 (0.50)	7.40* (262)	0.55

FFMQ = Five Faceted Mindfulness Questionnaire, range: 1 (Never/Very Rarely True) to 5 (Very Often/Always True).
d = Effect size measured in Cohen's *d*; *t* = *t*-test value.
* $p < 0.05$ with Bonferroni corrections.

Table 2
Items meeting Bonferroni adjusted criteria for differential item functioning.

Subscale	Item	Mantel χ^2	L-A LOR[a]	Cox's *B*[b]
Observe	FFMQ 1	21.67*	−1.38*	−0.83*
	FFMQ 11	20.03*	−1.26*	−0.73*
	FFMQ 15	21.74*	−1.44*	−0.84*
	FFMQ 31	32.14*	−2.01*	−0.85*
	FFMQ 36	18.72*	−1.41*	−0.77*
Describe	FFMQ 2	17.38*	−1.25*	−0.67*
	FFMQ 12[R]	16.45*	1.77*	0.94*
Nonjudge	FFMQ 14[R]	17.31*	1.19*	0.70*
	FFMQ 10[R]	12.48*	1.07*	0.69*
	FFMQ 17[R]	20.35*	1.21*	0.77*
	FFMQ 25[R]	21.45*	1.40*	0.93*
	FFMQ 35[R]	24.64*	1.58*	0.88*
	FFMQ 39[R]	18.21*	1.25*	0.77*
Nonreact	FFMQ 29	18.44*	−1.30*	−0.88*
	FFMQ 33	16.85*	−1.21*	−0.71*
Act w/awareness	FFMQ 5[R]	22.82*	1.58*	0.85*
	FFMQ 13[R]	27.50*	1.77*	0. 94*
	FFMQ 34[R]	14.17*	1.23*	0.73*

N.B. Penfield (2007a) suggests that |L-A LOR| > 0.64 indicates large DIF.
* Significant at Bonferroni adjusted $p < .05$; Adjusted $\chi^2 = 12.41$, Adjusted $Z = 3.33$.
[a] Liu-Agresti Common Log Odds Ratio; negative values indicate bias against meditators, positive values indicate bias against students; Significance test based on conversion to a standardized distribution where L-A LOR = 1.06 approximates $Z = 3.33$.
[b] Cox's Noncentrality Parameter; negative values indicate bias against meditators, positive values indicate bias against students; Significance test based on conversion to a standardized distribution where Cox's $B = 0.61$ approximates $Z = 3.33$.
[R] Indicates that item is negatively worded and reverse-scored.

Table 3
Items not meeting Bonferroni adjusted criteria for all test statistics.

Subscale	Item	Mantel χ^2	L-A LOR[a]	Cox's *B*[b]
Describe	FFMQ 7	10.93	−1.14*	−0.60
	FMQ 27	11.99	−1.06	0.59*
	FFMQ 37	12.18	−1.06*	−0.59*
Nonreact	FFMQ 24	11.99	−0.98	−0.57*
Act w/awareness	FFMQ 8[R]	11.19	1.07*	0.62*
	FFMQ 38[R]	12.00	1.11*	0.69*

N.B. Penfield (2007a) suggests that |L-A LOR| > 0.64 indicates large DIF.
* Significant at Bonferroni adjusted $p < .05$; Adjusted $\chi^2 = 12.41$, Adjusted $Z = 3.33$.
[a] Liu-Agresti Common Log Odds Ratio; negative values indicate bias against meditators, positive values indicate bias against students; Significance test based on conversion to a standardized distribution where L-A LOR = 1.06 approximates $Z = 3.33$.
[b] Cox's Noncentrality Parameter; negative values indicate bias against meditators, positive values indicate bias against students; Significance test based on conversion to a standardized distribution where Cox's $B = 0.61$ approximates $Z = 3.33$.
[R] Indicates that item is negatively worded and reverse-scored.

the stringent Bonferroni corrected alpha level (see Table 3). These additional items exhibited the same directionality pattern of bias.

A test with 25% or more of the items indicating moderate to large DIF exhibits DTF (Penfield & Algina, 2006). In the current study, 46% of the items displayed large DIF (see Penfield, 2007a) at a stringent Bonferroni cut-off. The FFMQ exhibited strong DTF, weighted $v^2 = 1.06$ (see Penfield & Algina, 2006). The weighted $v^2 = 1.06$ corresponds to $Z = 4.04$, $p < .00003$.

