Running Head: BALANCED INVENTORY OF MINDFULNESS-RELATED SKILLS

Development and Validation of the Balanced Inventory of Mindfulness-Related Skills (BIMS)

^{1,2} Nicholas T. Van Dam*

² Zarina Rovane Bilgrami

³ Tory Eisenlohr-Moul

¹ School of Psychological Sciences, University of Melbourne, VIC 3010, Australia

² Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY
10029, USA

³ Department of Psychiatry, University of Illinois at Chicago, Chicago, IL 60612, USA

* To whom correspondence should be addressed:

Nicholas T. Van Dam, Ph.D.

Melbourne School of Psychological Sciences,
University of Melbourne,

Victoria 0210, Australia

Email: nicholas.vandam@unimelb.edu.au

Abstract

Despite a number of psychometric limitations, the Five Facet Mindfulness Questionnaire (FFMQ) is one of the most widely used measures of mindfulness. Despite its potential to guide improvement of the scale, no multidimensional item response theory (IRT) analysis of the FFMQ has been conducted to date. The present studies fit competing Confirmatory Factor models to different versions of the FFMQ. A 23-item short-form of the FFMQ exhibited the most robust model-to-data fit, however, polytomous bifactor IRT exhibited poor properties, suggesting the need for scale revision. The authors undertook content revision, creating the Balanced-Inventory of Mindfulness-related Skills (BIMS). The BIMS used a structured alternative response format, wherein participants rate which of two response options is more like them, as well as the extent to which it is like them. Confirmatory factor and IRT analyses revealed reasonably good psychometric properties to a 22-item BiFactor scale, consisting of a general and 4 specific (ActAware, Describe, NonJudge, NonReact) factors. The BIMS appears to be a more psychometrically sound revision of the FFMQ, however, it's representation of the underlying construct of mindfulness continues to be a question for debate.

Keywords: Mindfulness; Psychometrics; Item Response Theory; Construct Validity

Introduction

Research into mindfulness and efforts to measure the construct via self-report are contentious (N.T. Van Dam et al., in press). Scholars have only loosely agreed that mindfulness entails aspects of attention, awareness, memory/retention, and acceptance/discernment (Davidson & Kaszniak, 2015). Despite inconsistencies in measurement, self-report assessment of mindfulness comprises something like half of all research currently being conducted in the field (cf. N.T. Van Dam et al., in press). Thus, it behooves us to understand how well these self-report measures reflect the construct of interest. There are over 10 different self-report measures of mindfulness, relationships between which are often inconsistent (Sauer, Walach, et al., 2013). The Mindful Attention Awareness Scale (MAAS; Brown & Ryan, 2003) and Five Facet Mindfulness Scale (FFMQ; Baer, Smith, Hopkins, Krietemeyer, & Toney, 2006) are by far the most popular of all self-report mindfulness measures (cf. Van Dam et al., 2018).

Interestingly, a literature search over the past decade on “psychometrics” and “mindfulness” returned 161 articles, while adding “IRT”, “Item Response Theory”, or “Rasch” resulted in only 11 articles (i.e., 6.8%). Despite many advantages over Classical Test Theory (CTT) approaches, Item Response Theory (IRT) seems not to have taken root in the measurement of mindfulness. Among the benefits of IRT approaches include sample invariance, estimates of item “difficulty”, estimates of item discriminability, item and scale information functions, and for Likert-type response formats, category response curves that provide information about the likelihood of endorsing a given response option for varying levels of the latent trait (cf. de Ayala, 2009; Embretson & Reise, 2000). To elaborate, sample invariance implies that while measurement error applies across scores,

it should generalize across populations. Moreover, unbiased estimates of item properties can be obtained from unrepresentative samples (Embretson & Reise, 2000). “Difficulty” estimates provide response category anchors, such that one can estimate the latent trait level at which a certain proportion of individuals are likely to endorse a given response option. This property is extremely informative in that it can suggest that only those with very high (or very low) levels of the latent trait are likely to select a (set of) response option(s). Discriminability suggests how well a given item or set of response options separate individuals who vary along the latent trait. High discrimination items or response options can useful separate individuals of very similar latent trait levels, while low discrimination items or response options may be useful for broader separation. Taken together, IRT provides a means of constructing combinations of items towards optimal differentiation of a given latent trait.

Efforts to use IRT for mindfulness scales have been made on the predecessor to the FFMQ (Medvedev, Siegert, Kersten, & Krägeloh, 2016), the MAAS (Goh, Marais, & Ireland, 2017; Van Dam, Earleywine, & Borders, 2010), and the Freiburg Mindfulness Inventory (Sauer, Ziegler, Danay, Ives, & Kohls, 2013). Notably, most implementations of IRT models have required either scale unidimensionality or (in the case of the Rasch model) treat multidimensionality as unique cases of multiple unidimensionality (Briggs & Wilson, 2003). The sole effort at applying a formal IRT model to the whole FFMQ (not just modeling certain aspects of item function, e.g., Differential Item Function) used the multiple unidimensional approach (Medvedev, Siegert, & Krägeloh, 2016). Notably, the authors of this study deleted two poorly performing items, combined locally dependent items, and merged response options where poor separation was observed. Modifications

such as these, permissible under the Rasch model approach, create difficulty regarding how to evaluate and use the scale going forward. It is important to note that there are some critical differences in how the Rasch modeling approach is viewed. Notably, proponents of logistic IRT models would typically suggest that the Rasch model is a unique case of the 1 parameter logistic model, where item difficulty can vary and item discrimination is fixed (Hambleton, Swaminathan, & Rogers, 1991). In contrast to primary emphasis on model-to-data fit, proponents of the Rasch model argue that model-to-data fit is important, but not primary (Andrich, 2004). The implications of this difference in perspective is that Rasch approaches allow for item-level analysis of misfit and exclusion of certain items if they can ‘credibly’ be argued not to measure the latent trait. Given the lack of agreement regarding the construct of mindfulness, we argue that such approaches may fail to account for known interrelatedness among the constructs that reflect mindfulness. Unfortunately, multidimensional IRT analysis of scales with polytomous (i.e., many response options) items has proven quite difficult and computationally intensive, especially when there are more than 3 dimensions to a given measure.

Fortunately, recent advances in IRT modeling using a bifactor model (i.e., a model wherein items load on a single general and one or more specific latent variables) have made it possible to fit a polytomous bifactor IRT model to multidimensional questionnaires using Likert-style response scales (Gibbons et al., 2007; Toland, Sulis, Giambona, Porcu, & Campbell, 2017). While full discussion of the polytomous bifactor IRT model is outside the scope of this paper, considerable detail about application of this

model to popular psychology questionnaires has recently been discussed (Toland et al., 2017).

The FFMQ has exhibited a number of concerning psychometric properties, including inconsistent factor structure (Baer et al., 2006; Baer et al., 2008), difficulties replicating and fitting proposed factor structures in large representative samples (e.g., Bohlmeijer, ten Klooster, Fledderus, Veehof, & Baer, 2011; Morgan, Dalrymple, Multach, & Zimmerman, 2017), reports of item response differences between meditators and non-meditators (Baer, Samuel, & Lykins, 2011; Van Dam, Earleywine, & Danoff-Burg, 2009), and potential concerns about method effects and/or response bias (e.g., Baer et al., 2011; Van Dam, Hobkirk, Danoff-Burg, & Earleywine, 2012). Fortunately, the bifactor polytomous IRT model is well-suited to address some of these concerns, and where necessary provide insight into possible scale revisions.

Thus, the present study aimed (1) to determine the optimal factor structure of the FFMQ based on previous suggestions, (2) fit the bifactor polytomous IRT model to the FFMQ to examine scale properties, (3) revise the scale format and content to address previously raised concerns about its psychometric properties, (4) test the CTT and IRT properties of the revised scale, and (5) examine construct validity of a revised scale, as well as make recommendations about further efforts to measure mindfulness with variations of the FFMQ.

Methods

Participants

Data was collected from four separate samples, two of which were from an undergraduate research pool at the University at Albany, State University of New York,

and two of which were collected via Amazon's Mechanical Turk (MTURK). MTURK is widely considered a valid means of acquiring reliable data (Buhrmester, Kwang, & Gosling, 2011; Hauser & Schwarz, 2016; Mason & Suri, 2012); while users are experienced survey takers, there are approximately 7,300 users available at any given time (Stewart et al., 2015). Participation by those in Sample 1 was completely voluntary (i.e., no reimbursement). Participants in Sample 1 ($N=586$) averaged 21 years of age ($SD=3.55$) and most (69%, $n=405$) were female. The majority (78%, $n=460$) of participants self-identified as Caucasian, while 6% identified as African American ($n=34$), 4% as Indian/Asian ($n=24$), 7% as Hispanic/Latino ($n=44$), 0.1% as Native American ($n=1$), and 4% as Other ($n=23$). Most participants (99%) were undergraduate students (99.1%), though a limited minority was not (Post-Baccalaureate: 0.5%, $n=3$; Graduate students: 0.3%, $n=2$; "Other": 0.1%, $n=1$).

Sample 2 received course credit for participation. The majority of participants (70.4%; $n=323$) were female. The average age of participants in sample 2 was 18.9 years of age ($SD=2.0$) and the majority (69%, $n=317$) self-identified as Caucasian, with 10.3 % ($n=47$) identified as Asian, 8.6% ($n=40$) as African American, 7.4% ($n=34$) as Hispanic/Latino, and 4.6% ($n=21$) as "Other."

Sample 3 participants were compensated \$1.50 for completion of study measures via MTURK. Participants averaged 36 years in age ($SD=11.44$) and 74% were female ($n=319$). The majority was Caucasian (84%, $n=361$), and 0.7% of participants ($n=3$) identified as American Indian/Alaskan Native, 3% ($n=13$) as Asian, 7% ($n=30$) as African American, 3.5% ($n=15$) as more than one race, and 1.7% ($n=15$) did not disclose race/ethnicity.

Sample 4 participants were compensated between \$0.10 and \$0.50 for completion of study measures via MTURK. Slightly different questionnaire batteries were administered to the sample, with approximately 1/3 getting a slightly shorter battery with some discrepant measures ($n = 208$, 34.3%) from the rest ($n = 398$, 65.7%). Participants were an average of 34.3 years in age ($SD = 11.1$) and 69.3% ($n=382$) were female. The majority self-identified as Caucasian (79.5%, $n=438$), and 0.7% ($n=5$) as American Indian/Alaskan Native, 4.5% ($n=25$) as Asian, 9.6% ($n=53$) as African America, 3.8% ($n=21$) as More than one race, and 1.6% ($n=9$) did not disclose race. The majority (90.0%, $n=496$) declined being of Hispanic/Latino ethnicity.

