Mindfulness
https://doi.org/10.1007/s12671-019-01105-x

ORIGINAL PAPER

Item Response Theory Analysis of the Five Facet Mindfulness Questionnaire and Its Short Forms

William E. Pelham III[1] · Oscar Gonzalez[2] · Stephen A. Metcalf[3] · Cady L. Whicker[3] · Emily A. Scherer[3] · Katie Witkiewitz[4] · Lisa A. Marsch[3] · David P. Mackinnon[1]

Abstract

Objectives The Five Facet Mindfulness Questionnaire (FFMQ) is a self-report measure of mindfulness with forms of several different lengths, including the FFMQ-39, FFMQ-24, and FFMQ-15. We use item response theory analysis to directly compare the functioning of these three forms.

Methods Data were drawn from a non-clinical Amazon Mechanical Turk study ($N = 522$) and studies of aftercare treatment of individuals with substance use disorders (combined $N = 454$). The item and test functioning of the three FFMQ forms were studied and compared.

Results All 39 items were strongly related to the facet latent variables, and the items discriminated over a similar range of the latent mindfulness constructs. Items provided more information in the low-to-medium range of latent mindfulness than in the high range. Scores in three of the five FFMQ-39 facets were unreliable when measuring individuals in the high range of latent mindfulness, resulting from ceiling effects in item responses. Reliability in the high range of mindfulness was further reduced in the FFMQ-24 and FFMQ-15, such that short forms may be ill-suited for applications that require reliable measurement in the high range.

Conclusions Results suggest the existing FFMQ item pool cannot be reduced without negatively affecting either overall reliability or the span of mindfulness over which reliability is assessed. Conditional test reliability curves and item functioning parameters can aid investigators in tailoring their choice of FFMQ form to the reliability they hope to achieve and to the range of latent mindfulness over which they must reliably measure.

Keywords Item response theory · Short form · Mindfulness

Mindfulness has become a topic of great interest in many areas of psychology (van Dam et al. 2018). To date, mindfulness has been most commonly measured via self-report questionnaires (van Dam et al. 2018), which are inexpensive and place low burden on participants. One mindfulness questionnaire that has been identified as promising (Park et al. 2013; Sauer et al. 2013) is the Five Facet Mindfulness Questionnaire (FFMQ-39; Baer et al. 2006). The FFMQ includes 39 items that together measure five different dimensions of mindfulness.

The original Five Facet Mindfulness Questionnaire (FFMQ-39) was designed to facilitate investigation of different dimensions of the mindfulness construct separately (Baer et al. 2006). The measure was developed by administering five existing mindfulness questionnaires in a sample of 613 undergraduate students and then including the items from all five scales (112 total items) in an exploratory factor analysis (Baer et al. 2006). Results indicated a five-factor solution, with the five factors corresponding to the FFMQ five *facets* of mindfulness: (1) acting with awareness, (2) describing, (3) nonjudging, (4) nonreactivity, and (5) observing. For each

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12671-019-01105-x) contains supplementary material, which is available to authorized users.

✉ William E. Pelham III
wpelham@asu.edu

[1] Department of Psychology, Arizona State University, Tempe, AZ 85281, USA

[2] Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[3] Department of Psychiatry, Dartmouth College, Lebanon, NH 03766, USA

[4] Center on Alcoholism, Substance Abuse, and Addictions, University of New Mexico, Albuquerque, NM 87131, USA

Published online: 02 March 2019

factor, the eight items with the highest loadings were retained to form the FFMQ-39, which was then subjected to confirmatory factor analysis on a new sample. For the nonreactivity scale, only seven items with sufficiently high factor loadings were identified, and hence, the total number of items is 39. Results indicated that a model with five correlated factors fit the data well. The psychometric structure of the FFMQ has since been replicated in many different samples (e.g., Bohlmeijer et al. 2011; Christopher et al. 2012), and the FFMQ-39 has become one of the most popular self-report measures of mindfulness.

Unfortunately, the inclusion of 39 items limits the practicality of the FFMQ-39 for rapid or repeated administration. When asking participants to complete the FFMQ as part of an extensive battery of questionnaires or on a daily basis (e.g., for ecological momentary assessment), investigators may prefer an abbreviated item set that can approximate the complete form. The use of shorter forms can increase response rates and response quality, but may also affect the reliability of scores and the relation of scores with other criteria. At least five different short forms of the FFMQ have been created (Baer et al. 2012; Bohlmeijer et al. 2011; Hou et al. 2014; Medvedev et al. 2018; Tran et al. 2013). The two oldest and most frequently used of these short forms are the FFMQ-15 and the FFMQ-24.

The FFMQ-15 (Baer et al. 2012) was originally created to track weekly changes in mindfulness over the duration of an 8-week course in mindfulness-based stress reduction. The investigators returned to data from the original FFMQ development (Baer et al. 2006) and retained for each facet the three items with the highest factor loadings in the exploratory factor analysis. The authors did not report a detailed psychometric evaluation of the 15-item form, since it was not the primary focus of their study. Recently, the FFMQ-15 was evaluated in a sample of 238 participants with recurrent major depressive disorder (Gu et al. 2016). Results indicated that the FFMQ-15 exhibited psychometric structure, reliability, and sensitivity to change similar to that of the FFMQ-39.

The FFMQ-24 (Bohlmeijer et al. 2011) was created using data from samples recruited in the Netherlands. A group of 376 adults with clinically relevant symptoms of depression and anxiety completed the full FFMQ-39. Items were then evaluated for high factor loadings, minimal cross-loadings, low residual error correlations, and preserved content span (Marsh et al. 2005; Smith et al. 2000). Twenty-four items were chosen, and the resulting FFMQ-24 was validated on a separate sample of 146 adults with self-reported fibromyalgia. Results indicated that the FFMQ-24 exhibited psychometric structure, reliability, and sensitivity to change similar to that of the FFMQ-39.

Together, the FFMQ-39, FFMQ-24, and FFMQ-15 provide investigators three useful options for the measurement of self-reported mindfulness. However, there is limited information to guide the choice among them, such as how much reliability is sacrificed by dropping items, or how a score on either short form is comparable to score on the complete form. We are aware of only one psychometric comparison of the FFMQ-39 and the FFMQ-15 (Gu et al. 2016) and of no psychometric comparison of the FFMQ-24 and the FFMQ-15.

One approach that may help clarify the relative functioning of the FFMQ-39, FFMQ-24, and FFMQ-15 is item response theory (IRT). IRT comprises a family of latent variable models used to analyze discrete item responses and the resulting test scores. These models underlie many of the modern methods for scale development and high-stakes testing. Most psychologists are familiar with the methods of classical test theory for scale development and evaluation, such as test-total correlations, proportion endorsed on each item, and reliability coefficients (e.g., coefficient alpha). IRT complements classical test theory to permit comprehensive investigation of item properties, facilitate scale development by using item information functions, and enable the linkage of scores from participants who responded to different versions of the scale (de Ayala 2009; Edelen and Reeve 2007; Embretson and Reise 2000; Reise et al. 2005).

A key property of IRT models is that the reliability (or measurement precision) of a score varies as a function of the respondent's value on the latent construct being assessed. This contrasts with classical test theory approaches, in which reliability is defined by a single number (e.g., coefficient alpha) that applies to every respondent's score. In IRT, a score may be more reliable for individuals that are in the high range of the latent variable (denoted *theta*, or θ, in the IRT framework) than for individuals that are in the low range of the latent variable. When comparing forms of three different lengths (FFMQ-39, FFMQ-24, and FFMQ-15), an IRT analysis would enable the comparison of not only whether but also where the forms differ in score reliability. For example, it could be that the FFMQ-15 produces scores with similar reliability to those of the FFMQ-39 for participants that are near the latent mean of nonreactivity, but produces substantially less reliable scores for respondents in the upper or lower extremes of nonreactivity. Such information could be used to tailor the choice of FFMQ form to the specific research question and sample at hand.

IRT analysis also yields richer understanding of item properties. *Discrimination* parameters indicate how strongly the items relate to the latent variable being measured. *Severity* parameters indicate at what point along the latent variable continuum the items are differentiating respondents. These parameters can be combined to yield an item information curve, or visual depiction of how useful an item is in estimating test scores across the range of the latent construct. For example, it might be the case that item 3 of the FFMQ is only useful for distinguishing individuals in the low range of nonjudging, whereas item 17 is only useful for distinguishing those in the mid-to-high range of nonjudging. This type of information could be used to improve the scale (e.g., by writing new items that would be useful in a neglected range of the latent variable), to shorten the scale (e.g., by selecting a subset of items that efficiently reproduces the properties of the full scale), or simply to guide the scale's effective use (e.g., by indicating over what range of the latent variable the scale scores are reliable). In summary, item response theory analyses could provide valuable information about the absolute and relative psychometric properties of all three forms of the FFMQ.