4. Discussion

Experienced meditators received higher scores on the FFMQ than non-meditators, as found in previous work (e.g., Baer et al., 2008). Scores also increased with meditation practice. In the student non-meditator sample, familiarity with mindfulness increased scores by ¼ standard deviation, perhaps indicating greater familiarity with the terms used in the items. Greater familiarity with mindfulness might increase perceptions of its value and create demand characteristics contributing to responses on self-report questionnaires of mindfulness (Grossman, 2008). Alternatively, this result could actually indicate greater mindfulness in this sample because of exposure, practice, or both.

Differential item functioning (DIF) analyses suggest that comparisons between meditators and non-meditators may be problematic. Eighteen items, representing every factor, showed strong evidence of DIF, even under limited power conditions (see Teresi, 2006). Further, an additional 6 items provided evidence of DIF, though not at the specified Bonferroni corrected alpha level. The pattern of DIF suggested that the direction of the potential for bias related to the wording of items. On negatively worded items that showed DIF, meditators scored lower in mindfulness than non-meditators who were comparably mindful (based on total FFMQ score). On positively worded items that showed DIF, meditators scored higher in mindfulness than non-meditators who were comparably mindful.

The correlation between positively and negatively worded items was substantially larger in meditators than students. These positively and negatively worded items may be tapping a single construct in meditators. Student responses were less correlated, perhaps because of more error or the presence of different demand characteristics. The significant correlation of both negatively and positively worded items with meditation history in meditators may support the notion that the increased meditation practice is associated with improved meta-consciousness generally and recognition of attention lapses specifically. Alternatively, the DIF pattern may result from differential demand characteristics. Items showing the largest bias against students (e.g., 12 – "It's hard for me to find the words to describe what I'm thinking", 13 – "I am easily distracted"; see Table 2), if highly endorsed, might *suggest* levels of attention or awareness low enough to represent clinical problems. In contrast, items showing the largest bias against meditators may only appear valuable to those trained in mindfulness. For example, items 15 – "I pay attention to sensations, such as the wind in my hair or sun on my face", and 31 – "I notice visual elements in art or nature, such as colors, shapes, textures, or patterns of light and shadow"(see Table 2), may seem trivial to non-meditators but essential to meditators (Grossman, 2008). These item characteristics may create response sets subject to general (negatively worded items) and group-specific (positively worded items) demand characteristics (e.g., Rosnow, 2002) that might explain the pattern of DIF revealed in these data.

Many studies of DIF reveal potential bias on an item or two that researchers can easily remove from the scale (e.g., Van Dam et al., 2009). The current results present a unique challenge to the construct validity of the scale. DIF was sufficiently pervasive and effects sufficiently large that the entire questionnaire met criteria for differential test functioning (DTF). Despite good classical psychometric properties, the FFMQ functions differently in meditators and non-meditators. These findings question the appropriateness of comparisons between meditators and non-meditators using self-report mindfulness scales. Due to demand characteristics and inherently limited awareness of attentional lapses (Schooler, 2002), the validity of the FFMQ as a pre-post measure following mindfulness practices also may prove problematic.

4.1. Limitations

While meditators did not receive any direct reward for participating, students were given course credit; potentially creating differential demand characteristics regarding "volunteer subject" effects (see Rosnow, 2002). The comparison of groups with varying levels of voluntary participation (e.g., altruistic vs. course credit incentives) represents an abstruse impact on variable demand characteristics that clearly requires further experimental investigation. Similar methodological limitations are not uncommon in research with meditators (e.g., Kohls et al., 2009).

Because meditators were older and potentially more educated than non-meditators, the DIF results may stem from age or education rather than meditator status. The extremely high correlation between group and age prevented us from including it as a covariate. However, age did not covary with FFMQ total score, while meditation history did. Additionally, endorsement of positively and negatively worded items had larger correlations with meditation history than age, suggesting that the correlation between age and positively worded items may have been an artifact of meditation history. Previous work suggests that age and education may have relatively minor effects on measures of mindfulness (Baer et al., 2008), but the possibility that DIF is a result of age or education cannot be negated. Differences in gender distributions could also present a potential limitation, though relative to other comparisons of mindfulness across samples (e.g., Baer et al., 2008), our study represents a closer approximation to equal gender distribution.