Procedures

Sample 1 data ($N = 586$) was acquired prior to several lectures from different undergraduate psychology courses. Potential participants from sample 1 were randomly given either the standard 5-option FFMQ or a modified 6-option version. Participation was voluntary; there was no reimbursement. All participants were provided with an explanation of the study purpose at its conclusion. Hereafter we refer to those who completed the 5-option FFMQ as dataset 1 (D1) and those who received the 6-option FFMQ as dataset 2 (D2). Sample 2 was given the standard 5-option FFMQ (as described in (Van Dam et al., 2012), yielding Dataset 3 (D3). Sample 3 (Dataset 4) was given a version attempting to mitigate wording bias in the FFMQ. Sample 4 (Dataset 5) was given a further modified version of the questionnaire administered to sample 3.

Procedures for all samples were deemed exempt by respective institutional review/ethics boards.

Measures

Balanced Index of Psychological Mindedness. Psychological Mindedness (PM) reflects an interest and ability to examine and monitor internal states (especially psychological ones). It could be considered to be a broad index of introspective tendency and skill, which arguably is quite important to mindfulness. The Balanced Inventory of Psychological Mindedness (BIPM; Nyklíček & Denollet, 2009) is a 14-item scale, comprised of 2 factors: Interest – an interest in internal phenomena, especially emotional experiences; Insight – a lack of insight into internal phenomena.

Balanced Inventory of Mindfulness-related Skills. Given prior reports of issues with original Five Facet Mindfulness Questionnaire (FFMQ; Baer et al., 2006), we aimed to find a question format that was developed to minimize desirable responding. We chose to use a variation of the “structured alternative format” (Harter & Messer, 2012), which presents two anchor statements of opposing directions (e.g., “Some teenagers do very well at their classwork” vs. “Other teenagers don’t do very well at their classwork”). Among other things, the format “normalizes” the response options in that it implies that people are generally like both of the anchors and that one anchor is not necessarily better than the other (see Harter & Messer, 2012). In addition to having theoretical benefits the original format showed minimal correlations with social desirability, while an alternative format of the same construct showed much larger correlations with social desirability (Harter & Messer, 2012). Despite being slightly more time-consuming than traditional scales, the “structured alternative format” has benefits over more traditional scales, when understood and administered correctly (Harter & Messer, 2012).

For a preliminary version of the BIMS, we adopted the “structured alternative format” and carefully revised 26 items from the original FFMQ (trying to represent those

items previously shown to optimally differentiate between mediators and non-mediators (Soler et al., 2014) and additionally representing most items from the FFMQ short form (Bohlmeijer et al., 2011). For the preliminary version, each item was presented as a combination of two anchor statements, from which the participant had to select the option that was more characteristic of him/her. Subsequently, the participant rated how characteristic that item was of them on a scale of 1 (a little characteristic of me) to 3 (very characteristic of me). The initial version comprised 26 items, a selection of 4-6 items that had been reworded for clarity from each of the 5 subscales.

In a revision of the preliminary BIMS, items were re-designed to present a single item with 2 anchor statements along with rankings. Participants were instructed to select between 1 (very characteristic of me) and 3 (a little characteristic of me) for the option on the left or 4 (a little characteristic of me) and 6 (very characteristic of me) for the option on the right. There were 30 total items (including further revisions of the 26 items in the preliminary BIMS, plus 4 additionally revised items), providing 6 options from each of the original subscales. It is important to note that for the within study revision, lower scores indicate higher levels of mindfulness-related skills.

Beck Depression Inventory, 2nd Edition. The BDI-II (Beck, Steer, & Brown, 1996) is a 21-item self-report measure of depression. Each item presents 4 options of varying magnitudes of symptom severity for a given symptom (e.g., anhedonia, general sadness). The BDI-II exhibits reasonably good psychometric properties and is one of the most widely used measures of depression (Beck, Steer, & Carbin, 1988).

Difficulties in Emotion Regulation Scale, Short Form. The DERS-SF (Kaufman et al., 2016) is a shortened (18-item) version of the longer (36-item), original DERS

(Gratz & Roemer, 2004). While it regains all 6 of the original subscales (Strategies, Non-acceptance, Impulse, Goals, Awareness, Clarity), they are represented by 3, rather than 6, items.

Five Facet Mindfulness Questionnaire. The FFMQ was developed via factor analysis of prior mindfulness scales. While is one of the most widely used measures of mindfulness (see e.g., Van Dam et al., 2018), it exhibits known differences in factor structure among different populations (see e.g., Baer et al., 2006; Baer et al., 2008; Bohlmeijer et al., 2011; Van Dam et al., 2009). In its original form, it consists of 39-items, comprising 5 subscales. Individual items are rated on a 5-option Likert-style response format ranging from 1 (never or very rarely true) to 5 (very often or always true). Approximately half of the items are reverse-scored and higher total scores are said to reflect higher levels of “dispositional mindfulness.”

NEO – Five Factor Inventory, Neuroticism subscale. The NEO-FFI (McCrae & Costa Jr, 2004) is one of the most widely used measures of personality. The total scale consists of 60 items (12 items for each of the Big 5 Personality constructs), each rated on a Likert-type scale from 1 (*Strongly disagree*) to 5 (*Strongly agree*) with respect to how much an individual thinks each item is characteristic of them. Herein only the 12-item neuroticism subscale was used as it has exhibited the strongest relationship to anxiety and depression, as well as to mindfulness (Giluk, 2009).

State Trait Inventory of Cognitive and Somatic Anxiety. The STICSA (Ree, French, MacLeod, & Locke, 2008) is a 21-item alternative to the State Trait Anxiety Inventory (STAI; Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1970). The STICSA was developed to overcome limitations of the STAI; namely poor differentiation of

anxiety from depression and lack of separate subscales for cognitive and somatic symptoms of anxiety. Unlike the STAI, the STICSA has shown stronger correlations with other measures of anxiety vs. depression (Gros, Antony, Simms, & McCabe, 2007). The STICSA has excellent psychometric properties and is rapidly gaining popularity (Van Dam, Gros, Earleywine, & Antony, 2013).

Zuckerman-Kuhlman Personality Questionnaire – Infrequency Index. The ZKPQ (Zuckerman, 2002) is a personality measure designed to assess traits that have strong biological-evolutionary relationships. Importantly, the infrequency index is a set of 10 true/false items that are designed to capture careless responding. In a large validation study of the ZKPQ ($N > 3000$), only 4% of subjects exhibited scores greater than 3, providing a useful index for responses of questionable validity (Zuckerman, 2002).

Analysis and Statistical Plan

Participants who were missing more than 5% of responses were excluded from the final datasets used for analysis. If less than 5% of data was missing, multiple imputation was used and their data was retained. For the remaining data, univariate and multivariate outliers were calculated. Participants with standardized scores $\geq |3.29|$ were considered to be univariate outliers and excluded. Similarly, participants whose data had Mahalanobis distances equivalent to $p \leq 0.001$ were excluded as multivariate outliers.

Exploratory and Confirmatory Factor Analyses were performed for a number of the datasets. Prior to performing EFA, parallel analysis (Horn, 1965) was run using the ‘nFactors’ package in R. Number of factors was ascertained by visual examination of actual eigenvalues relative to 95th percentile values of permuted data (5,000 permutations) on a screen plot. EFA was implemented using the factor number obtained

from parallel analysis, using maximum likelihood and promax rotation in the ‘psych’ package in R. CFA models were fit to the data using MLR estimation, implemented using the ‘lavaan’ package in R (v3.4.1).

As discussed above, several models have been proposed to fit the FFMQ, with varying success in different samples and with different numbers of items (e.g., Aguado et al., 2015; Baer et al., 2006; Baer et al., 2008; Morgan et al., 2017; Van Dam et al., 2012). In an effort to find the best model fit to our data, 6 different models were fit to 5 different configurations of items using CFA. The following models were also fit to the BIMS, though item numbers have been changed in the revision process. For each configuration of items, we fit a hierarchical five-factor model (cf. Baer et al., 2006), a correlated five-factor model (cf. Van Dam et al., 2012), a bi-factor model (cf. Aguado et al., 2015), a correlated five-factor, bi-method effects model (i.e., 5 factors plus 1 negative method effects factor and 1 positive method effects factor), a five-factor model plus a negative method effects factor, and a five-factor model plus a positive method effect factor (cf. Van Dam et al., 2012). When the item configuration called for removal of a full subscale, we fit a four-factor model (cf. Baer et al., 2008) in place of the five-factor analogue.

Cutoff criteria for good model fit were attained from existing literature. Comparative fit index (CFI) and Tucker Lewis index (TLI) $>.95$ was considered good and $>.90$ was considered reasonable, root mean square error of approximation (RMSEA) $<.06$ was considered good, standardized root mean residual (SRMR) $<.08$ was considered good (Hu & Bentler, 1999).

Bifactor polytomous IRT was implemented in R using the ‘mirt’ package (Chalmers, 2012) as recently recommended (Toland et al., 2017). Model assumptions about item independence (assessed by the oft-used Q3 index: Chen & Thissen, 1997; Yen, 1984), model-to-item fit (assessed using the $S-\chi^2$ statistic; (Bjorner, Chang, Thissen, & Reeve, 2007; Maria & David, 2003), and overall model fit (assessed using RMSEA, TLI, and CFI, computed from M2; Maydeu-Olivares & Joe, 2005), were examined. Only positive values of Q_3 are considered problematic and many use a value of 0.2 or higher to screen for local dependence. High local dependence suggests correlated errors, or similarities among items even after accounting for the latent variable(s) (see e.g., (S. P. Reise, Horan, & Blanchard, 2011)). The Bifactor polytomous IRT model yields two types of parameters for each item – an a value (corresponds to discrimination in traditional 2-parameter logistic (2PL) IRT models) and b values for the intercept between each contiguous response option pair (corresponds to difficulty in traditional 2PL IRT models). As a result of the way the model is fit, b values are only provided for the model as a whole (not separately for the general and specific factor(s)). However, equations are available (had were used) to convert the single set of b values to provide estimates of item discrimination and difficulty for both the general and specific factor corresponding to each item.