Despite the promise of an IRT approach, there have been only two previous IRT analyses of the FFMQ (Medvedev et al. 2017; Medvedev et al. 2018). Both analyses used Rasch modeling, a type of item response theory model in which the discrimination parameters of all items are constrained to be equal (Embretson and Reise 2000). In practice, it is difficult for all the items in a scale to meet this constraint, so part of a Rasch analysis entails finding a subset of items that conforms to the assumptions of the Rasch model (see Andrich 2004 for a discussion of the philosophical and measurement issues surrounding this practice). In the first study, Medvedev et al. (2017) conducted a Rasch analysis of the FFMQ-39 in a sample of 296 university students and community members in New Zealand. Medvedev et al. modified the FFMQ as needed to satisfy the Rasch assumptions, resulting in a new version: the FFMQ-37. The authors then provided tables to convert a total score on each facet of the FFMQ-37 into a more continuous, interval measurement scale. Second, Medvedev et al. (2018) applied the same Rasch procedures to four different short forms, including the FFMQ-24 and FFMQ-15 ($N = 400$, subsuming the sample in Medvedev et al. 2017). After modifying each scale as needed to satisfy the Rasch assumptions, the authors determined that a modified version of the FFMQ-24, the FFMQ-18, exhibited the best psychometric properties. Thus, they recommend the use of their FFMQ-18 when choosing among short forms and the use of their FFMQ-37 when maximum reliability is needed.

These two studies (Medvedev et al. 2017, 2018) illustrate the potential of using the IRT to evaluate the FFMQ. Their findings yielded increased precision of measurement without any change to the original item response format, plus a shorter short form with superior psychometric functioning. However, these studies were limited in several ways. First, neither study included a clinical sample, with which the FFMQ is commonly used (e.g., in trials of mindfulness-based stress reduction). Second, both studies produced and then evaluated modified versions of the FFMQ (i.e., the FFMQ-37 and FFMQ-18), rather than studying the properties of the existing forms investigators are already using. Third, the studies did not directly compare the reliability and score recovery of existing long and short forms, and thus provided limited guidance to investigators seeking to choose among them. Finally, due to its strict requirements, the Rasch approach does not yield some of the benefits of IRT discussed above, such as evaluation of how each item independently functions. A more flexible IRT analysis that compares existing forms as they are may complement the work of Medvedev and colleagues in understanding the psychometric functioning of the FFMQ.

The purpose of the current study was to compare the functioning of the FFMQ-39, FFMQ-24, and FFMQ-15 using item response theory analysis. Analyses were expected to clarify the conditions under which an investigator might choose one of these forms over the others. We studied the item and test functioning of each FFMQ facet and compared these properties across the 39-, 24-, and 15-item forms. Finally, we evaluated the consistency of findings across two large datasets: adults recruited via Amazon Mechanical Turk ($N = 522$) and individuals receiving aftercare treatment for substance use disorder ($N = 456$).

Method

Participants

Primary Sample (MTurk) Five hundred twenty-two participants completed the FFMQ as part of a larger battery of questionnaires on Amazon Mechanical Turk (MTurk). The participants completed 21 different self-report measures of constructs related to self-regulation as part of a larger, multisite project aiming to develop an "ontology" of measures of impulsivity, time perspective, grit, mindfulness, sensation seeking, willpower, and related concepts. Only responses passing quality checks were retained for analysis (see Eisenberg et al. 2018 for a description of these checks and of the larger study protocol). Participants were adults between 20 and 59 years old who lived in the USA. Mean age was 34 years old (SD = 8), 51% of participants were female, 86% of participants were Caucasian, and 44% of participants were at least college-educated. There were no missing data on the FFMQ items.

Replication Sample (Clinical) Two smaller, secondary samples were combined and used to replicate the findings observed on the primary sample. Both were drawn from randomized, controlled trials of mindfulness-based relapse prevention (Bowen et al. 2011) for individuals with substance use disorders. In each study, participants had recently completed inpatient or intensive outpatient treatment for substance use disorders and were randomized to different aftercare conditions. The first sample (Bowen et al. 2009) comprised 168 adults who were randomized to mindfulness-based relapse prevention or treatment as usual with the following characteristics: mean age of 41 years old (SD = 10), 64% male, 54% non-Hispanic White, 30% African American, 15% Native American, 5% Hispanic or Latino/a, 41% unemployed, and 72% having a high school degree. The second sample (Bowen et al. 2014) comprised 286 adults who were randomized to mindfulness-based relapse prevention, relapse prevention, or treatment as usual with the following characteristics: mean age of 38 years old (SD = 11), 75% male, 53% non-Hispanic White, 25% African American, 6% Native American, 7% Hispanic or Latino/a, 66% unemployed, and 66% having a high school degree. Only data collected at baseline (i.e., prior to randomization) are used in this report, so the participants had not yet been exposed to the mindfulness-based intervention material that might be expected to affect their responses (Quaglia et al. 2016). The two samples were combined to yield a sizeable validation dataset despite missing data. Most participants (324 of 454 cases, or 71%) had complete data, with data missing sporadically across items (mean of 2.8% of item responses missing, maximum of 4.0%).

Procedures

In the primary sample (MTurk), participants completed the FFMQ online as one component of a larger battery of questionnaire and cognitive tasks measuring self-regulation (Eisenberg et al. 2018). The battery was delivered online via the Experiment Factory platform (Sochat et al. 2016). In the secondary sample (clinical), participants completed the FFMQ via a web-based survey platform (DatStat Illume, DatStat, Incorporated, Seattle, Washington) as part of the baseline intake battery, prior to randomization to treatment (Bowen et al. 2009, 2014).

Measures

FFMQ-39 The FFMQ-39 (Baer et al. 2006) consists of 39 items asking the individual to rate the extent to which a statement pertaining to mindfulness is applicable, on a scale from 1 (never or rarely true) to 5 (very often or always true). Nineteen of the 39 items are reverse-scored. These 19 items were reverse-scored prior to analysis, such that higher scores indicate higher mindfulness throughout this manuscript. Table 1 lists the item prompts and provides descriptive statistics in the primary sample. Seven of the items comprise the nonreactivity facet, and eight items comprise each of the observing, describing, acting with awareness, and nonjudging facets. Total scores for each facet are computed by summing the items after reverse scoring.

FFMQ-24 and FFMQ-15 The FFMQ-15 (Baer et al. 2012) and FFMQ-24 (Bohlmeijer et al. 2011) were not administered separately from the FFMQ-39 in this study. Instead, responses on these two short forms were reconstructed based on participants' responses to the complete, 39-item form. Table 1 shows which items are included on both the FFMQ-24 and FFMQ-15.

Data Analyses

All analyses were conducted first on the primary dataset (MTurk) and second on the replication dataset (clinical). The *mirt* package in R (Chalmer 2012) was used for all modeling. Our basic procedure mimics that described in Edwards' (2009) introduction to IRT; we direct readers to that and the following references for further background about IRT analysis and interpretation (de Ayala 2009; Edelen and Reeve 2007; Embretson and Reise 2000; Reise et al. 2005). Unidimensional IRT models were fit to each of the five facets of the FFMQ-39 (Baer et al. 2006): acting with awareness, describing, nonjudging, nonreactivity, and observing. We analyzed item responses using the graded response model (GRM; Samejima 1969), which is appropriate for items with ordered-categorical responses. The FFMQ items have five response options, so five item parameters were estimated for each item: one slope parameter and four threshold parameters (often called "severity" parameters; there are $k - 1$ thresholds for an item with k categories). The slope parameter is analogous to a factor loading and indicates the strength of relationship between the item and the latent variable. An item with a larger slope parameter has a stronger relationship with the latent variable and contributes more to the precise estimation of a participant's value on the latent variable. The threshold parameters correspond to the location of boundaries between two response options on an item. Since each FFMQ item has five response options, there are four boundaries, and thus four threshold parameters per item. The thresholds are on the same metric as the latent variable, so they can be interpreted as indicating at what value of the latent variable a participant has a 50% chance of endorsing that response category or a higher one. Lower threshold values indicate that the responses to the corresponding item separate those at lower values of the latent variable, and higher threshold values indicate that responses to the corresponding item separate those at higher values of the latent variable. Taken together, the positioning of each item's set of four thresholds indicates over what range of the latent variable that item is most useful.

Item and Test Information Functions Slope and threshold parameters may be difficult to interpret in isolation. To ease interpretation, the estimated parameters from the FFMQ-39 can be transformed into item information functions that indicate over what range of the latent variable each item is most useful. For example, an item might provide more precise estimation for people below the mean of the mindfulness facet latent variable than those above the mean. Item information functions are additive, so the test information function can be estimated to investigate the range of the latent variable in which the scale is most useful. Test information functions were computed separately for the FFMQ-39, FFMQ-24, and FFMQ-15 by summing information from only the items present on each form.

