Mindful in a random forest: Assessing the validity of mindfulness items using Random Forests methods

Sebastian Sauer[1,2*], Jana Lemke[2,3], Winfried Zinn[4,5], Ricardo Buettner[1], Niko Kohls[2,6]

1 FOM University of Applied Sciences, Essen, Germany

2 Samueli Institute, Alexandria, USA

3 Viadrina University, Frankfurt (Oder) , Germany

4 Forschungsgruppe Metrik, Bermuthshain, Germany

5 Applied University for Health and Sports, Berlin, Germany

6 Coburg University of Applied Sciences, Coburg, Germany

Author Note

*Corresponding author. Institute of Management & Information Systems, FOM University of Applied Sciences, Leimkugelstraße 6, 45141 Essen, Germany. Email: Sebastian.sauer@fom.de, Phone: 0049-89/202452-27

Abstract

Whereas the number of studies supporting the efficacy of mindfulness as a health intervention is increasing, the measurement of mindfulness remains a subject of debate. Given the importance of measurement in this field, this paper aims to further our understanding of the assessment of mindfulness by employing an approach referred to as "Random Forests" (RF). RF is an ensemble learning method that is based on decision trees. RF is well known in biological research, for example, but is practically unknown in psychometrics. In this study, RF was used to gauge the predictive validity of the items from two mindfulness instruments concerning their ability to estimate group allocation (i.e., mindfulness practitioners vs. nonpractitioners). To allow for a better generalization of the results, we examined the research questions in two samples (N = 76 and N = 202) of different quality. We investigated two instruments: the Freiburg Mindfulness Inventory (FMI) and the Mindfulness Attention and Awareness Scale. Although results indicated that both instruments were capable of distinguishing practitioners from nonpractitioners, the predictive quality of most items on both scales was determined to be insufficient.

Keywords: mindfulness, assessment, measurement, psychometrics, random forests

1 Introduction

Mind full or mindful? This pun speaks of complementary states of mind. A "full mind" can be associated with cognitive processes, subserving forms of mind-wandering such as preoccupied thinking (e.g., resentment or contentment regarding past experiences and desire or dissatisfaction regarding anticipated future events). Two characteristics of mind-wandering are that they are (a) either past- or future-centered, and correspondingly not anchored in the experience of the present moment, and (b) invoke some kind of predominantly negative emotional reaction such as displeasure or distress. By contrast, mindfulness can be conceived as a mental state in which attention is systematically focused in the present moment with a stance of equanimity such as openness, curiosity, and nonjudgmental awareness (Walach, Ferrari, Sauer, & Kohls, 2012).

The hypothesis that mindfulness may increase health and well-being has been empirically corroborated in recent years. Several (meta-analytical) reviews have documented its clinical effectiveness (e.g., Mars & Abbey, 2010). Especially for psychosomatic diagnoses (e.g., anxiety and stress), mindfulness-based interventions have been found to be effective.

Despite promising empirical observations concerning the effects of mindfulness, its measurement has been a target of criticism (Grossman, 2011). As measurement approaches for mindfulness differ considerably from a conceptual point of view (Sauer, Walach, et al., 2013), it seems appropriate to state that research has thus failed to provide a theoretically accepted notion of how mindfulness should be measured.

It is therefore necessary for the advancement of this field to improve its current state of measurement. Hence, it is the aim of the present research to contribute to improving the measurement of mindfulness by scrutinizing the predictive ability of mindfulness scales to gauge whether or not a given individual practices mindfulness training or not. The rationale behind this idea is that an individual with a regular practice of mindfulness should report a higher (self-reported) mindfulness level than an individual who does not regularly practice mindfulness. Although it has occasionally been proposed that mindfulness practitioners may describe themselves as less mindful than individuals naïve to the idea of mindfulness due to a shifting of the internal reference point, the majority of published experimental studies suggest that regular mindfulness training leads to higher self-reported mindfulness levels (Bohlmeijer, Prenger, Taal, & Cuijpers, 2010; Chadwick, Hughes, Russell, Russell, & Dagnan, 2009; Gaylord et al., 2011; Mars & Abbey, 2010). This is not to say that there are no other sizeable problems associated with the quantitative assessment of mindfulness: For example, a positive relation between mindfulness levels and smoking/frequent binge-drinking behavior was found (Leigh, Bowen, & Marlatt, 2005), suggesting that mindfulness instruments may actually tap into other constructs such as conscious sensitivity or bodily awareness. Keeping this limitation in mind, the ability to estimate group allocation can naturally not be expected to be perfect.