Results

Data Cleaning

As a number of the datasets were obtained using online platforms, cleaning/screening procedures were implemented to mitigate concerns about random response, careless responding, and other issues. Datasets 1 (D1) and 2 (D2) were obtained

in person, thus no cleaning procedures were used on this data. There was no missing data in either of these datasets. D1 (5-option response format FFMQ) had no univariate outliers and 10 multivariate outliers, final $n=297$. D2 (6-option response format FFMQ) contained 13 univariate outliers and 14 multivariate outliers, final $n=252$. Multiple participants attached to the same IP address in D3 (same data as described in Van Dam et al. 2012; traditional 39-item FFMQ) were excluded to ensure all data came from different participants. D3 was not examined for univariate/multivariate outliers on its own, as the dataset was only intended for use in combination with D1. For those participants from D3 for whom less than 5% of the data was missing and there was insufficient evidence to reject the hypothesis that the data was missing at random, multiple imputation was used. Coming from the same population, D1 ($n=307$) was combined with D3 ($n=459$). While there were no univariate outliers, 40 multivariate outliers were removed, leaving a total n of 726 in this combined dataset.

Dataset 4 (D4) resulted from $N=499$ participants, who completed the online survey and met inclusion criteria. Though participants were provided the option to state if they were impaired at the time of survey completion and/or that they did not think we should use their data, no participants endorsed either of these options. Three “checks” were included throughout the surveys. Any participant who selected the wrong answer on more than one of these checks, was excluded ($n=19$), resulting in $n=480$. The ZKPQ-Inf was also included. As validation data has shown Scores > 3 are very unlikely, occurring in less than 5% of a large population (Zuckerman, 2002), $n=12$ were additionally excluded for responses on this index. An inconsistency index (see Conners, Erhardt, & Sparrow, 1999) comprised of highly correlated item pairs from the NEO-FFI

(determined using an independent dataset of $n=900$) was also used to evaluate data. In simulation of 100,000 random datasets, the 95th percentile of inconsistency scores was equivalent to 41. We excluded 8 participants for exhibiting scores greater than this value, resulting in $n = 460$. An Attentiveness VAS scale was also included at the end of the surveys to capture self-report of effortful attention to the questionnaires. The 25th percentile (first quartile)-3*IQI was equivalent to a score of 76. Examination of the full data ($N=499$) revealed that scores ≤ 76 only occurred in 3.8% of cases. Any participant scoring 76 or below was excluded ($n=14$; $N=446$). Finally, we examined the amount of time it took for participants to complete the full set of questionnaires. Excluding any questionnaire that took longer than 60 min, suggesting a break, the average time was 1hr 4min ($SD = 34\text{min}$). Assuming approximately 3s per question, the total survey would have taken roughly 28min. Thus, anyone completing the survey in less than 28 minutes was excluded from additional analysis. This excluded 18 participants, leaving a sample of 429. There was no missing data, no univariate, and 2 multivariate outliers (who were removed), leaving a final n of 418.

As the second round of MTURK data (Dataset 5; D5) were time-constrained, we did not examine completion time. A total of 606 participants met inclusion criteria and completed at least 95% of the measures across the longer ($n=398$) and shorter ($n=208$) versions of the battery. As before, three checks were embedded in the surveys and a score of greater than 1 was considered problematic; 27 participants were removed for this reason (remaining $N = 579$). A total of 9 participants positive endorsed being impaired during survey completion and/or that we should not include their data; they were thus excluded (remaining $n = 570$). We additionally examined a VAS attentiveness score; data

from individuals who did not provide a score of at least 75 (to yield consistency with D4; $n = 19$) were excluded from further analyses (remaining $n = 551$). No data was missing. There were no univariate outliers, but 40 multivariate outliers, the latter of which were excluded (final $N = 511$).

Response Formatting

Because the middle response option does not always represent a neutral response, this can undermine the analysis of the latent trait and present false correlations, as well as provide a sort of “dumping ground” for when participants are uncertain of the item or their response (Kulas, Stachowski, & Haynes, 2008; Kulas & Stachowski, 2009). We gave one sample of participants the original 5-option survey and another sample a 6-option version, forcing people to pick a direction of endorsement. When the means and standard deviations were adjusted for range, they were essentially equivalent (D1: $M = 0.55$, $SD = 0.20$; D2: $M = 0.66$, $SD = 0.19$). Further, while endorsement patterns were shifted, the 6-option response format actually resulted in worse overall model fit (Hierarchical Model: $\chi^2 = 1284.96$, $df = 697$, CFI = 0.818, TLI = 0.807, RMSEA = 0.060, SRMR = 0.082) relative to the 5-option response format (Hierarchical Model: $\chi^2 = 1390.70$, $df = 697$, CFI = 0.843, TLI = 0.834, RMSEA = 0.060, SRMR = 0.076).

Factor Structure of FFMQ

Models fit to the standard 39-item dataset (D1-D3 combined, $n = 725$), showed that the correlated five-factor model with both positive and negative method effects provided the best fit to the data (see **Table 1**), consistent with prior work using a smaller subset of this sample (Van Dam et al. 2012). One of the benefits of the Bi-Factor model is that it permits examination of scale dimensionality, as well as the extent to which a

given factor and item is related to both a general dimension and a specific dimension (cf. Rodriguez, Reise, & Haviland, 2016a, 2016b). Noticing that the Bi-Factor model fit the 39-item data about as well as single method effects model, we further examined the relative magnitude of variance explained between a general factor and specific factors using the explained common variance (ECV; Steven P. Reise, Moore, & Haviland, 2010; Ten Berge & Sočan, 2004). We also used an item-level version of explained common variance (IECV; Stucky & Edelen, 2015; Stucky, Thissen, & Edelen, 2013) to examine the extent to which a given item loads on the general factor vs. its specific factor.

Examination of standardized factor loadings, ECV, and IECV, suggested that the General Factor explained approximately 30.5% of the common variance in responses. Moreover, items on the ActAware subscale exhibited very high general factor loadings and extremely high IECV values (ranging from ~ 74 – 99). In contrast, items on the Observe subscale exhibited very low general factor loadings and very low IECV values (see **Table 2**). In addition to the common approach of eliminating the Observe subscale (fitting a four-factor model instead; cf Baer et al. 2006), we also examined a model eliminating the ActAware subscale. All models without the Observe items, and separately, without the ActAware items, fit the data better than their counterparts in the full 39-item models (see **Table 1**). It is worth noting that items from the ActAware subscale have exhibited among the highest standardized factor loadings on the negative method effect factor (cf. Van Dam et al., 2012), which could underpin similarities in the Bifactor and Negative Method Effects model fit indices (see **Table 1**). As studies have reported an inability to fit the full five-factor model in community and psychiatric samples (cf. Bohlmeijer et al., 2011; Morgan et al., 2017), we also fit all models to a 24-

item short-form (as proposed by Bohlmeijer et al., 2011). The short form exhibited some of the best overall fit values, especially for the Correlated five-factor plus negative method effects model and the Bi-Factor model. We additionally tested a Short-form, without the Observe items, which showed further fit improvement, but only on the four-factor plus negative method effects model. We examined, but do not report a similar approach removing the ActAware items from the short-form as the conclusions were similar to those for removing Observe items from the short-form.

Item Response Theory Analysis of FFMQ.

To advance understanding of the FFMQ, and considering prior work using Rasch model approaches to the FFMQ (Medvedev et al., 2016), we fit a BiFactor Polytomous Item Response Theory Graded Response Model (cf. Gibbons et al., 2007; Reise, 2012; Toland et al., 2017) to the 24-item short-form configuration of the D1-D3 data. We used the 24-item short form because it resulted in the best model fit values. Among the items of the FFMQ-SF, Q_3 values > 0.2 occurred a minimum of 3 times (13% of pairings; per item) and a maximum of 9 times (39.1% of pairings; per item), with a mean of 5.17 (22.5% of pairing). Item-level fit was good; only 1 item (i.e., FFMQ 27) exhibited a value corresponding to $p < .05$ after applying False Discovery Rate correction. Overall model fit was also good, $RMSEA_{M2} = 0.034$, 90% CI = 0.027-0.055, $TLI_{M2} = 0.967$, $CFI_{M2} = 0.975$. Examination of adjusted IRT item parameters (cf. (Toland et al., 2017), however, indicated a number of problems (see **Table 4**), consistent with those observed in the BiFactor model parameters (**Table 3**). While the item discrimination parameters (i.e., a parameter) were generally good across the specific factors, they were more varied across the general a parameter, with very poor discrimination among the observe subscale items

(consistent with poor general factor loadings for these items). Response option intercepts (b parameters; i.e., location along latent trait at which probability of endorsing one response option shifts to next) were again acceptable for the specific factors, albeit with a fairly truncated range (median values ranged from -3.35 for b_1 to 2.36 for b_4), but extremely broad for the general factor (median values ranged from -8.73 for b_1 to 5.63 for b_4) – see Table 4. Consistent with the above parameters, we also examined test information for the general and each of the specific factors (cf. Toland et al., 2017). While the overall information acquired was relatively well distributed for the specific factors, there was a sharp decline in the acquired information above ~ 1.5 along the general latent trait, suggesting the scale (as a whole) does little to differentiate moderately high from high ‘mindfulness’.

Factor Structure and IRT properties of preliminary BIMS

Parallel analysis of the preliminary BIMS strongly suggested a 5-factor solution. While there was improvement (over what was seen in the FFMQ), the method effects model fit the data about as well as the bifactor model (see **Table 3**), though only at a level that would be considered acceptable to good, not good to excellent. Moreover, there were some relatively poor factor loadings and indications of inadequate separation between the general and specific factors for some items. Examination of IRT model assumptions indicated high local dependence, Q_3 values > 0.2 ranged from 4 (16.0% of pairs) to 13 (52.0% of pairs), with an average of 7.23 (28.9% of pairs). Overall model fit was reasonably good ($RMSEA_{M2} = 0.038$, 90% CI = 0.030-0.046, $TLI_{M2} = 0.939$, $CFI_{M2} = 0.953$, $SRMR = 0.063$), as was individual item fit (no items were significant after FDR correction). Discrimination parameters were reasonably good (only 4 items with $a < 0.3$

on general factor and 1 item on specific factor), but the response option intercepts exhibited ranges comparable to those observed with the FFMQ-SF (general: median values of -3.09 to 1.18, specific: median values of -2.19 to 1.37), corresponding to a lack of information at higher levels of theta. As a result of these issues, we decided to further revise the BIMS, as well as to add items to balance representation of all factors.