Table 1 Descriptive statistics for FFMQ items in primary sample

Facet	Item	On FFMQ-24	On FFMQ-15	Mean	SD	% per response value	Item label
Acting with awareness	5			3.42	1.08	0.06/0.15/0.27/0.37/0.15	When I do things, my mind wanders off and I am easily distracted
	8		✓	3.86	0.96	0.01/0.07/0.25/0.38/0.29	I do not pay attention to what I am doing because I am daydreaming, worrying, or otherwise distracted
	13			3.52	1.10	0.05/0.13/0.26/0.36/0.20	I am easily distracted
	18	✓		3.82	0.98	0.01/0.09/0.24/0.38/0.28	I find it difficult to stay focused on what's happening in the present
	23	✓		3.78	1.00	0.02/0.10/0.24/0.39/0.26	It seems I am 'running on automatic' without much awareness of what I am doing
	28	✓		4.03	0.90	0.01/0.05/0.21/0.39/0.35	I rush through activities without being really attentive to them
	34	✓	✓	3.82	0.94	0.01/0.07/0.28/0.38/0.27	I do jobs or tasks automatically without being aware of what I am doing
	38	✓	✓	3.86	0.98	0.02/0.07/0.23/0.39/0.29	I find myself doing things without paying attention
Describing	2	✓	✓	3.48	1.10	0.05/0.14/0.28/0.34/0.19	I am good at finding words to describe my feelings
	7	✓		3.67	1.02	0.02/0.11/0.26/0.38/0.23	I can easily put my beliefs, opinions, and expectations into words
	12	✓		3.74	1.08	0.04/0.10/0.20/0.39/0.26	It's hard for me to find the words to describe what I am thinking
	16		✓	3.68	1.11	0.04/0.13/0.21/0.36/0.26	I have trouble thinking of the right words to express how I feel about things
	22	✓		3.82	1.01	0.02/0.10/0.20/0.40/0.28	When I have a sensation in my body, it's difficult for me to describe it because I cannot find the right words
	27	✓	✓	3.41	1.10	0.06/0.14/0.29/0.35/0.16	Even when I am feeling terribly upset, I can find a way to put it into words
	32			3.27	1.10	0.06/0.20/0.29/0.33/0.13	My natural tendency is to put my experiences into words
	37			3.37	1.14	0.07/0.16/0.27/0.34/0.16	I can usually describe how I feel at the moment in considerable detail
Nonjudging	3			3.50	1.18	0.06/0.15/0.25/0.31/0.24	I criticize myself for having irrational or inappropriate emotions
	10	✓	✓	3.50	1.06	0.03/0.15/0.30/0.33/0.19	I tell myself I should not be feeling the way I am feeling
	14		✓	3.86	1.07	0.02/0.11/0.22/0.31/0.35	I believe some of my thoughts are abnormal or bad and I should not think that way
	17	✓		3.27	1.13	0.05/0.22/0.30/0.26/0.16	I make judgments about whether my thoughts are good or bad
	25	✓		3.57	1.05	0.02/0.15/0.26/0.35/0.20	I tell myself that I should not be thinking the way I am thinking
	30	✓	✓	3.75	1.06	0.02/0.12/0.21/0.36/0.28	I think some of my emotions are bad or inappropriate and I should not feel them
	35			3.65	1.09	0.02/0.15/0.28/0.28/0.28	When I have distressing thoughts or images, I judge myself as good or bad, depending what the thought or image is about
	39	✓		3.42	1.14	0.04/0.20/0.27/0.28/0.21	I disapprove of myself when I have irrational ideas
Nonreactivity	4			3.18	0.93	0.05/0.16/0.43/0.31/0.06	I perceive my feelings and emotions without having to react to them
	9	✓		3.24	0.94	0.05/0.13/0.41/0.34/0.07	I watch my feelings without getting lost in them
	19	✓	✓	3.16	0.99	0.07/0.16/0.39/0.32/0.07	When I have distressing thoughts or images, I "step back" and am aware of the thought or image without getting taken over by it
	21			3.40	0.94	0.03/0.12/0.37/0.37/0.11	In difficult situations, I can pause without immediately reacting
	24	✓		3.02	1.00	0.08/0.21/0.38/0.29/0.05	When I have distressing thoughts or images, I feel calm soon after
	29	✓	✓	3.16	0.92	0.05/0.14/0.46/0.28/0.06	When I have distressing thoughts or images, I am able just to notice them without reacting
	33	✓	✓	3.05	0.96	0.07/0.18/0.45/0.25/0.06	When I have distressing thoughts or images, I just notice them and let them go
Observing	1			2.84	1.04	0.11/0.26/0.37/0.22/0.05	When I am walking, I deliberately notice the sensations of my body moving
	6		✓	3.20	1.10	0.06/0.20/0.34/0.26/0.13	When I take a shower or bath, I stay alert to the sensations of water on my body
	11		✓	2.92	1.11	0.12/0.22/0.35/0.24/0.07	I notice how foods and drinks affect my thoughts, bodily sensations, and emotions
	15	✓	✓	3.35	1.02	0.04/0.14/0.36/0.32/0.13	I pay attention to sensations, such as the wind in my hair or sun on my face
	20	✓		3.48	1.02	0.04/0.11/0.35/0.33/0.17	I pay attention to sounds, such as clocks ticking, birds chirping, or cars passing
	26	✓		3.86	0.96	0.02/0.05/0.25/0.39/0.28	I notice the smells and aromas of things
	31	✓		3.58	1.01	0.03/0.09/0.33/0.36/0.19	I notice visual elements in art or nature, such as colors, shapes, textures, or patterns of light and shadow
	36			3.47	0.92	0.02/0.12/0.34/0.41/0.11	I pay attention to how my emotions affect my thoughts and behavior

"% per response value" indicates the percentage of participants responding in the first, second, third, fourth, and fifth category on the response scale. Items were reverse-scored as indicated prior to calculating descriptive statistics. $N = 522$ for all items

For example, there are only three items on the observing facet of the FFMQ-15. The item parameters estimated on the FFMQ-39 (i.e., those in Table 2) for those three items were used to estimate a three-item test information function for the observing facet of the FFMQ-15. Test information functions can be transformed into the standard error of measurement (SEM = 1/√Information) and score precision, or reliability (Reliability = 1 − SEM^2), so they can indicate over what range of the latent variable the scale produces reliable scores.

Score Linking When two scales contain overlapping items, the factor scores they produce can be linked in the IRT framework. Summed scores on the FFMQ-39 observing facet (range = 8–40) and on the FFMQ-15 observing facet (range = 3–15) are not directly comparable (cf. Hambleton and Swaminathan 1985), but the constituent item responses can be scored using the IRT item parameters to produce factor scores on the same scale (i.e., theta). Thus, score linking analyses can provide a sense of how well the FFMQ-24 and FFMQ-15 recover the latent facet scores that would have been produced using the full FFMQ-39, making scores across all of these forms comparable. To prevent overfitting, we split the sample in half, fit an IRT model to the FFMQ-39 facet in the first half of the sample, and then estimated *expected* a posteriori (EAP; Thissen and Wainer 2001) scores for both the FFMQ-39 facet and the short-form facet in the second half of the sample. We then calculated (a) the correlation of the factor scores produced by the short-form facet with those produced by the FFMQ-39 facet and (b) the mean square error of the factor score of the short-form facet.

Results

We report results from the primary sample first, and then compare them to results from the replication sample.

Item Functioning

Estimated item slopes and thresholds are reported in Table 2. Fig. 1 shows the estimated slope parameters of each item. Across 39 items, slopes in the primary sample ranged from 1.40 to 4.60 with a median value of 2.58, indicating positive and reasonably strong associations of responses on the items with the latent variable. Slopes were lower on the nonreactivity (mean = 1.96) and observing facets (mean = 2.19) than on the other facets (means = 3.23 [describing], 3.11 [nonjudging], and 2.88 [acting with awareness]). The most discriminating items on each facet were as follows: item number 38 for acting with awareness (“I find myself doing things without paying attention”), item number 7 for describing (“I can easily put my beliefs, opinions, and expectations into words”), item number 30 for nonjudging (“I think some of my emotions are bad or inappropriate and I shouldn’t feel them”), item number 29 for nonreactivity (“When I have distressing thoughts or images I am able to just notice them without reacting”), and item number 15 for observing (“I pay attention to sensations, such as the wind in my hair or sun on my face”).

Figure 2 shows the estimated thresholds for each item. Across 39 items, thresholds ranged from − 3.29 to 2.67, although most of the thresholds were negative or close to the mean of the latent variable. The third thresholds (“b3”) were typically close to the mean of the latent variable (indicated by the vertical dashed line in Fig. 2). Thus, for most items, someone who is average on the latent facet of mindfulness will have about 50% probability of responding to positively valenced statements with either *often true* (response value of 4) or *very true/always true* (response value of 5). In other words, many of the items in the FFMQ-39 did not discriminate participants high on the facet of mindfulness. The nonreactivity facet was the only one to include multiple items with high (> 2 SDs) threshold values.