We employed a statistical procedure referred to as “Random Forests” (RF). Although RF is widely used in biologically driven research (Bosch, Zisserman, & Muoz, 2007), it is still scarcely used in psychological studies (but see Strobl, Malley, & Tutz, 2009). Evidence supports the predictive strength of the method (Abu-Nimeh, Nappa, Wang, & Nair, 2007). We submitted data

from two mindfulness instruments to an RF procedure in order to gauge the predictive quality of the respective instrument (i.e., predicting class membership of the binary criterion "regular mindfulness practice or not").

2 Methods

2.1 Samples

Two samples were included in this study to allow for replication of the results and in order to investigate the influence of data quality. The total sample size was $N = 278$. Sample 2 was of higher quality and Sample 1 of lower quality (see below).

Sample 1. Sample 1 ($N = 202$) was collected as part of an unpublished online study investigating the relation between mindfulness, health, and emotion; these data have not been published before. About two thirds of the sample ($n = 129$) reported having no prior mindfulness practice, whereas $n = 72$ individuals (36%) reported practicing mindfulness on a regular basis (1 missing value). Examples of types of mindfulness training included Buddhist meditation, Thai Chi, or Yoga exercises. The mean age was 35 years for nonpractitioners ($SD = 13$) and 39 years for practitioners ($SD = 11$); $n = 142$ (71%) persons were female, and $n = 59$ (29%) persons were male (1 missing value).

Sample 2. Sample 2 (N = 76) was described in Sauer, Lemke, et al. (2012). The sample consisted of 38 expert mindfulness practitioners (21 female, 17 male) and an age- and gender-matched group of 38 nonmindfulness practitioners (28 female, 10 male). Practitioners had trained in different Buddhist mindfulness traditions. The mean age was 51 years in both groups (SD = 10). Inclusion criteria were at least 5 years of daily meditation practice for the meditation group and no meditation experience for the control group.

2.2 Instruments

Freiburg Mindfulness Inventory (FMI-14). The conceptual foundation of the FMI-14 is rooted in Buddhist Psychology (Walach, Buchheld, Buttenmüller, Kleinknecht, & Schmidt, 2006), and the instrument was designed to measure mindfulness as a stable trait. Whereas the scale was first developed as a unidimensional scale, recent research supported a two-factor solution (Kohls, Sauer, & Walach, 2009). The two factors have been labeled "Presence," indicating the awareness of stimuli in the subjective now (Sauer et al., 2012), and "Acceptance," indicating a nonjudgmental stance toward all kinds of experience (Kohls et al., 2009). One strength of the instrument is that it has been validated using not only classical psychometric methods such as exploratory and confirmatory factor analysis (Kohls et al., 2009) but also item response theory (Sauer, Walach, Offenbächer, Lynch, & Kohls, 2011a; Sauer, Ziegler, Danay, Ives, & Kohls, 2013). The instrument consists of 14 items with four answer options.

Mindfulness Attention and Awareness Scale (MAAS). The MAAS is one of the most widely used instruments measuring mindfulness as a unidimensional construct with 15 items and six answer options. It is noteworthy that all items are negatively formulated to assess "mindlessness" rather

than “mindfulness.” Although a study suggested that mindlessness should not be conceived as the inverted construct of mindfulness (Van Dam, Earleywine, & Borders, 2010), there are several studies supporting the practical usefulness of the MAAS. We deem the MAAS an adequate candidate for our analysis in addition to the FMI, given that data support its usefulness.

2.3 Statistical method - Random Forests

We employed RF, a statistical method stemming from machine learning contexts to assess the predictive accuracy of the FMI-14 or the MAAS items with regard to their ability to estimate class membership as operationalized by a regular mindfulness practice or a lack thereof (i.e., mindfulness practice as a binary criterion). RF allows statistical properties to be delineated (e.g., nonlinear trends, high-degree interaction, and correlated predictors). In addition, assumptions necessary for classical multivariate analyses such as homoscedasticity (homogeneity of variance), linear associations between variables, or metric variable levels are not warranted (Breiman, 2001).