Factor Structure and IRT properties of revised 30-item BIMS. Parallel analysis indicated 5 factors. Factor analysis indicated a reasonably good fit for the BiFactor model to the data (considerably better than method effects model; see **Table 4**), though there were some issues with the factor loadings (see **Table 5**). While ECV values were acceptably concentrated across specific and general factors in the bifactor model, only one of the observe items had a standardized factor loading > 0.25 on the general factor and that same item had a relatively low standardized loading on the Observe subscale. Four other items on other factors exhibited loadings < 0.30 . With regards to IRT model assumptions, local independence was quite high. Q_3 values > 0.2 occurred a minimum of 4 times (13.8% of pairings; per item) and a maximum of 20 times (69.0% of pairings; per item), with a mean of 11.53 (39.8% of pairing). Notably, local dependency was substantially higher regarding intra-subscale item correlations ($MDN Q_3 = 0.409$) relative to inter-subscale item correlations ($MDN Q_3 = 0.128$). Overall model fit was good, as was item level fit was good (no values significant at $p < .05$ after FDR correction). IRT parameter estimates were problematic however (see **Table 6**), most notably items from the Observe subscale. Nearly all of these items exhibited poor general slopes and broad intercepts. Since the BiFactor IRT model is specified as a two-step model wherein an item can be modeled as loading on the general and not the specific factor, but not vice

versa, all Observe items were subsequently omitted from the model. Initial IRT analyses were run on a 24-item version of the scale, but some additional modifications were necessary. Notably, items 1, 6, and 19 were set to load only on the general factor and item 4 was omitted.

Factor Structure and IRT properties of revised 23-item BIMS.

The adjusted scale exhibited good standardized factor loadings; all specific loadings were > 0.3 while all general loadings were ≥ 0.24 . Model fit was also good (see **Table 4**). With regards to the IRT model, there were no major violations of the assumption of local independence. Q_3 values > 0.2 occurred a total of 4 times (1.6% of all pairs), with the maximum value equivalent to 0.287. Overall model fit was excellent ($RMSEA_{M2} = 0.041$, 90% CI = 0.032-0.049, $TLI_{M2} = 0.961$, $CFI_{M2} = 0.972$, $SRMR = 0.060$), as was item level fit was good (all $S-\chi^2$ values were smaller than the equivalent of $p < .05$ after FDR correction). IRT parameter estimates were reasonably good as well (see **Table 7**; cf. de Ayala, 2009).

Intra- and Inter- Scale Correlations

Latent trait estimates for the IRT model were computed using the default settings in the R package mirt (i.e., expected a-posteriori; cf. Chalmers, 2016). Sum scores and regression estimates were examined for the raw data (Grice, 2001). Sum scores resulted in modest-high correlations among subscales (r 's = .295-.503), which were absent in the IRT theta estimates and in the regression estimates. Correlations of the scale and subscale scores for latent variable, factor regression, and sum score estimates are reported in **Table 8**. As displayed in the upper left quadrant of the table, the latent variable representations exhibit good separation of the constructs. Regression scores, based on the

BiFactor model, resulted in high correlations with the latent variable estimates from IRT (upper right quadrant, **Table 8**) and good separation of the scales (upper triangle, lower right quadrant, **Table 8**). Sum scores resulted in subscale overlap otherwise absent in IRT and factor model estimates.

Relationships of factor model scores to external measures suggested that the total score carried the most weight with regards to relationships to other scales (e.g., neuroticism, depression, anxiety, emotion regulation; see **Table 9**). Use of sum scores resulted in significantly increased correlation (and poorer separation of constructs) with a number of external measures (**Table 9**). Of particular note, there was no correlation to the Interest component of Psychological-mindedness, but there was a correlation to the Insight component of Psychological-mindedness.

In an effort to provide a practical solution for use of the BIMS subscales, we computed partial correlations of subscales with other measures, including the BIMS total score as a covariate (see **Table 10**). This approach resulted in correlations that were more similar to the Factor Score relationships. A noticeable difference was apparent, however, with regards to global personality (e.g., neuroticism) and psychopathology measures (e.g., depression, anxiety), wherein the partial correlations of the NonJudge and NonReact subscales were not significant despite their factor score counterparts exhibiting a significant relationship.

Discussion

Using a combination of previously acquired (Van Dam et al., 2012) and novel data across 5 samples with a total of n=1834 participants, we sought to ascertain the optimal factor structure of the FFMQ and to apply BiFactor Item Response Theory

analysis to the data. Efforts to fit a number of previously proposed models, including various short forms, to the data, resulted in relatively poor model fits with the exception of Method Effects models, (Van Dam et al., 2012). These models suggest that there is something inherent in the wording of the items themselves (over and above the content of the items) that leads to particular patterns of endorsement. These findings suggest that there may be factors other than those which we intend to measure that influence how participants select response options on the FFMQ.

Notably, of the previously proposed models, the 24-item FFMQ short-form (Bohlmeijer et al., 2011) performed the best when a BiFactor model was fit to the data. Unfortunately, the items of the ActAware subscale exhibited very high Item-level Explained Common Variance, indicating poor separation of the subscale from the broader global construct. Moreover, IRT parameters were poor for the 24-item short form, with extremely broad ranges in the parameter estimates for the general construct and highly truncated ranges for parameter estimates of the specific constructs (i.e., subscales). For the latter, the subscales were often unable to assess much beyond a theta of 2.

Generation of a revised scale, titled the Balanced Inventory of Mindfulness-related Skills (BIMS) was undertaken to address the challenges previously discussed. The BIMS utilized the structured alternative response format (Harter & Messer, 2012). Although this format can be somewhat cumbersome with regards to additional efforts required in formatting and providing responses, it is well-known for its ability to minimize response bias. A preliminary version of the scale with 26 items performed acceptably, but required revision. After additional iterations, an acceptable version of the scale was developed. However, items related to the construct of Observe loaded very

poorly on the general construct, consistent with prior observations in general populations (Baer et al., 2006). Elimination of a few poorly performing items, only allowing select items to load on the general (and not specific factors), and eliminating the Observe subscale, resulted in good model fit from both a factor analytic and IRT perspective. Additionally, item information and category response curves looked good for the 23-item scale.

Three methods for calculating latent variable estimates were examined (IRT, Factor regression, and sum score). Both IRT and Factor regression estimates resulted in good factor separation, while the sum score approach exhibited some less than favorable features (especially for the subscales). The sum score approach also exhibited correlational patterns to external measures that appeared to conflate the general and specific constructs. Including the total sum score as a covariate in subscales analyses appeared to mitigate some issues with these scales. Thus, we recommend total sum score be used preferably and when subscales are of interest, total scale score should be included (as a covariate) to remove variance reflective of the general factor, thereby providing effects more comparable to the factor model. Unlike the FFMQ, it is acceptable in this case to enter all subscales, as well as the total score, into a multiple regression as they appear to reflect somewhat unique latent variables.

Consistent with prior work, correlations for the BIMS total score were highest with Neuroticism, Anxiety (especially cognitive), and overall Difficulties with Emotion Regulation. Interestingly, while there was no correlation with interest in Psychological-mindedness, there was a significant negative relationship with the Insight component of Psychological-mindedness, perhaps suggesting that those with higher levels of

Mindfulness-related skills exhibit more awareness of internal processes (whether for better – e.g., more accurate interoceptive awareness – or worse – e.g., heightened sensitive to negative thoughts – remains to be determined). Looking at the Factor regression estimates, the ActAware subscale exhibited only a small correlation with the Insight component of Psychological-mindedness and a small correlation to the *DERs Awareness* scale. The Describe subscale exhibited significant correlations to the *DERs Awareness* and *DERs Clarity* subscales, as well as the Insight component of Psychological-mindedness. The NonJudge subscale exhibited significant correlations to the *DERs NonAccept* subscale, along with small correlations to depression and cognitive anxiety. The NonReact subscale exhibited significant correlations to Neuroticism, as well as the *DERs Goals* and *Strategies* subscales. These unique correlational patterns (only evident in the factor regression model) indicate that the separate facets of mindfulness-related skills may represent unique constructs, as originally intended (Baer et al., 2006). Such separation of constructs, however, may be unattainable with the common sum score approach, though a proxy might be attained using total score covariates. If a sufficiently large normative sample is available, standard weightings could potentially be proposed to facilitate consistent and valid scoring.

The Bifactor structure IRT models fit herein for the 23-item BIMS were not validated in a separate sample. Thus, generalizability of these models awaits examination of the scale's properties in additional samples. Of particular importance, given past differences in the observed factor structure among experienced meditating and non-meditating samples (Van Dam et al., 2009), scale properties must be examined in these samples. Such examination is especially important for the *Observe* subscale, which was

omitted from the 23-item BIMS. Given recent lack of association between the FFMQ and actual meditation practice (Manuel, Somohano, & Bowen, 2017), such relationships should also be explored with the BIMS. Finally, the BIMS should be examined in relation to more objective measures of attention to the breath and other aspects of mindfulness (cf. (Levinson, Stoll, Kindy, Merry, & Davidson, 2014) and in relationship to those other characteristics of meditation practice that have been identified as important defining features, such as arousal, orientation of attention, motivation and goals, etc. (cf. Table 2; Van Dam et al., 2018).

Acknowledgements

The authors thank Dr. James Murrough for facilitation of the data collection and Ms. Georgina Costello for help with the figures.

Figure Legends

Figure 1. Total Information Function for Full Scale and 4 Subscales. Upper left panel represents total information across $\theta = -5$ to 5 for the general factor. Other panels represent total information across same θ range for subscales. Note that values for information are somewhat arbitrary while θ can be considered as comparable to a standard normal distribution.

Figure 2. Category Response Curves for General Trait Estimate. Category response curves represent probability of endorsing a particular response option at a certain trait (θ) level. Y axis represents probability of selecting a given response, while x axis represents latent trait estimate (θ). Each response option is color-coded, though insight is best gained by looking for a general pattern. Note that items 9 and 24 are examples where there is relatively clear differentiation of response likelihoods across latent trait estimates, while items 5 and 12 are examples of relatively poor differentiation of response options by latent trait estimates.