Figure 3 (top row) displays the item information curves for each of the five facets. Within each facet, the items conveyed information over a similar range of the latent variable, but varied in the amount of information conveyed at each point within this range. For acting with awareness, describing, and nonjudging, none of the items conveyed much information above 1.5 standard deviations from the mean of the latent variable. For observing and nonreactivity, none of the items conveyed much information above 2.0 standard deviations from the mean of the latent variable. Across all facets, items conveyed more information in the low range of the latent variable than in the high range.

We next compared the item information functions of the items retained in (versus excluded from) each of the two short forms, the FFMQ-24 and FFMQ-15. This can be achieved by cross-referencing Table 1 and Fig. 3, but Figures S2 and S3 facilitate more direct comparison of the information functions of included versus excluded items. The FFMQ-24 had generally retained the items with higher slopes, all of which discriminated over a very similar range of latent mindfulness. Item number 1 on the observing facet (“When I’m walking, I deliberately notice the sensations of my body moving”) was not included in the FFMQ-24, but it appears it might have improved discrimination in the upper range of the latent variable (Figure S2). Like the FFMQ-24, the FFMQ-15 had generally retained the items with higher slopes, all of which discriminated over a similar range of latent mindfulness. Item number 7 on the describing facet (“I can easily put my beliefs, opinions, and expectations into words”) was the most discriminating item on that facet, but was not included in the FFMQ-15 (Figure S3). In summary, inspection of the item information curves revealed only minor opportunities for strategic improvement of the FFMQ-

Table 2 Estimated item parameters for FFMQ-39

Facet	Item	Primary sample (MTurk)					Replication sample (clinical)				
		a	b1	b2	b3	b4	a	b1	b2	b3	b4
Acting with awareness	5	2.48	−1.95	−1.04	−0.12	1.26	2.20	−1.34	−0.50	0.67	1.68
	8	2.78	−2.62	−1.63	−0.51	0.66	2.34	−2.05	−1.14	−0.12	0.86
	13	2.98	−1.91	−1.06	−0.17	0.97	2.32	−1.48	−0.68	0.32	1.36
	18	2.41	−2.74	−1.55	−0.52	0.70	2.24	−1.86	−0.99	0.13	0.97
	23	2.91	−2.47	−1.42	−0.47	0.71	2.00	−2.43	−1.44	−0.04	0.96
	28	2.80	−2.97	−1.88	−0.72	0.44	1.71	−2.56	−1.45	0.20	1.54
	34	3.04	−2.67	−1.64	−0.44	0.69	1.26	−3.12	−1.83	−0.14	1.24
	38	3.61	−2.29	−1.49	−0.51	0.59	2.37	−2.23	−1.11	0.05	1.02
Describing	2	3.83	−1.84	−0.96	−0.11	0.91	2.29	−2.06	−1.18	−0.04	1.21
	7	4.31	−2.14	−1.15	−0.28	0.78	2.57	−1.96	−1.08	−0.10	1.00
	12	3.61	−1.92	−1.16	−0.44	0.67	1.26	−2.98	−1.62	−0.03	1.41
	16	3.52	−1.94	−1.05	−0.37	0.68	1.42	−2.50	−1.22	0.17	1.44
	22	2.41	−2.53	−1.42	−0.55	0.70	1.20	−3.34	−1.94	−0.15	1.46
	27	2.76	−1.88	−1.05	−0.09	1.12	1.92	−2.28	−1.10	0.08	1.26
	32	2.04	−2.05	−0.82	0.15	1.46	2.15	−1.89	−0.87	0.26	1.32
	37	3.54	−1.66	−0.83	−0.03	1.07	2.28	−1.96	−0.77	0.27	1.38
Nonjudging	3	2.65	−1.82	−0.98	−0.16	0.84	1.57	−2.05	−0.79	0.63	1.65
	10	2.79	−2.12	−1.09	−0.09	1.00	1.55	−2.44	−0.97	0.70	1.82
	14	3.53	−2.28	−1.26	−0.46	0.40	1.71	−2.24	−1.24	0.16	1.11
	17	2.23	−2.07	−0.82	0.17	1.23	1.46	−2.11	−0.87	0.81	1.97
	25	3.96	−2.11	−1.03	−0.19	0.89	2.14	−2.14	−1.21	0.31	1.21
	30	4.60	−2.06	−1.11	−0.41	0.59	2.44	−2.13	−1.22	0.16	1.09
	35	2.55	−2.51	−1.19	−0.22	0.67	1.98	−2.03	−1.05	0.24	1.12
	39	2.58	−2.13	−0.89	−0.02	0.94	1.66	−1.88	−0.97	0.34	1.31
Nonreactivity	4	1.55	−2.56	−1.23	0.44	2.34	1.27	−2.64	−1.29	0.70	2.26
	9	1.81	−2.31	−1.21	0.30	2.11	1.43	−2.70	−1.26	0.56	2.29
	19	1.59	−2.28	−1.11	0.40	2.26	1.57	−2.22	−1.14	0.41	1.75
	21	1.83	−2.56	−1.37	0.07	1.73	1.72	−2.45	−1.46	0.10	1.69
	24	1.40	−2.29	−0.90	0.64	2.67	1.43	−1.88	−0.68	0.89	2.50
	29	3.34	−1.84	−0.93	0.40	1.76	1.88	−2.03	−0.95	0.62	2.06
	33	2.17	−1.91	−0.86	0.61	2.08	1.37	−2.21	−0.89	0.89	2.51
Observing	1	2.32	−1.55	−0.45	0.74	2.11	1.58	−1.12	−0.41	0.73	1.60
	6	2.25	−1.93	−0.79	0.31	1.40	2.04	−1.38	−0.65	0.18	1.15
	11	1.63	−1.68	−0.58	0.70	2.19	1.21	−1.85	−0.65	0.54	2.07
	15	4.04	−1.82	−0.96	0.10	1.22	2.76	−1.39	−0.75	0.20	1.22
	20	2.26	−2.16	−1.25	−0.02	1.20	1.56	−1.96	−1.10	0.12	1.12
	26	1.75	−2.86	−1.97	−0.61	0.80	1.80	−2.20	−1.69	−0.28	0.87
	31	1.86	−2.50	−1.54	−0.15	1.18	1.76	−1.90	−1.11	−0.01	1.00
	36	1.43	−3.29	−1.63	−0.07	1.92	1.28	−2.46	−1.37	0.15	1.76

“a” indicates item slope and “b1” through “b4” indicate item thresholds. Model fit separately for each of five facets

24 and FFMQ-15, and suggests the existing item selection is similar (although not identical) to the items that would be selected by an IRT analysis.

Results for the FFMQ-39 in the replication sample mirrored those from the primary sample. Figure 1 compares the slopes of each item in the primary and replication samples. Across all items, slopes in the replication sample ranged from 1.20 to 2.76 with a median value of 1.72, suggesting reasonable associations between the items and the latent variable. For three of the five facets (describing, nonjudging, and nonreactivity), the top two most discriminating items were the same in the primary and replication samples, but

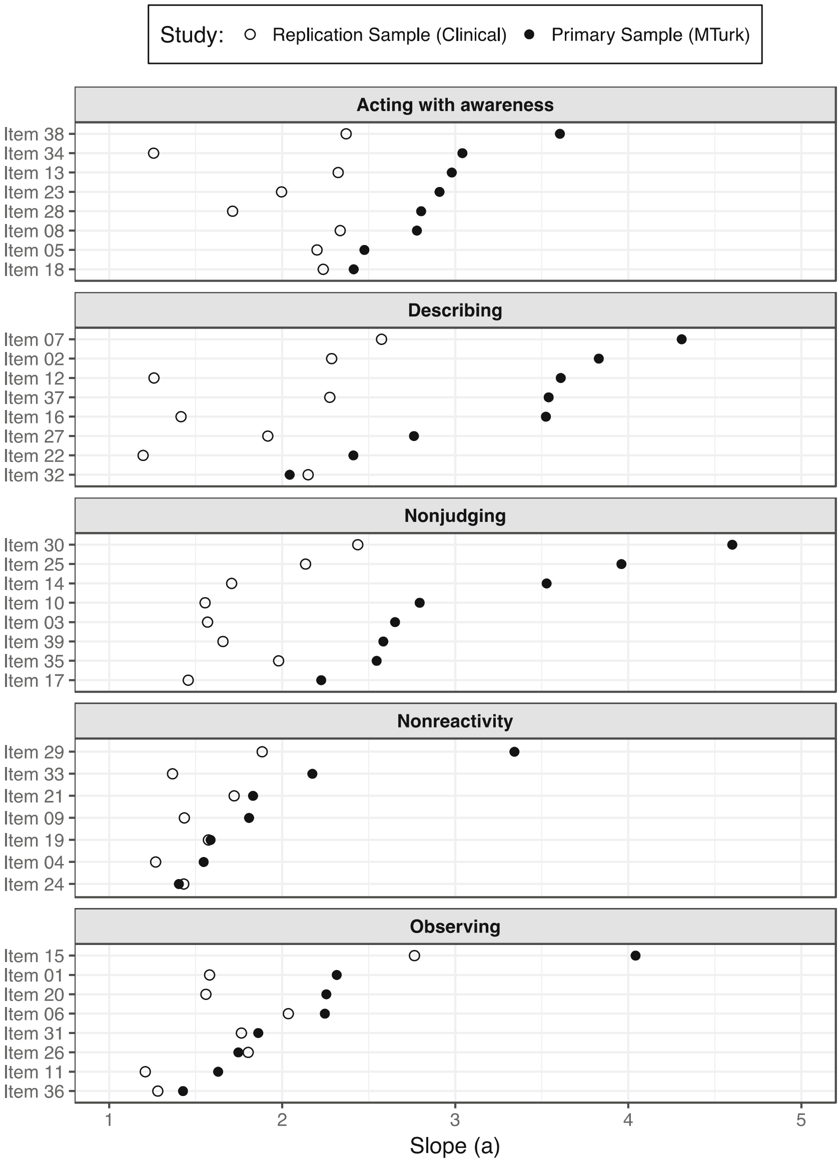


Fig. 1 Slopes of items on FFMQ-39 in both samples. Within facet, items are sorted in descending order based on slope in primary sample. A slope of zero indicates no relation of the item to the latent mindfulness facet. Thus, all items were related to their latent mindfulness facet, with slopes farther to the right in each panel indicating stronger relations