The term “random forests” is well chosen because in the respective procedure, random subsamples of decision trees are drawn, building together a “forest” of “trees” on a random basis. The RF procedure therefore represents an ensemble, averaging the results of several decision trees based on a majority vote principle. RF predictions of group or class membership are based on the part of the sample that was not used to build the RF framework. This excluded part of the sample is called the “out of bag” (OOB) sample. The advantage of the OOB procedure is that the results can be seen as stemming from a cross-validation sampling procedure, thereby increasing confidence in the results (Breimann, 2001).

2.4 Procedure

We conducted four analyses that could provide answers to the following four pivotal research questions. First, as a baseline test, we wanted to investigate if the group mean value can be seen as a suitable indicator for distinguishing between mindfulness practitioners and non-practitioners. Correspondingly, for each of the two instruments, total mean scale scores were calculated and compared. For the FMI, we additionally computed the mean scores for both factors (i.e., *Presence* and *Acceptance*). We wanted to see whether the mere aggregation of item mean scores into a total mean scale score would contain enough information to allow us to differentiate between the subgroups (practitioners vs. nonpractitioners). We expected that practitioners would exhibit higher mindfulness levels than nonpractitioners.

Second, we tried to determine the (positive and negative) predictive accuracy for each instrument in each sample for predicting class membership (mindfulness practitioners vs. non-practitioners). In other words, we assessed whether the total mean scores correctly predicted regular mindfulness practice or not. Recall that positive accuracy (sensitivity), negative accuracy (specificity), and average accuracy may not necessarily be identical.

Third, the importance of each item for predictive accuracy was investigated by computing a mean decrease in the accuracy (MDA) statistic using the OOB sample. This allowed us to identify the most important items for allocating group membership to the two groups. We were additionally able to scrutinize whether some items such as FMI Item 13, which was identified as

"problematic" by conventional psychometric analyses, would also be recognized as problematic by the RF method.

Fourth, we investigated whether predictive accuracy could be increased by combining the MAAS and FMI items. Due to the substantial differences and only partial conceptual overlap of the scales, it seems plausible that neither of the two scales carries all conceptual information useful for the given classification task.

2.5 Analysis, software, and settings

Given that the FMI and the MAAS use different scales, item scores were first standardized to a range of [0; 1] where higher values indicate higher mindfulness levels. Alpha was set to .05. We used R V 3.0 (R-Core-Team, 2008) on a Windows computer. RF analyses were performed using the package *party* (Hothorn, Hornik, & Zeileis, 2006; Strobl, Boulesteix, Zeileis, & Hothorn, 2007). Only raw data were analyzed, and missing values were not replaced. We grew 1,000 trees, and drew three variables from each tree at each node (default values).

3 Results

3.1 For which of the instruments does the group mean distinguish between mindfulness practitioners and nonpractitioners?

As can be seen in the boxplots (Figure 1), for both the FMI and MAAS, the self-attributed median mindfulness levels of the practitioners were found to be consistently larger than the

respective scores of the nonpractitioners. Especially in the group of mindfulness practitioners, the interquartile distance of the MAAS was determined to be smaller than the interquartile distance of the FMI. As expected due to the differences in sampling quality, the interquartile distance in Sample 2 was found to be smaller than in Sample 1, and extreme values occurred more frequently in Sample 1. Furthermore, an examination of the density plots supported the notion that group allocation could be estimated on the basis of the scale's mean scores (see Figure 2). Moreover, a comparison of the mean scores of the two groups for each scale (see Table 1) showed that the average mindfulness levels of practitioners per scale exceeded the mindfulness levels of the nonpractitioners; that is, all confidence intervals (95%) excluded zero (indicating a statistically significant result at the given α-level).