Figure 3. Category Response Curves for Specific Trait Estimates. Category response curves represent probability of endorsing a particular response option at a certain trait (θ) level. Y axis represents probability of selecting a given response, while x axis represents latent trait estimate (θ). Each response option is color-coded, though insight is best gained by looking for a general pattern. Note that items 9 and 24 are examples where there is relatively clear differentiation of response likelihoods across

latent trait estimates, while items 5 and 12 are examples of relatively poor differentiation of response options by latent trait estimates. Please note that latent traits herein are differentially represented. In other words, the latent trait for items 2, 3, and 5 corresponds to the ActAware subscale, items 7-12 to the Describe subscale, items 13-18 to the NonJudge subscale, and items 20-24 to the NonReact subscale.

Figure 1.

Figure 2.

Figure 3.

Table 1. Comparative fit indices for confirmatory factor analysis of FFMQ in a combined convenience sample (n=725)

		χ^2	df	χ^2/df	CFI	TLI	RMSEA	90% CI	SRMR
Original (39 items)	Hier-5F	2174.29	697	3.12	0.874	0.866	0.056	0.054, 0.057	0.066
	Corr-5F	2114.84	692	3.06	0.879	0.871	0.055	0.053, 0.058	0.054
	BiFac-5F	1747.12	663	2.64	0.908	0.897	0.049	0.047, 0.052	0.060
	Corr-5F + Bi-Method	1428.75	652	2.19	0.935	0.926	0.042	0.039, 0.045	0.044
	Corr-5F + Neg method	1767.24	673	2.63	0.908	0.899	0.049	0.046, 0.052	0.052
	Corr-5F + Pos Method	1788.15	672	2.66	0.906	0.896	0.050	0.047, 0.053	0.048
No Observe (31 total items)	Hier-4F	1539.93	430	3.58	0.889	0.88	0.063	0.059, 0.066	0.059
	Corr-4F	1525.48	428	3.56	0.890	0.881	0.062	0.059, 0.066	0.054
	BiFac-4F	1074.25	403	2.67	0.926	0.914	0.049	0.046, 0.053	0.050
	Corr-4F + Bi-Method	942.74	396	2.38	0.947	0.937	0.045	0.041, 0.049	0.045
	Corr-4F + Neg method	1212.22	409	2.96	0.921	0.910	0.054	0.050, 0.058	0.051
	Corr-4F + Pos Method	1281.288	416	3.08	0.914	0.903	0.056	0.053, 0.060	0.051
No ActAware (31 total items)	Hier-4F	1381.21	430	3.21	0.894	0.885	0.057	0.054, 0.061	0.066
	Corr-4F	1334.73	428	3.12	0.898	0.889	0.056	0.053, 0.060	0.055
	BiFac-4F	1229.54	433	2.84	0.914	0.901	0.052	0.048, 0.055	0.075
	Corr-4F + Bi-Method	927.04	396	2.34	0.942	0.932	0.044	0.040, 0.048	0.044
	Corr-4F + Neg method	1132.25	418	2.71	0.920	0.911	0.050	0.047, 0.054	0.051
	Corr-4F + Pos Method	1010.69	408	2.48	0.933	0.923	0.047	0.043, 0.050	0.047
FFMQ-SF (24 total items)	Hier-5F	617.08	247	2.50	0.938	0.931	0.047	0.043, 0.052	0.059
	Corr-5F	568.98	242	2.35	0.945	0.937	0.045	0.040, 0.050	0.047
	BiFac-5F	497.60	228	2.18	0.956	0.947	0.041	0.037, 0.046	0.051
	Corr-5F + Bi-Method	model would not converge.....							
	Corr-5F + Neg Method	477.67	230	2.08	0.959	0.951	0.040	0.035, 0.045	0.041
	Corr-5F + Pos Method	505.61	230	2.20	0.954	0.945	0.042	0.037, 0.047	0.045
FFMQ-SF No Observe (20 items)	Hier-4F	433.92	166	2.61	0.948	0.941	0.049	0.044, 0.055	0.057
	Corr-4F	414.21	164	2.53	0.952	0.944	0.048	0.042, 0.054	0.050
	BiFac-4F	420.07	151	2.78	0.949	0.936	0.051	0.046, 0.057	0.083
	Corr-4F + Bi-Method	model would not converge.....							
	Corr-4F + Neg Method	322.02	152	2.12	0.968	0.959	0.041	0.035, 0.047	0.041
	Corr-4F + Pos Method	model would not converge.....							

χ^2 = Chi-Square Statistic; df = degrees of freedom; CFI = Comparative Fit Index; TLI = Tucker Lewis Index; RMSEA = Root Mean Square Error of Approximation; 90% CI = 90% Confidence Interval of RMSEA; SRMR = Standardized Root Mean Residual;

Hier = Hierarchical; Corr = Correlated; BiFac = Bifactor; 5F = Five Factor; 4F = 4 Factor; Bi-Method = Positive & Negative Method Effect

Model; Neg Method = Negative Method Model Only; Pos Method = Positive Method Model Only; N.B. Fit estimates are Robust, except SRMR

Table 2. Standardized loadings and explained common variance of FFMQ bifactor model

Item	General	ActAware	Describing	NonJudge	NonReact	Observe	IECV
FFMQ5	0.735	0.436					73.97
FFMQ8	0.726	0.276					87.37
FFMQ13	0.731	0.411					75.98
FFMQ18	0.697	0.059					99.29
FFMQ23	0.674	-0.296					83.83
FFMQ28	0.691	-0.174					94.04
FFMQ34	0.628	-0.302					81.22
FFMQ38	0.769	-0.128					97.30
FFMQ2	0.24		0.755				9.18
FFMQ7	0.277		0.745				12.45
FFMQ12	0.385		0.766				20.17
FFMQ16	0.395		0.752				21.62
FFMQ22	0.396		0.484				40.10
FFMQ27	0.155		0.526				8.00
FFMQ32	0.137		0.634				4.46
FFMQ37	0.285		0.658				15.80
FFMQ3	0.287			0.604			18.42
FFMQ10	0.331			0.654			20.39
FFMQ14	0.364			0.679			22.32
FFMQ17	0.244			0.662			11.96
FFMQ25	0.332			0.775			15.51
FFMQ30	0.387			0.753			20.89
FFMQ35	0.352			0.63			23.79
FFMQ39	0.282			0.599			18.14
FFMQ4	0.066				0.408		2.55
FFMQ9	0.242				0.465		21.31
FFMQ19	0.245				0.652		12.37
FFMQ21	0.156				0.497		8.97
FFMQ24	0.158				0.596		6.57
FFMQ29	0.066				0.706		0.87
FFMQ33	0.114				0.654		2.95
FFMQ1	-0.036					0.523	0.47
FFMQ6	0.013					0.642	0.04
FFMQ11	-0.036					0.493	0.53
FFMQ15	-0.046					0.744	0.38
FFMQ20	-0.067					0.695	0.92
FFMQ26	0.025					0.574	0.19
FFMQ31	-0.002					0.624	0.00
FFMQ36	0.069					0.482	2.01
ECV (%)	30.51	3.51	19.14	19.10	12.34	15.41	

ECV = Explained Common Variance – the common variance explained by a given factor; IECV = Item-level explained common variance – the common variance of an item explained by the general factor.

Table 3. Adjusted IRT parameters for FFMQ-SF

scale	item	a.gen	b1.gen	b2.gen	b3.gen	b4.gen	a.spec	b1.spec	b2.spec	b3.spec	b4.spec
aa	18	1.89	-2.96	-1.45	0.04	1.61	0.34	-16.4	-8.07	0.20	8.94
aa	23	1.42	-4.02	-2.20	-0.11	2.13	0.86	-6.64	-3.64	-0.18	3.51
aa	28	1.40	-4.37	-2.27	0.45	2.81	0.90	-6.80	-3.52	0.70	4.37
aa	34	0.85	-7.04	-3.79	0.29	4.40	1.59	-3.76	-2.02	0.15	2.35
aa	38	1.38	-3.94	-1.77	0.82	3.34	1.14	-4.79	-2.15	0.99	4.06
de	2	0.59	-12.46	-5.40	0.49	7.87	2.38	-3.07	-1.33	0.12	1.94
de	7	0.70	-10.63	-5.12	0.09	5.76	2.21	-3.39	-1.63	0.03	1.84
de	12	1.00	-7.25	-3.45	0.23	4.74	1.80	-4.05	-1.92	0.13	2.65
de	22	1.01	-5.96	-2.85	-0.14	3.01	0.90	-6.74	-3.22	-0.15	3.41
de	27	0.45	-9.68	-4.02	0.86	6.91	1.32	-3.32	-1.38	0.30	2.37
nj	10	0.80	-5.80	-2.66	1.07	4.74	1.47	-3.15	-1.44	0.58	2.57
nj	17	0.54	-8.85	-2.81	2.00	6.25	1.69	-2.82	-0.90	0.64	1.99
nj	25	0.84	-10.39	-5.31	0.07	5.54	2.35	-3.70	-1.89	0.02	1.97
nj	30	0.96	-9.03	-4.64	-0.37	3.89	2.00	-4.33	-2.23	-0.18	1.87
nj	39	0.73	-6.25	-2.95	-0.29	2.94	1.25	-3.67	-1.73	-0.17	1.73
nr	9	0.61	-6.40	-2.96	0.87	5.72	0.84	-4.65	-2.15	0.63	4.15
nr	19	0.55	-8.62	-3.46	1.46	6.88	1.66	-2.86	-1.15	0.49	2.28
nr	24	0.33	-10.08	-2.22	5.44	13.22	1.59	-2.08	-0.46	1.12	2.72
nr	29	0.22	-21.48	-7.59	6.08	21.20	1.79	-2.58	-0.91	0.73	2.54
nr	33	0.30	-13.25	-2.46	7.02	16.40	1.92	-2.08	-0.39	1.10	2.57
obs	15	0.01	-654.3	-345.6	39.65	421.0	1.92	-2.24	-1.18	0.14	1.44
obs	20	-0.06	97.24	46.67	-0.57	-55.8	2.53	-2.29	-1.10	0.01	1.31
obs	26	0.14	-37.18	-23.95	-6.94	13.54	1.63	-3.17	-2.04	-0.59	1.16
obs	31	0.06	-60.58	-31.98	1.16	33.97	1.67	-2.33	-1.23	0.04	1.31