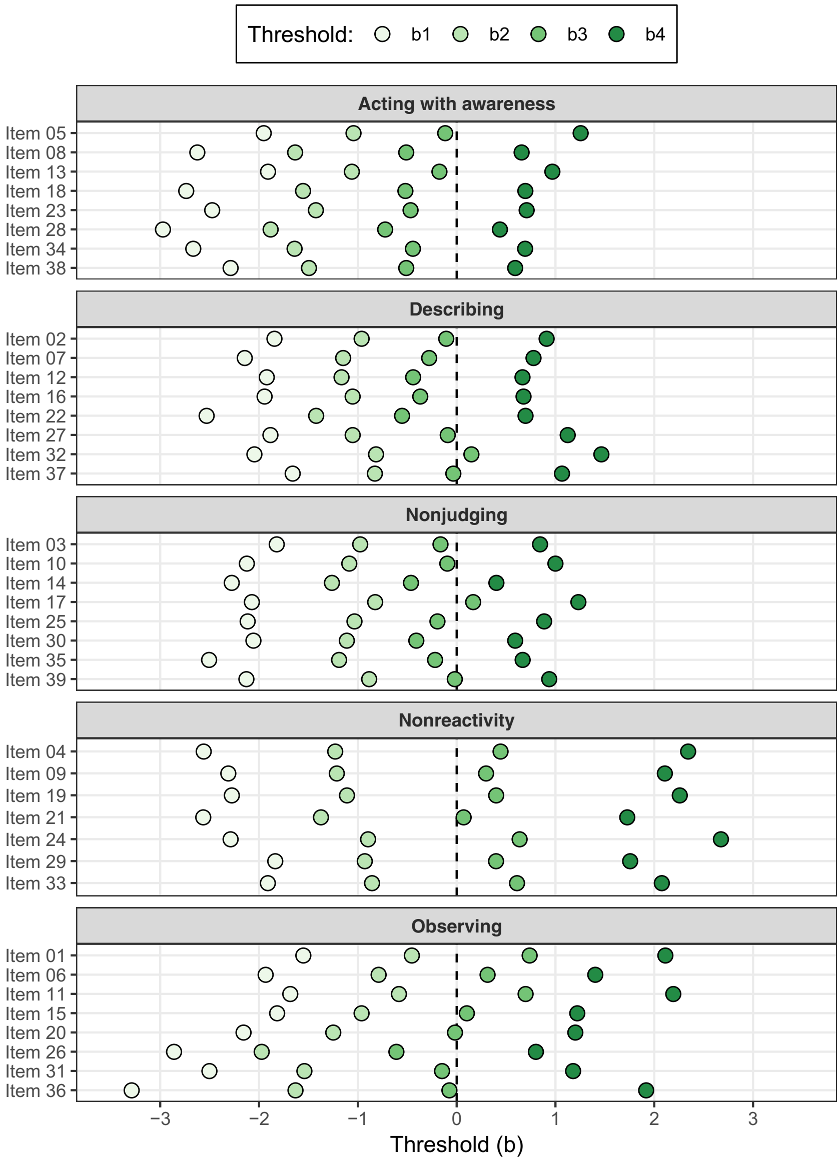
consistency was lower for the acting with awareness and observing facets. However, the relation of item responses to the latent construct was considerably weaker in the replication sample (mean slope of 1.8) than in the primary sample (mean slope of 2.7). The thresholds ranged from − 3.34 to 2.51, with a median value of − 0.34 (Table 2). As in the primary sample, most thresholds were below the mean of the latent variable metric, and very few thresholds discriminated individuals located in the high range of latent mindfulness.

Test Functioning

We first consider test functioning of the facets of the FFMQ-39. For reference, coefficient alphas were 0.84 for nonreactivity, 0.88 for observing, and 0.93 for each of the three remaining facets (Table S1). Moving to the IRT framework, the item parameters reported in Table 2 can be used to calculate the information, standard error of measurement, and score reliability for a person at any value of latent mindfulness (Edelen and Reeve 2007). For example, for a person who is one standard deviation above the mean on latent acting with awareness (i.e., $\theta = 1$), test information on the FFMQ-39 is 13.4, standard error of measurement is 0.27, and score reliability is 0.93. These attributes would vary across forms: the score for the same person on the FFMQ-15 would have test information of 5.3, standard error of measurement of 0.43, and score reliability of 0.82.

We focus on score reliability, since it is likely the metric most familiar to readers. Figure 4 (top row) shows the conditional score reliability (i.e., measurement precision) as a function of the latent variable for each of the five facets. The acting with awareness, describing, and nonjudging facets all

Fig. 2 Thresholds of items of FFMQ-39 in primary sample. "b1" is interpreted as the value of theta at which the individual has a 50% chance of responding with a response value above 1. "b2" through "b4" are interpreted analogously. For an example interpretation, consider item 36 at the very bottom of the figure. The dot for "b3" is located at approximately 0 on the *x*-axis. This implies that on item 36, we estimate that a participant will have 50% chance of responding with a value greater than 3 (i.e., "sometimes true") when they are at 0 (i.e., the mean) on the latent observing facet. Figure in color online



exhibited reliability above 0.90 through the span of approximately − 2.5 to + 1.5 SD from the mean of the latent variable. The observing facet exhibited similar reliability, but through a smaller range of the latent variable, from approximately − 2.0 to + 1.5 SD. Finally, the nonreactivity facet exhibited lower reliability (~ 0.85) across a wider range of the latent variable, from approximately − 2.0 to + 2.0 SD. In summary, while the five facets exhibit good to excellent reliability through much of the range in the latent variable, three of five exhibit reliability below 0.70 in the upper range (i.e., > 2 SDs).

We now turn to test functioning on the short forms, the FFMQ-24 and FFMQ-15. For reference, coefficient alphas across facets ranged from 0.80 to 0.90 for the FFMQ-24 and from 0.75 to 0.85 for the FFMQ-15 (Table S1). Moving to the IRT framework, Fig. 4 (top row) also shows conditional score reliability for the short forms, the FFMQ-24 and FFMQ-15. For all five facets, the FFMQ-39 was more reliable than the two short forms, through a wider range of theta. The reliability of the FFMQ-24 was generally about halfway between that of the FFMQ-15 and that of the FFMQ-39. Like the FFMQ-39, both short forms displayed adequate reliability (> 0.70) between approximately − 2.5 to + 1.5 SD from the mean of the latent variable. However, the short forms exhibited a steeper drop in reliability in the upper range of the latent variable.