Assessing responses in small samples with unequal ability distributions warrants cautious interpretation of DIF identification (e.g., Zwick, 1990). Limited group overlap of response distributions might have contributed to directionality and patterns of bias. However, simulations suggest the current approach produced valid results (e.g., Fidalgo et al., 2007). The conservative item flagging rules decrease potentially invalid DIF identification (e.g., Teresi, 2006). Further exploration of potential item bias with larger samples and alternative statistical approaches is warranted. Despite potential limitations, the results represent a serious challenge to construct validity in mindfulness self-report.

5. Conclusions

While meditators scored higher than non-meditators, supporting the validity of the FFMQ, the functionality of the items was different across these two groups. Non-meditators showed a significant difference between items that ask about mindfulness and those that ask about the absence of mindfulness; meditators did not. These results raise an important question: Does the absence of the opposite of a quality indicate its presence? The presence of qualities negatively related to mindfulness may help identify whether or not mindfulness *could* be present, but may do little to indicate whether it *is* present. Considerations of mindfulness present and mindfulness absent items require careful attention in studies using self-report methods to measure mindfulness.

The fact that DTF is present may suggest that self-report indices of mindfulness may not be appropriate for comparisons of meditators and non-meditators nor for pre-post measures following meditation practice. Even within populations, there may be limitations to self-report of such a complex construct associated with inherent difficulties in measurement (see Grossman, 2008; Schooler, 2002). Meditators, as volunteers, may attempt to be more objective and accepting of their self-representation (e.g., the "good subject", Rosnow, 2002), paradoxically something that their meditative practice may instruct them to value (Grossman, 2008). Separating demand characteristics relative to ability represents an epistemological problem for assessing mindfulness in meditators.

More evaluation of the qualitative aspects associated with mindfulness practice, changes in meta-consciousness, and careful examination of the numerous facets typically associated with the traditional Buddhist notion of mindfulness could prove useful in converging upon ways of tapping the construct while limiting bias.

Acknowledgements

The authors thank Paul Grossman and Ron Hambleton for comments on earlier drafts of this manuscript.

References

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., et al. (2008). Construct validity of the five facet mindfulness questionnaire in meditating and nonmeditating samples. *Assessment, 15*(3), 329–342.

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.

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics, 24*(4), 323–341.

DiStefano, C., & Motl, R. W. (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.

Fidalgo, Á. M., Hashimoto, K., Bartram, D., & Muñiz, J. (2007). Empirical Bayes versus standard Mantel-Haenszel statistics for detecting differential item functioning under small sample conditions. *The Journal of Experimental Education, 75*, 293–314.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling, 13*, 378–402.

Grossman, P. (2008). On measuring mindfulness in psychosomatic and psychological research. *Journal of Psychosomatic Research, 64*, 405–408.

Kohls, N., Sauer, S., & Walach, H. (2009). Facets of mindfulness – Results of an online study investigating the Freiburg mindfulness inventory. *Personality and Individual Differences, 46*, 224–230.

Liu, I., & Agresti, A. (1996). Mantel-Haenszel type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics, 52*(4), 1223–1234.

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.

Penfield, R. D. (2007a). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education, 20*(3), 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*(4), 295–312.

Rosnow, R. L. (2002). The nature and role of demand characteristics in scientific inquiry. *Prevention and Treatment, 5*(1), 37.

Schooler, J. W. (2002). Re-representing consciousness: Dissociations between experience and meta-consciousness. *TRENDS in Cognitive Sciences, 6*(8), 339–344.
Teresi, J. A. (2006). Different approaches to differential item functioning in health applications. *Medical Care, 44*(11), S152–S170.
Van Dam, N. T., Earleywine, M., & DiGiacomo, G. (2008). Polydrug use, cannabis, and psychosis-like symptoms. *Human Psychopharmacology Clinical and Experimental, 23*(6), 475–485.
Van Dam, N. T., Earleywine, M., & Forsyth, J. P. (2009). Gender bias in the sixteen-item anxiety sensitivity index: An application of polytomous differential item functioning. *Journal of Anxiety Disorders, 23*, 256–259.
Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item function coincide? *Journal of Educational Statistics, 15*(3), 185–197.