Table 4. Comparative fit indices for confirmatory factor analysis of BIMS

Data	Model	χ^2	df	χ^2/df	CFI	TLI	RMSEA	90% CI	SRMR
Prelim-BIMS (26 items)	Hier-5F	646.19	294	2.20	0.901	0.890	0.056	0.051, 0.062	0.075
	Corr-5F	605.49	289	2.10	0.911	0.900	0.054	0.048, 0.060	0.066
	BiFac-5F	459.01	273	1.68	0.948	0.938	0.042	0.036, 0.049	0.056
	Corr-5F + Bi-Method	448.54	263	1.71	0.949	0.937	0.043	0.036, 0.050	0.055
BIMS (30 items)	Hier-5F	1279.75	400	3.20	0.858	0.845	0.069	0.065, 0.074	0.108
	Corr-5F	1209.92	395	3.06	0.868	0.855	0.067	0.063, 0.071	0.098
	BiFac-5F	726.23	374	1.94	0.943	0.934	0.045	0.040, 0.050	0.058
	Corr-5F + Bi-Method	1001.12	368	2.72	0.898	0.880	0.061	0.056, 0.066	0.121
BIMS (23 items)	Hier-4F ^a	476.26	166	2.87	0.927	0.917	0.065	0.058, 0.071	0.071
	Corr-4F ^a	439.34	164	2.68	0.935	0.925	0.061	0.054, 0.068	0.061
	BiFac-4F	432.77	210	2.06	0.955	0.945	0.048	0.042, 0.055	0.048
	Corr-4F + Bi-Method	731.89	204	3.59	0.893	0.867	0.075	0.069, 0.081	0.150

^a In both cases for the traditional models, these values reflect the omission of items 1, 6, and 19, permitting optimal fit indices

χ^2 = Chi-Square Statistic; df = degrees of freedom; CFI = Comparative Fit Index; TLI = Tucker Lewis Index; RMSEA = Root Mean Square Error of Approximation; 90% CI = 90% Confidence Interval of RMSEA; SRMR = Standardized Root Mean Residual;

Hier = Hierarchical; Corr = Correlated; BiFac = Bifactor; 5F = Five Factor; 4F = 4 Factor; Bi-Method = Positive & Negative Method Effect

Model; Neg Method = Negative Method Model Only; Pos Method = Positive Method Model Only; N.B. Fit estimates are Robust, except SRMR

Table 5. Standardized loadings and explained common variance of 30-item BIMS

Item	General	ActAware	Describing	NonJudge	NonReact	Observe	IECV
bims1	0.681	<u>0.005</u>					<u>1.000</u>
bims2	0.560	0.323					0.750
bims3	0.362	0.663					0.230
bims4	0.371	0.884					0.150
bims5	0.439	0.636					0.323
bims6	0.509	<u>0.015</u>					<u>0.999</u>
bims7	0.480		0.709				0.314
bims8	0.565		0.564				0.501
bims9	0.572		0.598				0.478
bims10	0.315		0.399				0.384
bims11	0.486		0.655				0.355
bims12	<u>0.259</u>		0.614				0.151
bims13	0.444			0.494			0.447
bims14	0.363			0.728			0.199
bims15	0.375			0.754			0.198
bims16	0.494			0.457			0.539
bims17	0.530			0.494			0.535
bims18	0.613			0.402			0.699
bims19	0.641				<u>0.204</u>		0.908
bims20	0.642				0.396		0.724
bims21	0.451				0.358		0.613
bims22	0.579				0.526		0.548
bims23	0.446				0.592		0.362
bims24	0.648				0.562		0.571
bims25	<u>0.134</u>					0.458	0.079
bims26	<u>0.135</u>					0.611	0.047
bims27	<u>-0.094</u>					0.689	0.018
bims28	<u>-0.049</u>					0.594	0.007
bims29	<u>-0.041</u>					0.698	0.003
bims30	0.414					<u>0.271</u>	0.700
ECV (%)	41.17	11.22	13.90	12.69	8.23	12.78	

ECV= explained common variance; IECV=item explained common variance.

Table 6. Adjusted IRT parameter estimates for 30-item BIMS

scale	item	a.gen	b1.gen	b2.gen	b3.gen	b4.gen	b5.gen	a.spec	b1.spec	b2.spec	b3.spec	b4.spec	b5.spec
	1	1.84	-1.96	-1.17	-0.72	-0.18	1.06	<u>-0.10</u>	<u>36.49</u>	<u>21.72</u>	<u>13.36</u>	<u>3.27</u>	<u>-19.73</u>
	2	1.40	-2.01	-0.72	0.01	0.86	2.45	0.58	-4.87	-1.75	0.02	2.07	5.92
	3	0.84	-2.86	0.04	1.89	3.59	6.87	1.69	-1.42	0.02	0.94	1.78	3.40
	4	0.93	-7.09	2.06	6.28	12.31	20.73	2.89	-2.27	0.66	2.01	3.95	6.65
	5	1.07	-2.26	0.42	1.92	3.26	5.62	1.46	-1.65	0.31	1.40	2.38	4.10
	6	1.22	-1.70	-0.39	0.15	1.21	2.36	<u>0.04</u>	<u>-52.82</u>	<u>-12.27</u>	<u>4.67</u>	<u>37.58</u>	<u>73.52</u>
	7	1.06	-3.64	-1.19	-0.20	1.01	3.30	1.84	-2.10	-0.68	-0.11	0.58	1.90
	8	1.35	-1.85	-0.29	0.67	1.51	3.10	1.22	-2.05	-0.32	0.74	1.67	3.42
	9	1.33	-2.80	-0.92	0.16	1.40	3.20	1.34	-2.78	-0.91	0.16	1.39	3.17
	10	0.77	-2.96	-0.51	0.72	2.02	4.01	0.89	-2.54	-0.44	0.62	1.73	3.44
	11	1.13	-2.99	-0.78	0.09	1.09	2.98	1.50	-2.26	-0.59	0.07	0.83	2.25
	12	0.59	-4.27	-1.09	0.36	2.08	4.56	1.52	-1.65	-0.42	0.14	0.80	1.76
	13	0.94	-4.37	-2.22	-0.64	0.76	2.81	1.11	-3.70	-1.88	-0.54	0.64	2.38
	14	0.73	-6.96	-3.79	-1.90	0.41	4.60	2.08	-2.44	-1.33	-0.67	0.14	1.61
	15	0.78	-5.71	-3.12	-1.42	0.97	4.97	2.07	-2.15	-1.17	-0.53	0.37	1.87
	16	1.06	-2.60	-0.93	0.06	1.33	3.49	0.99	-2.79	-0.99	0.06	1.42	3.74
	17	1.11	-3.91	-1.94	-1.00	0.06	1.90	1.11	-3.93	-1.95	-1.01	0.06	1.91
	18	1.49	-1.98	-0.62	0.15	1.01	2.49	0.85	-3.46	-1.09	0.27	1.76	4.36
	19	1.61	-2.37	-1.05	-0.28	0.50	1.73	0.42	-9.00	-3.99	-1.07	1.89	6.55
	20	1.70	-2.35	-0.94	-0.20	0.64	1.99	0.77	-5.22	-2.09	-0.44	1.42	4.41
	21	1.00	-2.67	-0.81	-0.04	0.64	2.53	0.74	-3.62	-1.10	-0.05	0.87	3.43
	22	1.32	-3.21	-1.29	-0.33	0.66	2.24	1.18	-3.58	-1.44	-0.37	0.73	2.50
	23	0.94	-4.62	-1.98	-0.52	1.36	4.08	1.42	-3.08	-1.32	-0.35	0.91	2.72
	24	1.57	-4.40	-1.88	-0.56	0.78	2.61	1.27	-5.45	-2.32	-0.69	0.96	3.23
	25	<u>0.21</u>	<u>-12.85</u>	<u>-5.84</u>	<u>-1.46</u>	<u>2.12</u>	<u>10.59</u>	0.98	-2.73	-1.24	-0.31	0.45	2.25
	26	<u>0.25</u>	<u>-7.96</u>	<u>-1.70</u>	<u>3.64</u>	<u>6.91</u>	<u>13.61</u>	1.61	-1.22	-0.26	0.56	1.06	2.08
	27	<u>-0.20</u>	<u>10.09</u>	<u>2.00</u>	<u>-3.81</u>	<u>-7.49</u>	<u>-16.98</u>	1.95	-1.03	-0.20	0.39	0.77	1.74
	28	<u>-0.11</u>	<u>12.81</u>	<u>0.20</u>	<u>-8.45</u>	<u>-16.28</u>	<u>-32.22</u>	1.56	-0.93	-0.01	0.62	1.19	2.35
	29	<u>-0.11</u>	<u>20.67</u>	<u>2.43</u>	<u>-9.33</u>	<u>-18.18</u>	<u>-38.37</u>	1.98	-1.18	-0.14	0.53	1.03	2.18
	30	0.84	-1.81	0.44	1.46	2.29	4.09	0.57	-2.66	0.65	2.15	3.36	6.01