Score linking analyses suggested that both short forms, FFMQ-24 and FFMQ-15, recovered the factor scores of the FFMQ-39 well. Table S2 shows the correlations of each short-form factor score with its full-form counterpart, as well as the mean squared errors of the short-form scores. Factor scores from both the FFMQ-24 (*r*s = 0.95–0.98) and FFMQ-15 (*r*s = 0.94–0.97) were highly correlated with those from the FFMQ-39. Mean squared errors ranged from 0.03 to 0.08 for the

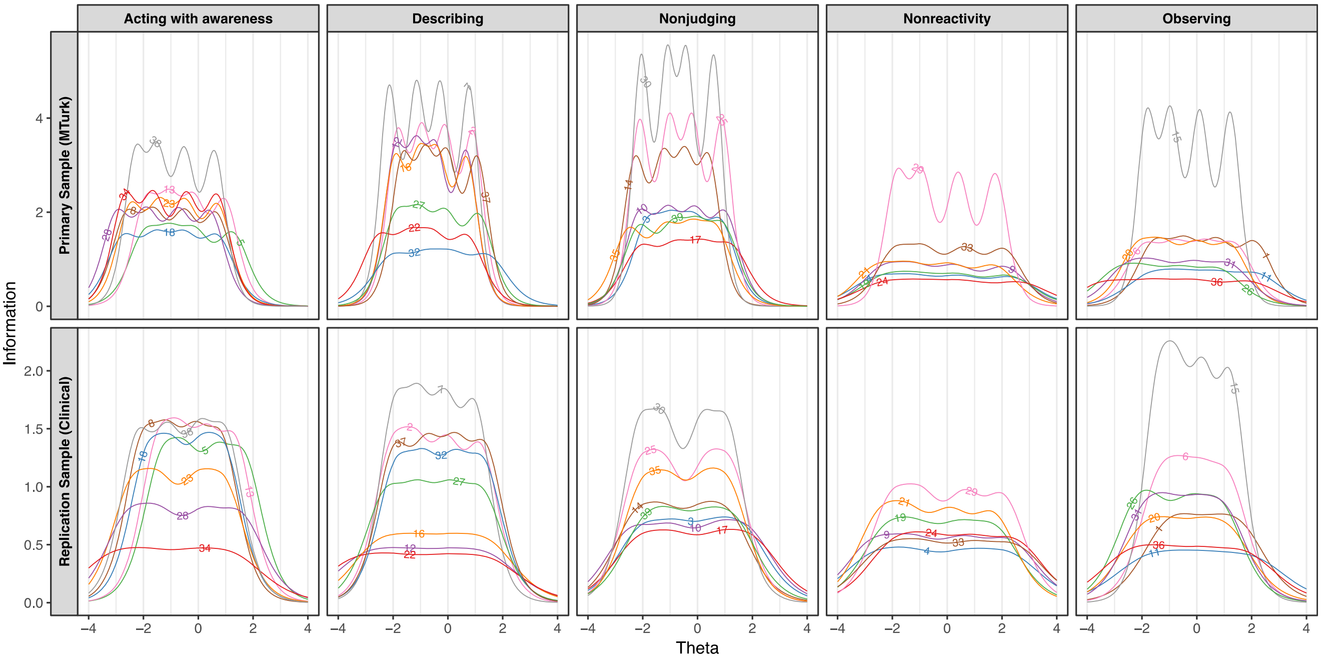


Fig. 3 Item information curves for FFMQ-39. Each curve indicates the information in a specific item, conditional on the individual's level of the latent variable ("theta"). Numbering indicates the item number, and item curve coloring is kept consistent across samples (i.e., within columns) to allow comparison of item information in the primary sample with item information in the secondary sample. Each panel indicates item functioning for a specific combination of facet (columns) and sample (rows). Figure in color online

FFMQ-24 and from 0.08 to 0.10 for the FFMQ-15. Thus, the factor scores produced by both short forms generally recovered the factor scores that would have arisen from the FFMQ-39 and did so with acceptable error.

We now turn to test functioning in the replication sample. Figure 4 (bottom row) shows conditional score reliabilities in the replication sample. Results for the FFMQ-39 generally mirrored those in the primary sample, with reliability being highest between approximately − 2.0 and + 1.5 SDs of theta, and lower in the upper range of latent mindfulness. However, reliability was considerably lower than it was in the primary sample, as suggested by the weaker item-construct relations (i.e., slopes) described earlier. Results for the FFMQ-24 and FFMQ-15 also mirrored those in the primary sample, with these forms conveying incrementally lower reliability, across incrementally smaller ranges of theta. Linking analyses suggested that factor scores produced by the short forms were still highly correlated with those produced by the FFMQ-39 (all *r*s > 0.87), but the mean squared errors were larger than in the primary sample (range = 0.14–0.21). In summary, results in the replication sample mirrored those from the primary sample, with reliabilities being lower in all cases.

Discussion

We used data from two samples to conduct item response theory analyses on the FFMQ-39, FFMQ-24, and FFMQ-15. All 39 items were significantly related to the latent mindfulness constructs they indicated, and all items conveyed information over a similar range of the latent constructs. Scores on the five facets were more reliable for individuals in the low range of latent mindfulness than in the high range. Reliability was poor for individuals in the upper range of latent mindfulness on three of the five facets—acting with awareness, describing, and observing. Reliability in the upper range of latent mindfulness was even poorer when using the short forms, FFMQ-24 and FFMQ-15. Findings were consistent across the primary and replication samples, although reliability was generally lower in the replication sample.

FFMQ Item Functioning

Results indicate that all 39 items of the FFMQ contribute useful information to the measurement of the facets to which they belong. All item slopes exceeded 1.0, on all three FFMQ forms, and in both the primary and replication samples. For readers more familiar with the factor loading metric, in the primary sample, the median loading was 0.84, and all loadings exceeded 0.60, demonstrating reasonable associations between the items and the latent variable measured. Results also indicate that the FFMQ items contribute information over a similar range of the latent mindfulness construct. As shown in Fig. 3, the item information functions are generally vertically aligned, differing only in their height. Thus, there appears to be limited potential

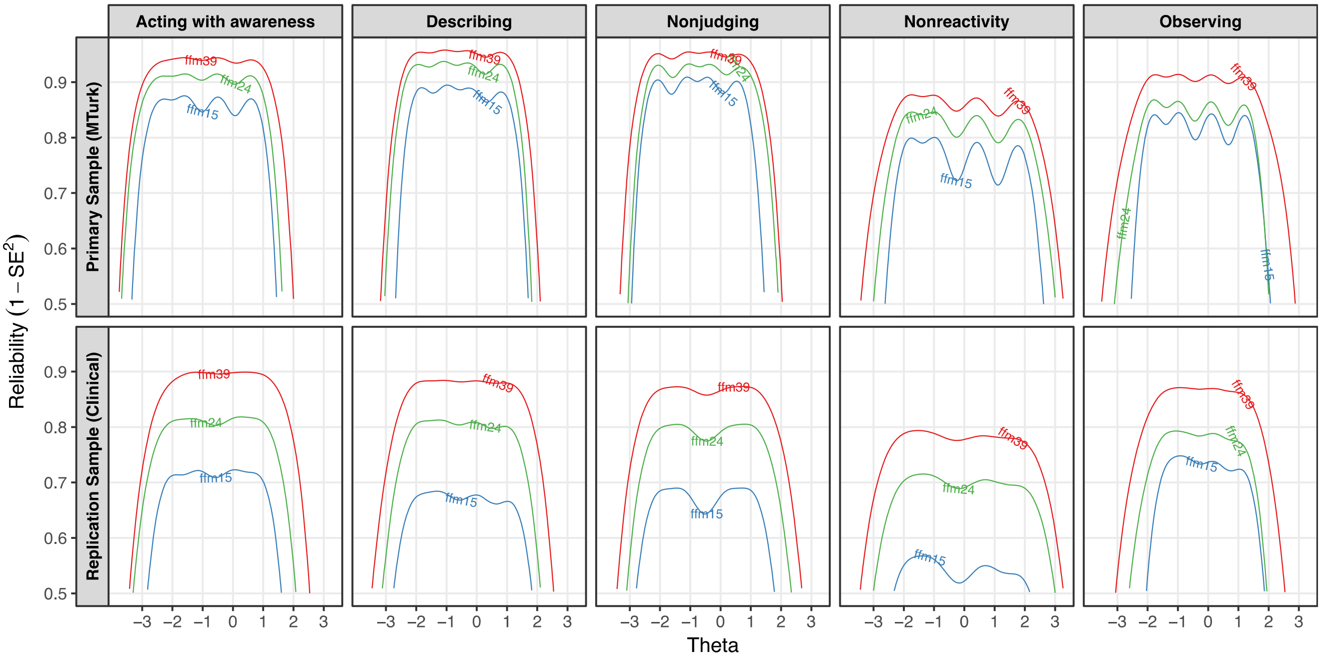


Fig. 4 Test reliability curves for FFMQ-39, FFMQ-24, and FFMQ-15. ffm39, FFMQ-39; ffm24, FFMQ-24; ffm15, FFMQ-15. Curves indicate estimated reliability for a specific form's measure of each facet, conditional on the respondent's level of the latent variable ("theta"). Each panel indicates conditional reliability for a specific combination of facet (columns) and sample (rows). Figure in color online

for strategically dropping items without materially affecting conditional reliability (cf. Van Dam et al. 2010).

Nearly all items conveyed more information in the low to medium range than the high range of the latent mindfulness construct. This may be related to the criticism of self-report measures of mindfulness as being too reliant on respondents' introspection into conceptually subtle mental states (Grossman 2008, 2011; Grossman and Dam 2011; van Dam et al. 2018). Perhaps respondents are better at noticing the gross absence of mindfulness (e.g., complete inattention to the task at hand) than distinguishing milder lapses in mindfulness (e.g., intermittent inattention to the task at hand). In this way, limitations of the self-report instruments used to measure mindfulness could be affecting our conceptualization of its structure.