Table 7. Adjusted IRT parameter estimates for 23-item BIMS

scale	item	a.gen	b1.gen	b2.gen	b3.gen	b4.gen	b5.gen	a.spec	b1.spec	b2.spec	b3.spec	b4.spec	b5.spec
AA	1	1.86	-1.95	-1.16	-0.71	-0.18	1.04	-	-	-	-	-	-
AA	2	1.28	-2.24	-0.82	0.00	0.95	2.74	0.75	-3.83	-1.40	0.00	1.63	4.67
AA	3	0.75	-2.88	0.00	1.89	3.59	6.94	1.54	-1.40	0.00	0.92	1.75	3.37
AA	5	0.92	-2.83	0.51	2.43	4.11	7.01	1.75	-1.48	0.27	1.27	2.15	3.67
AA	6	1.21	-1.71	-0.41	0.14	1.21	2.38	-	-	-	-	-	-
De	7	0.92	-4.17	-1.36	-0.23	1.15	3.77	2.04	-1.88	-0.62	-0.10	0.52	1.70
De	8	1.18	-2.12	-0.34	0.76	1.72	3.53	1.39	-1.79	-0.28	0.64	1.45	2.99
De	9	1.21	-3.12	-1.03	0.17	1.54	3.55	1.49	-2.53	-0.84	0.14	1.25	2.88
De	10	0.71	-3.21	-0.56	0.77	2.18	4.34	0.95	-2.39	-0.41	0.58	1.63	3.24
De	11	0.99	-3.44	-0.91	0.10	1.25	3.41	1.68	-2.02	-0.53	0.06	0.74	2.01
De	12	0.52	-4.83	-1.23	0.41	2.35	5.16	1.57	-1.58	-0.40	0.14	0.77	1.69
NJ	13	1.01	-4.08	-2.07	-0.60	0.70	2.62	1.05	-3.94	-2.00	-0.58	0.68	2.53
NJ	14	0.83	-6.04	-3.28	-1.64	0.36	3.99	1.91	-2.63	-1.43	-0.72	0.16	1.74
NJ	15	0.84	-5.35	-2.92	-1.33	0.91	4.64	1.99	-2.25	-1.23	-0.56	0.38	1.96
NJ	16	1.04	-2.63	-0.94	0.06	1.34	3.53	0.99	-2.77	-0.99	0.06	1.41	3.71
NJ	17	1.17	-3.71	-1.84	-0.95	0.05	1.80	1.05	-4.14	-2.06	-1.07	0.06	2.01
NJ	18	1.55	-1.90	-0.60	0.15	0.96	2.39	0.79	-3.71	-1.18	0.29	1.88	4.67
NR	19	1.93	-2.05	-0.91	-0.24	0.43	1.49	-	-	-	-	-	-
NR	20	1.91	-2.10	-0.84	-0.18	0.57	1.77	0.60	-6.63	-2.66	-0.56	1.79	5.60
NR	21	1.05	-2.55	-0.78	-0.04	0.61	2.42	0.69	-3.88	-1.18	-0.06	0.92	3.68
NR	22	1.50	-2.79	-1.12	-0.29	0.57	1.95	1.00	-4.18	-1.68	-0.44	0.85	2.93
NR	23	1.09	-4.03	-1.73	-0.45	1.19	3.56	1.27	-3.45	-1.48	-0.38	1.02	3.05
NR	24	1.82	-3.84	-1.64	-0.49	0.68	2.29	1.08	-6.46	-2.76	-0.83	1.14	3.84

Table 8. Correlations within and between IRT, regression, and sum estimates

			1	2	3	4	5	6	7	8	9	10
<i>IRT Scores</i>	<i>M</i>	<i>SD</i>	<i>A.</i>					<i>B. IRT Scores x Regression Estimates</i>				
1. General	0.00	0.92	-					0.982	0.093	0.097	0.134	0.26
2. ActAware	-0.01	0.81	0.098	-				0.965	0.113	-0.175	-0.263	
3. Describe	0.00	0.89	0.107	0.131	-			0.967	-0.182	-0.282		
4. NonJudge	-0.01	0.87	0.117	-0.159	-0.152	-		0.976	0.026			
5. NonReact	-0.01	0.78	0.167	-0.212	-0.262	0.044	-	0.957				
<i>Sum Scores</i>	<i>M</i>	<i>SD</i>	<i>C. IRT Scores x Sum Scores</i>					<i>D. <u>Lower:</u> Sum Scores <u>Upper:</u> Regression Est.</i>				
6. General	81.29	20.45	0.936					0.944	0.098	0.110	0.113	0.195
7. ActAware	8.90	3.50	0.593	0.810				0.610	0.839	0.149	-0.154	-0.244
8. Describe	20.10	7.91	0.573	0.154	0.842			0.728	0.421	0.863	-0.159	-0.270
9. NonJudge	22.60	7.08	0.660	-0.066	-0.033	0.784		0.742	0.276	0.295	0.796	0.038
10. NonReact	18.46	6.09	0.781	-0.065	-0.094	0.074	0.692	0.757	0.343	0.315	0.503	0.724

Table 9. Nomothetic span of Balanced Inventory of Mindfulness-related Skills in subset of Sample 5 (n=359).

	<i>Total</i>	<i>ActAware</i>	<i>Describe</i>	<i>NonJudge</i>	<i>NonReact</i>
	<i>FacScr SumScr</i>	<i>FacScr SumScr</i>	<i>FacScr SumScr</i>	<i>FacScr SumScr</i>	<i>FacScr SumScr</i>
<i>NEO-Neuroticism</i>	0.676 0.666	0.033 0.340*	0.120 0.441*	0.177 0.506*	0.256 0.562*
<i>BDI-2</i>	0.602 0.611	0.045 0.307*	0.098 0.385*	0.239 0.526*	0.190 0.481*
<i>STICSA-Tot</i>	0.674 0.633*	0.053 0.347*	0.101 0.409*	0.149 0.485*	0.209 0.517*
<i>STICA-Cognitive</i>	0.718 0.683*	0.033 0.348*	0.110 0.433*	0.187 0.538*	0.242 0.565*
<i>STICSA-Somatic</i>	0.483 0.443*	0.069 0.276*	0.069 0.297*	0.070 0.321*	0.126 0.354*
<i>DERS-Tot</i>	0.673 0.653	0.083 0.375*	0.196 0.471*	0.157 0.502*	0.183 0.507*
<i>DERS-NonAccept</i>	0.589 0.588	0.010 0.277*	0.042 0.333*	0.290 0.568*	0.144 0.424*
<i>DERS-Goals</i>	0.551 0.541	0.039 0.289*	0.068 0.338*	0.131 0.399*	0.263 0.496*
<i>DERS-Impulse</i>	0.470 0.449	0.070 0.267*	0.021 0.259*	0.020 0.280*	0.235 0.436*
<i>DERS-Aware</i>	0.123 0.105	0.149 0.176	0.298 0.216*	-0.007 0.087	-0.184 -0.052*
<i>DERS-Clarity</i>	0.482 0.458	0.088 0.297*	0.349 0.463*	0.069 0.327*	0.021 0.284*
<i>DERS-Strategies</i>	0.646 0.632	0.024 0.308*	0.113 0.424*	0.140 0.460*	0.261 0.543*
<i>BIPM-Interest</i>	0.046 0.038	-0.024 -0.001	-0.186 -0.045	-0.026 -0.014	0.109 0.094
<i>BIPM-Insight</i>	-0.401 -0.373	-0.194 -0.343*	-0.364 -0.436*	-0.020 -0.247*	0.092 -0.160*

N.B. BIMS was scored such that lower scores reflected more mindfulness;

FacScr = correlations for Bifactor regression estimates; *SumScr* = correlations for sum scores

As r values > 0.1 were significant at $p < .05$, we indicate meaningful correlations in bold ($r \geq .25$).

* Indicates a significant difference in correlation values at $p < .05$ using a test for correlated correlations

Table 10. Nomothetic span of BIMS subscales via Factor Score and Partial Correlation estimates

	<i>ActAware</i>	<i>Describe</i>	<i>NonJudge</i>	<i>NonReact</i>
	<i>FacScr/PartCor</i>	<i>FacScr/PartCor</i>	<i>FacScr/PartCor</i>	<i>FacScr/PartCor</i>
<i>NEO-Neuroticism</i>	0.033 -0.065	0.120 -0.023	0.177* -0.031	0.256* -0.091
<i>BDI-2</i>	0.045 -0.064	0.098 -0.054	0.239* -0.098	0.190* -0.008
<i>STICSA-Tot</i>	0.053 -0.023	0.101 -0.038	0.149* -0.020	0.209* -0.049
<i>STICA-Cognitive</i>	0.033 -0.070	0.110 -0.059	0.187* -0.011	0.242* -0.071
<i>STICSA-Somatic</i>	0.069 0.032	0.069 -0.007	0.070 -0.046	0.126 0.014
<i>DERS-Tot</i>	0.083 0.004	0.196* -0.049	0.157* -0.018	0.183* -0.014
<i>DERS-NonAccept</i>	0.010 -0.088	0.042 -0.113	0.290* -0.215*	0.144* -0.005
<i>DERS-Goals</i>	0.039 -0.028	0.068 -0.048	0.131 -0.046	0.263* -0.140*
<i>DERS-Impulse</i>	0.070 0.015	0.021 -0.071	0.020 -0.127	0.235* -0.154*
<i>DERS-Aware</i>	0.149* -0.142*	0.298* -0.199*	-0.007 0.009	-0.184* -0.215*
<i>DERS-Clarity</i>	0.088 0.050	0.349* -0.233*	0.069 -0.056	0.021 -0.131
<i>DERS-Strategies</i>	0.024 -0.083	0.113 -0.010	0.140* -0.071	0.261* -0.103
<i>BIPM-Interest</i>	-0.024 -0.028	-0.186* -0.097	-0.026 -0.072	0.109 0.102
<i>BIPM-Insight</i>	-0.194* -0.170*	-0.364* - 0.267*	-0.020 0.077	0.092 0.226*

N.B. BIMS was scored such that lower scores reflected more mindfulness;

FacScr = correlations for Bifactor regression estimates; *PartCor* = partial correlations for sum scores with total score as a covariate

Meaningful correlations are bold ($r \geq .25$); * Indicates a significant correlation at $p < .01$

References

- Aguado, J., Luciano, J. V., Cebolla, A., Serrano-Blanco, A., Soler, J., & García-Campayo, J. (2015). Bifactor analysis and construct validity of the five facet mindfulness questionnaire (FFMQ) in non-clinical Spanish samples. *Frontiers in Psychology*, 6, 404. doi:10.3389/fpsyg.2015.00404
- Andrich, D. (2004). Controversy and the Rasch Model: A Characteristic of Incompatible Paradigms? *Medical Care*, 42(1), 1-7. doi:10.1097/01.mlr.0000103528.48582.7c
- Baer, R. A., Samuel, D. B., & Lykins, E. L. B. (2011). Differential Item Functioning on the Five Facet Mindfulness Questionnaire Is Minimal in Demographically Matched Meditators and Nonmeditators. *Assessment*, 18(1), 3-10. doi:10.1177/1073191110392498
- 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(1), 27-45. doi:10.1177/1073191105283504
- Baer, R. A., Smith, G. T., Lykins, E., Button, D., Krietemeyer, J., Sauer, S., . . . Williams, J. M. G. (2008). Construct validity of the five facet mindfulness questionnaire in meditating and nonmeditating samples. *Assessment*, 15(3), 329-342. doi:10.1177/1073191107313003
- Beck, A. T., Steer, R., & Brown, G. (1996). *Manual for the Beck Depression Inventory II (BDI-II)*. San Antonio, TX: Psychology Corporation.
- Beck, A. T., Steer, R. A., & Carbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, 8(1), 77-100. doi:10.1016/0272-7358(88)90050-5
- Bjorner, J. B., Chang, C. H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: item banking and computerized adaptive assessment. *Qual Life Res*, 16 Suppl 1, 95-108. doi:10.1007/s11136-007-9168-6
- Bohlmeijer, E., ten Klooster, P. M., Fledderus, M., Veehof, M., & Baer, R. A. (2011). Psychometric Properties of the Five Facet Mindfulness Questionnaire in Depressed Adults and Development of a Short Form. *Assessment*, 18(3), 308-320. doi:10.1177/1073191111408231
- Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of applied measurement*, 4(1), 87-100.
- 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(4), 822-848.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk. *Perspectives on Psychological Science*, 6(1), 3-5. doi:10.1177/1745691610393980
- Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. 2012, 48(6), 29. doi:10.18637/jss.v048.i06
- Chalmers, R. P. (2016). Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. 2016, 71(5), 38. doi:10.18637/jss.v071.i05
- Chen, W.-H., & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265-289. doi:10.3102/10769986022003265
- Conners, C. K., Erhardt, D., & Sparrow, E. (1999). *Conners' adult ADHD rating scales: Technical manual*. North Tonawanda, NY: Multi-Health Systems, Inc.

- Davidson, R. J., & Kaszniak, A. W. (2015). Conceptual and methodological issues in research on mindfulness and meditation. *American Psychologist*, 70(7), 581-592. doi:10.1037/a0039512
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.
- Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., . . . Stover, A. (2007). Full-Information Item Bifactor Analysis of Graded Response Data. *Applied Psychological Measurement*, 31(1), 4-19. doi:10.1177/0146621606289485
- Giluk, T. L. (2009). Mindfulness, Big Five personality, and affect: A meta-analysis. *Personality and Individual Differences*, 47(8), 805-811. doi:https://doi.org/10.1016/j.paid.2009.06.026
- Goh, H. E., Marais, I., & Ireland, M. J. (2017). A Rasch Model Analysis of the Mindful Attention Awareness Scale. *Assessment*, 24(3), 387-398. doi:10.1177/1073191115607043
- Gratz, K., & Roemer, L. (2004). Multidimensional Assessment of Emotion Regulation and Dysregulation: Development, Factor Structure, and Initial Validation of the Difficulties in Emotion Regulation Scale. *Journal of Psychopathology and Behavioral Assessment*, 26(1), 41-54. doi:10.1023/b:joba.00000007455.08539.94
- Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, 6(4), 430-450. doi:10.1037/1082-989X.6.4.430
- Gros, D. F., Antony, M. M., Simms, L. J., & McCabe, R. E. (2007). Psychometric properties of the State-Trait Inventory for Cognitive and Somatic Anxiety (STICSA): comparison to the State-Trait Anxiety Inventory (STAI). *Psychological Assessment*, 19(4), 369-381. doi: 10.1037/1040-3590.19.4.369
- Hambleton, R. K., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: SAGE.
- Harter, S., & Messer, B. (2012). *The self-perception profile for adults: Manual and questionnaires*. Denver: University of Denver.
- Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48(1), 400-407. doi:10.3758/s13428-015-0578-z
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179-185.
- Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. doi:10.1080/10705519909540118
- Kaufman, E. A., Xia, M., Fosco, G., Yaptangco, M., Skidmore, C. R., & Crowell, S. E. (2016). The Difficulties in Emotion Regulation Scale Short Form (DERS-SF): Validation and Replication in Adolescent and Adult Samples. *Journal of Psychopathology and Behavioral Assessment*, 38(3), 443-455. doi:10.1007/s10862-015-9529-3
- Kulas, J. T., Stachowski, A., & Haynes, B. (2008). Middle Response Functioning in Likert-responses to Personality Items. *Journal of Business and Psychology*, 22(3), 251-259. doi:10.1007/s10869-008-9064-2
- Kulas, J. T., & Stachowski, A. A. (2009). Middle category endorsement in odd-numbered Likert response scales: Associated item characteristics, cognitive demands, and preferred

- meanings. *Journal of Research in Personality*, 43(3), 489-493.
doi:10.1016/j.jrp.2008.12.005
- Levinson, D. B., Stoll, E. L., Kindy, S. D., Merry, H. L., & Davidson, R. J. (2014). A mind you can count on: validating breath counting as a behavioral measure of mindfulness. *Frontiers in Psychology*, 5, 1202. doi:10.3389/fpsyg.2014.01202
- Manuel, J. A., Somohano, V. C., & Bowen, S. (2017). Mindfulness Practice and Its Relationship to the Five-Facet Mindfulness Questionnaire. *Mindfulness*, 8(2), 361-367.
doi:10.1007/s12671-016-0605-x
- Maria, O., & David, T. (2003). Further Investigation of the Performance of S - X2: An Item Fit Index for Use With Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, 27(4), 289-298. doi:10.1177/0146621603027004004
- Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1), 1-23. doi:10.3758/s13428-011-0124-6
- Maydeu-Olivares, A., & Joe, H. (2005). Limited- and Full-Information Estimation and Goodness-of-Fit Testing in 2n Contingency Tables. *Journal of the American Statistical Association*, 100(471), 1009-1020. doi:10.1198/0162145040000002069
- McCrae, R. R., & Costa Jr, P. T. (2004). A contemplated revision of the NEO Five-Factor Inventory. *Personality and Individual Differences*, 36(3), 587-596.
doi:http://dx.doi.org/10.1016/S0191-8869(03)00118-1
- Medvedev, O. N., Siegert, R. J., Kersten, P., & Krägeloh, C. U. (2016). Rasch Analysis of the Kentucky Inventory of Mindfulness Skills. *Mindfulness*, 7(2), 466-478.
doi:10.1007/s12671-015-0475-7
- Morgan, T. A., Dalrymple, K., Multach, M., & Zimmerman, M. (2017). Relations Between Mindfulness Processes, Psychiatric Symptoms, and Functioning in Psychiatric and Pre-bariatric Surgery Outpatients. *Mindfulness*, 8(2), 417-432. doi:10.1007/s12671-016-0614-9
- Nyklíček, I., & Denollet, J. (2009). Development and Evaluation of the Balanced Index of Psychological Mindedness (BIPM). *Psychological Assessment*, 21, 32-44.
- Ree, M. J., French, D., MacLeod, C., & Locke, V. (2008). Distinguishing Cognitive and Somatic Dimensions of State and Trait Anxiety: Development and Validation of the State-Trait Inventory for Cognitive and Somatic Anxiety (STICSA). *Behavioural and Cognitive Psychotherapy*, 36(03), 313-332.
- Reise, S. P. (2012). The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research*, 47(5), 667-696. doi:10.1080/00273171.2012.715555
- Reise, S. P., Horan, W. P., & Blanchard, J. J. (2011). The challenges of fitting an item response theory model to the Social Anhedonia Scale. *Journal of Personality Assessment*, 93(3), 213-224. doi:10.1080/00223891.2011.558868
- Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor Models and Rotations: Exploring the Extent to which Multidimensional Data Yield Univocal Scale Scores. *Journal of Personality Assessment*, 92(6), 544-559. doi:10.1080/00223891.2010.496477
- Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying Bifactor Statistical Indices in the Evaluation of Psychological Measures. *Journal of Personality Assessment*, 98(3), 223-237. doi:10.1080/00223891.2015.1089249
- Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, 21(2), 137-150.
doi:10.1037/met0000045

- 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(1), 3-17. doi:10.1007/s12671-012-0122-5
- Sauer, S., Ziegler, M., Danay, E., Ives, J., & Kohls, N. (2013). Specific Objectivity of Mindfulness—A Rasch Analysis of the Freiburg Mindfulness Inventory. *Mindfulness*, 4(1), 45-54. doi:10.1007/s12671-012-0145-y
- Soler, J., Cebolla, A., Feliu-Soler, A., Demarzo, M. M. P., Pascual, J. C., Baños, R., & García-Campayo, J. (2014). Relationship between Meditative Practice and Self-Reported Mindfulness: The MINDSENS Composite Index. *PLoS One*, 9(1), e86622. doi:10.1371/journal.pone.0086622
- Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1970). *State - Trait Anxiety Inventory*. Menlo Park, CA: Mind Garden, Inc.
- Stewart, N., Ungemach, C., Harris, A. J., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making*, 10(5), 479.
- Stucky, B. D., & Edelen, M. O. (2015). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 183-206). New York: Routledge.
- Stucky, B. D., Thissen, D., & Edelen, M. O. (2013). Using Logistic Approximations of Marginal Trace Lines to Develop Short Assessments. *Applied Psychological Measurement*, 37(1), 41-57. doi:10.1177/0146621612462759
- Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, 69(4), 613-625. doi:10.1007/bf02289858
- Toland, M. D., Sulis, I., Giambona, F., Porcu, M., & Campbell, J. M. (2017). Introduction to bifactor polytomous item response theory analysis. *Journal of School Psychology*, 60(Supplement C), 41-63. doi:https://doi.org/10.1016/j.jsp.2016.11.001
- 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(7), 805-810. doi:10.1016/j.paid.2010.07.020
- Van Dam, N. T., Earleywine, M., & Danoff-Burg, S. (2009). Differential item function across meditators and non-meditators on the Five Facet Mindfulness Questionnaire. *Personality and Individual Differences*, 47(5), 516-521. doi:http://dx.doi.org/10.1016/j.paid.2009.05.005
- Van Dam, N. T., Gros, D. F., Earleywine, M., & Antony, M. M. (2013). Establishing a trait anxiety threshold that signals likelihood of anxiety disorders. *Anxiety Stress Coping*, 26(1), 70-86. doi:10.1080/10615806.2011.631525
- 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(2), 198-204. doi:10.1177/1073191112438743
- Van Dam, N. T., van Vugt, M. K., Vago, D. R., Schmalzl, L., Saron, C. D., Olendzki, A., . . . Meyer, D. E. (2018). Mind the hype: A critical evaluation and prescriptive agenda for research on mindfulness and meditation. *Perspectives on Psychological Science*, 13, 36-61. doi: 10.1177/1745691617709589

- Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement*, 8(2), 125-145. doi:10.1177/014662168400800201
- Zuckerman, M. (2002). Zuckerman-Kuhlman Personality Questionnaire (ZKPQ): An alternative five-factorial model. In B. DeRaad & M. Persini (Eds.), *Big Five Assessment* (pp. 377-396). Seattle, WA: Hogrefe and Huber Publishers.