Item slopes were substantially weaker in the replication sample than in the primary sample. There are at least two potential explanations for this discrepancy. First, participants in the replication sample had recently completed inpatient or intensive outpatient treatment for substance use disorder, so the reliability of their self-report may plausibly have been reduced by either state (e.g., recent alcohol or drug use) or trait (e.g., cognitive difficulties induced by long-term substance use) variables. Second, the primary and replication samples differed in other ways (e.g., age, ethnic composition, and sex) that might explain the observed difference in reliability. Future research may clarify how reliability and item functioning may vary as a function of person characteristics. Until then, our results suggest that reducing scale length could have a greater impact on reliability when studying focal populations.

FFMQ Test Functioning

The facets of the FFMQ-39, FFMQ-24, and FFMQ-15 all exhibited at least adequate reliability across part of the range of the latent variable. Across forms, all five of the FFMQ facets were more reliable when measuring individuals in the low to medium range of latent mindfulness than when measuring individuals in the high range. On three of the five facets—acting with awareness, describing, and nonjudging—reliability in the upper range was particularly poor. On the FFMQ-39, score reliability in the primary sample dropped below 0.70 when latent mindfulness exceeded 1.8, 1.9, and 1.8 SDs above the mean on each of these three facets. The use of short forms further reduced score reliability in the upper range. The corresponding cutoffs for the FFMQ-24 were 1.4, 1.6, and 1.7 SDs, and for the FFMQ-15 were 1.2, 1.5, and 1.2 SDs. These results can be understood more concretely when assuming latent mindfulness scores to be normally distributed (as is common practice in IRT analyses). In that case, a cutoff of 1.2 SDs would suggest that the scores of the upper 12% of the sample are measured with reliability lower than 0.70. Thus, the use of a short form not only reduces overall score reliability but also reduces the range of latent mindfulness over which scores are reliable (compare the curves in each panel of Fig. 4).

This phenomenon is explained by ceiling effects in the distribution of item responses. Figure S1 shows the distributions of total facet scores on the FFMQ-39, FFMQ-24, and FFMQ-15, most of which exhibit left skew, and some of which exhibit bounds at the right upper limit. On the FFMQ-39, approximately 5–10% of the sample respond to all items on the acting with awareness, describing, and nonjudging facets with the maximum response value. This means that scores near the upper boundary may not reflect participant's true mindfulness, and prohibits the scale from discriminating those who are already achieving a maximum score. Since there are fewer items per facet on the FFMQ-24 and FFMQ-15, a higher proportion of respondents reach the maximum score, and the ceiling effect is exacerbated (Figure S1). For example, 15% of the sample achieved the maximum score on the acting with awareness and nonjudging facets of the FFMQ-15.

Several lines of evidence suggest that the ceiling effect is a general issue, rather than a finding specific to our data. First, ceilings were present in both the primary and replication samples, which differed in several substantial ways. Second, both samples were large ($N = 522$ and $N \sim 410$), reducing the likelihood that the ceilings simply reflect sampling error. Third, ceilings were present in all three FFMQ forms, suggesting that it was not introduced by the selection of specific FFMQ items. Fourth, the means and standard deviations of item responses in both samples were similar to those observed in other published data (e.g., Veehof et al. 2011), suggesting that response values in our data are not aberrantly high.

Ceiling effects in item responses would be most problematic when studying populations that are high in mindfulness. However, a ceiling effect was present even in participants recently completing inpatient or intensive outpatient treatment for substance use disorder (i.e., the replication sample), a population that is lower in mindfulness than the general population (e.g., Karyadi et al. 2014; Shorey et al. 2014). Moreover, the ceiling effect may introduce problems even in low-mindfulness samples if mindfulness is measured to assess change over time. For example, suppose that the FFMQ is administered before and after an eight-week mindfulness-based intervention. Intervention-related change for participants that began the study already in the above-average range of mindfulness may not be reflected in the FFMQ facet scores. Similarly, if the intervention moves a large percentage of participants into the high end of the mindfulness range by the end of treatment, then change scores reflecting each person's improvements from baseline to end of treatment will be unreliable. Such a pattern could explain inconsistencies in tests of the FFMQ's sensitivity to change following intervention (Goldberg et al. 2016; Quaglia et al. 2016), role as a mediator of intervention effects (Hsiao et al. 2018), or lack of factorial invariance before and after intervention (Gu et al. 2016).

Potential Remedies to Ceilings One way to address the ceiling effect would be to modify the FFMQ-39 to increase item "severity," in the sense that the items receive lower mean response values. This could be done by modifying item wording, by adding items, or by revisiting the original item pool with the explicit goal of retaining difficult items (Baer et al. 2006). These changes would likely decrease the proportion of respondents reaching a facet score ceiling and thereby improve discrimination in the upper range of latent mindfulness. Moreover, new short forms could be created with item severity in mind, to produce a lower-burden measure that is still appropriate for measuring mindfulness among those in the high range. Scores on these short forms could be equated with those from the FFMQ-39, as in our score linking analyses. As shown in Table 1, many of the items retained on the FFMQ-15 are high in mean response value, resulting in facets especially vulnerable to ceiling effects.

In the meantime, the curves shown in Fig. 4 can aid investigators in choosing the most appropriate form of the FFMQ for their specific research question. Facet definitions (i.e., the item list) from different forms might be mixed and matched in a customizable fashion. For example, one might use the FFMQ-15 (i.e., lowest burden) definition of four of the five facets, and then use the FFMQ-39 (i.e., highest reliability) definition of the facet most directly targeted by an intervention and thus most important to measure precisely. More generally, the item functioning parameters reported in Table 2 can be used to compute the conditional test reliability of any arbitrary set of FFMQ items, enabling investigators to customize their administration of the FFMQ to their specific measurement needs (see supplement for elaboration).

Score Linking Score linking analyses showed that factor scores on the FFMQ-24 and FFMQ-15 were highly correlated with the factor scores that would have been obtained using the FFMQ-39 (though differing in reliability). These results illustrate that investigators can use IRT models to link scores obtained using any of the three forms. Linking might be used to merge samples in which different forms were administered (e.g., when conducting integrative data analysis, Curran and Hussong 2009), or to standardize repeated measurements of the same individual using different forms (e.g., use of FFMQ-39 at baseline and then FFMQ-15 at weekly follow-ups). Scores on the FFMQ-39, FFMQ-24, and FFMQ-15 could be linked by scoring item responses (a) using the item parameters reported in this manuscript (i.e., Table 2) or (b) using new item parameters obtained in the target dataset. How to link scores across forms is beyond the scope of this paper, so we direct readers to Lee and Lee (2018) for a description of several different approaches to doing so.

Limitations

Findings must be considered in the context of our samples. Our MTurk sample produced item response patterns consistent with existing literature and passed a series of quality checks (Eisenberg et al. 2018), but is still vulnerable to the general limitations of a sample recruited online. Our clinical sample consisted of participants with a history of substance use disorder, who may differ from the broader clinical populations with whom the FFMQ is often used (e.g., individuals with depression, anxiety, or stress). Participants with a history of heavy substance use may conceptualize mindfulness dimensions like nonreactivity or nonjudging differently than participants without this history (Eisenlohr-Moul et al. 2012). Finally, neither sample included a group of experienced mindfulness meditators, for whom the FFMQ items may function differently (Christopher et al. 2009).

Findings must also be considered in the context of our modeling approach. The FFMQ is typically used in applied contexts by calculating sum-scores for each of the five facets, then analyzing these five sum-scores as separate variables. Accordingly, we analyzed each of the five facets in a separate, unidimensional model, with the relations among items treated as independent, conditional on the latent variable (i.e., no correlated residual variances). While this specification corresponds with the prevailing treatment of the FFMQ, there is some evidence that the facets may not exhibit strict unidimensional structure (Tran et al. 2013; Van Dam et al. 2012).

In summary, all 39 items of the FFMQ contribute information to the measurement of the facets to which they belong, and do so over a similar range of the latent mindfulness constructs. Thus, we did not discover opportunities for strategic shortening of the FFMQ without negative impact on score reliability (cf. Van Dam et al. 2010). Scores in three of the five FFMQ-39 facets were unreliable when measuring individuals in the high range of latent mindfulness, resulting from ceiling effects in item responses. Reliability in the upper range was further reduced in the FFMQ-24 and FFMQ-15, which were more vulnerable to ceiling effects. Taken together, results suggest that all three FFMQ forms are best suited for measuring mindfulness in the low to medium range. Short forms may be ill-suited for situations in which reliable measurement in the upper range is necessary. Conditional reliability curves (Fig. 4) and item functioning parameters (Table 2) can aid investigators in tailoring their choice of FFMQ form to the specific reliability they hope to achieve and to the range of latent mindfulness over which they must reliably measure. Results await replication in samples from different populations.

Author Contributions LM and DM designed and executed the study from which data for primary sample were drawn. KW designed and executed the larger study from which data for secondary sample were drawn. WP, OG, and DM conceptualized the current research question and planned analyses. WP and OG carried out analyses and wrote the first draft of the manuscript. SM, CW, ES, KW, LM, and DP critically reviewed and revised the manuscript. All authors approved the final version of the manuscript for submission.

Funding Data for the primary sample were collected via a grant from the National Institute on Drug Abuse (NIDA) (UH2 DA041713). Data for the replication sample were collected via grants from NIDA (R21 DA010562 and R01 DA025764). Additional support was provided by the NIDA (P30 DA029926, R37 DA009757), the National Institute of Alcohol Abuse and Alcoholism (F31 AA026768), and the National Science Foundation (DGE-1311230). The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding agencies.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflicts of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. IRB approval for the primary sample study was received at Stanford University. IRB approval for the secondary sample studies was received at University of Washington.

Informed Consent Informed consent/assent was obtained from all participants in this study.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care, 42*, I7–I16.

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., Carmody, J., & Hunsinger, M. (2012). Weekly change in mindfulness and perceived stress in a mindfulness-based stress reduction program. *Journal of Clinical Psychology, 68*, 755–765.

Bohlmeijer, E., ten Klooster, P. M., Fledderus, M., Veehof, M., & Baer, R. (2011). Psychometric properties of the Five Facet Mindfulness Questionnaire in depressed adults and development of a short form. *Assessment, 18*, 308–320.

Bowen, S., Chawla, N., Collins, S. E., Witkiewitz, K., Hsu, S. H., Grow, J., et al. (2009). Mindfulness-based relapse prevention for substance use disorders: a pilot efficacy trial. *Substance Abuse, 30*, 295–305.

Bowen, S., Chawla, N., & Marlatt, G. A. (2011). *Mindfulness-based relapse prevention for addictive behaviors: a clinician's guide*. New York, NY: Guilford Press.

Bowen, S., Witkiewitz, K., Clifasefi, S. L., Grow, J., Chawla, N., Hsu, S. H., et al. (2014). Relative efficacy of mindfulness-based relapse prevention, standard relapse prevention, and treatment as usual for substance use disorders: a randomized clinical trial. *JAMA Psychiatry, 71*, 547–556.

Chalmer, R. P. (2012). mirt: a multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*, 1–29.

Christopher, M. S., Christopher, V., & Charoensuk, S. (2009). Assessing "Western" mindfulness among Thai Theravāda Buddhist monks. *Mental Health, Religion & Culture, 12*, 303–314.

Christopher, M. S., Neuser, N. J., Michael, P. G., & Baitmangalkar, A. (2012). Exploring the psychometric properties of the Five Facet Mindfulness Questionnaire. *Mindfulness, 3*, 124–131.

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychological Methods, 14*, 81–100.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: Guilford Publications.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*, 5–18.

Edwards, M. C. (2009). An introduction to item response theory using the Need for Cognition Scale. *Social and Personality Psychology Compass, 3*, 507–529.

Eisenberg, I. W., Bissett, P. G., Canning, J. R., Dallery, J., Enkavi, A. Z., Whitfield-Gabrieli, S., et al. (2018). Applying novel technologies and methods to inform the ontology of self-regulation. *Behaviour Research and Therapy, 101*, 46–57.

Eisenlohr-Moul, T. A., Walsh, E. C., Charnigo, R. J., Lynam, D. R., & Baer, R. A. (2012). The "what" and the "how" of dispositional mindfulness: using interactions among subscales of the Five-Facet Mindfulness Questionnaire to understand its relation to substance use. *Assessment, 19*, 276–286.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associaties.

Goldberg, S. B., Wielgosz, J., Dahl, C., Schuyler, B., MacCoon, D. S., Rosenkranz, M., et al. (2016). Does the Five Facet Mindfulness Questionnaire measure what we think it does? Construct validity evidence from an active controlled randomized clinical trial. *Psychological Assessment, 28*, 1009–1014.

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

Grossman, P. (2011). Defining mindfulness by how poorly I think I pay attention during everyday awareness and other intractable problems for psychology's (re)invention of mindfulness: comment on Brown et al. (2011). *Psychological Assessment, 23*, 1034–1040.

Grossman, P., & Dam, N. T. V. (2011). Mindfulness, by any other name…: trials and tribulations of sati in western psychology and science. *Contemporary Buddhism, 12*, 219–239.

Gu, J., Strauss, C., Crane, C., Barnhofer, T., Karl, A., Cavanagh, K., & Kuyken, W. (2016). Examining the factor structure of the 39-item and 15-item versions of the Five Facet Mindfulness Questionnaire before and after mindfulness-based cognitive therapy for people with recurrent depression. *Psychological Assessment, 28*, 791–802.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Hou, J., Wong, S. Y.-S., Lo, H. H.-M., Mak, W. W.-S., & Ma, H. S.-W. (2014). Validation of a Chinese version of the Five Facet Mindfulness Questionnaire in Hong Kong and development of a short form. *Assessment, 21*, 363–371.

Hsiao, Y.-Y., Tofighi, D., Kruger, E. S., Lee Van Horn, M., MacKinnon, D. P., & Witkiewitz, K. (2018). The (lack of) replication of self-reported mindfulness as a mechanism of change in mindfulness-based relapse prevention for substance use disorders. *Mindfulness*. https://doi.org/10.1007/s12671-018-1023-z.

Karyadi, K. A., VanderVeen, J. D., & Cyders, M. A. (2014). A meta-analysis of the relationship between trait mindfulness and substance use behaviors. *Drug and Alcohol Dependence, 143*, 1–10.

Lee, W. C., & Lee, G. (2018). IRT linking and equating. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: a multidisciplinary reference on scale and test development* (pp. 639–673). Hoboken, NJ: John Wiley & Sons, Inc..

Marsh, H. W., Ellis, L. A., Parada, R. H., Richards, G., & Heubeck, B. G. (2005). A short version of the Self Description Questionnaire II: operationalizing criteria for short-form evaluation with new applications of confirmatory factor analyses. *Psychological Assessment, 17*, 81–102.

Medvedev, O. N., Siegert, R. J., Kersten, P., & Krägeloh, C. U. (2017). Improving the precision of the Five Facet Mindfulness Questionnaire using a Rasch approach. *Mindfulness, 8*, 995–1008.

Medvedev, O. N., Titkova, E. A., Siegert, R. J., Hwang, Y.-S., & Krägeloh, C. U. (2018). Evaluating short versions of the Five Facet Mindfulness Questionnaire using Rasch analysis. *Mindfulness, 9*, 1411–1422.

Park, T., Reilly-Spong, M., & Gross, C. R. (2013). Mindfulness: a systematic review of instruments to measure an emergent patient-reported outcome (PRO). *Quality of Life Research, 22*, 2639–2659.

Quaglia, J. T., Braun, S. E., Freeman, S. P., McDaniel, M. A., & Brown, K. W. (2016). Meta-analytic evidence for effects of mindfulness training on dimensions of self-reported dispositional mindfulness. *Psychological Assessment, 28*, 803–818.

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory: fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science, 14*, 95–101.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

Sauer, S., Walach, H., Schmidt, S., Hinterberger, T., Lynch, S., Büssing, A., & Kohls, N. (2013). Assessment of mindfulness: review on state of the art. *Mindfulness, 4*, 3–17.

Shorey, R. C., Brasfield, H., Anderson, S., & Stuart, G. L. (2014). Differences in trait mindfulness across mental health symptoms among adults in substance abuse treatment. *Substance Use & Misuse, 49*, 595–600.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12*, 102–111.

Sochat, V. V., Eisenberg, I. W., Enkavi, A. Z., Li, J., Bissett, P. G., & Poldrack, R. A. (2016). The experiment factory: standardizing behavioral experiments. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.00610.

Thissen, D., & Wainer, H. (2001). *Test Scoring*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Tran, U. S., Glück, T. M., & Nader, I. W. (2013). Investigating the Five Facet Mindfulness Questionnaire (FFMQ): construction of a short form and evidence of a two-factor higher order structure of mindfulness. *Journal of Clinical Psychology, 69*, 951–965.

van Dam, N. T., Earleywine, M., & Borders, A. (2010). Measuring mindfulness? An item response theory analysis of the Mindful Attention Awareness Scale. *Personality and Individual Differences, 49*, 805–810.

van Dam, N. T., Hobkirk, A. L., Danoff-Burg, S., & Earleywine, M. (2012). Mind your words: positive and negative items create method effects on the Five Facet Mindfulness Questionnaire. *Assessment, 19*, 198–204.

van Dam, N. T., van Vugt, M. K., Vago, D. R., Schmalzl, L., Saron, C. D., Olendzki, A., et al. (2018). Mind the hype: a critical evaluation and prescriptive agenda for research on mindfulness and meditation. *Perspectives on Psychological Science, 13*, 36–61.

Veehof, M. M., ten Klooster, P. M., Taal, E., Westerhof, G. J., & Bohlmeijer, E. T. (2011). Psychometric properties of the Dutch Five Facet Mindfulness Questionnaire (FFMQ) in patients with fibromyalgia. *Clinical Rheumatology, 30*, 1045–1054.