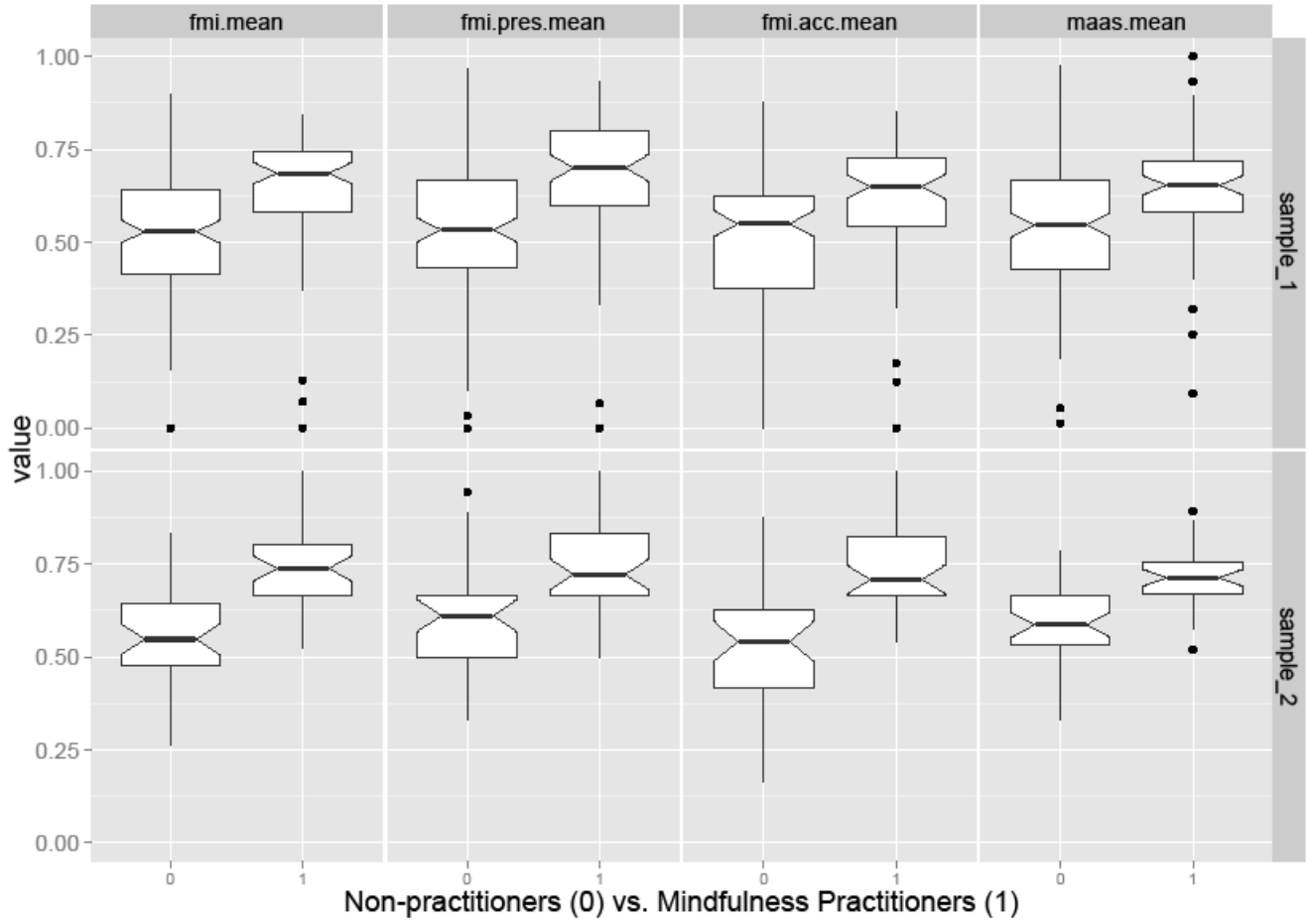

Figure 1. Box plots for the classification task for the two subgroups "nonpractitioners (0)" vs. "mindfulness practitioners (1)" (Sample 1: upper part; Sample 2: lower part). fmi.mean: FMI total mean score. fmi.pres.mean: mean score of the subscale “presence.” fmi.acc.mean: mean score of the subscale “acceptance.” maas.mean: MAAS mean scale score. Notches show 95% confidence intervals of the median.

Table 1. *Confidence intervals of mean differences comparing mindfulness practitioners and nonpractitioners*

sample	scale	CI (95%) low	CI (95%) high
1	FMI total	-0.23	-0.12
1	MAAS	-0.17	-0.08
2	FMI total	-0.14	-0.05
2	MAAS	-0.16	-0.07
1	FMI presence	-.20	-.09
1	FMI acceptance	-.15	-.05
2	FMI presence	-.20	-.08
2	FMI acceptance	-.27	-.14

Note. CI: Confidence interval. Negative values for the mean differences (CI borders) indicate that the nonpractitioner group had lower values than the practitioner group. As zero was not included in any of the four data sets, the difference between the two subsamples in all four data sets was statistically significant at $p \leq .05$ *(two-tailed).*

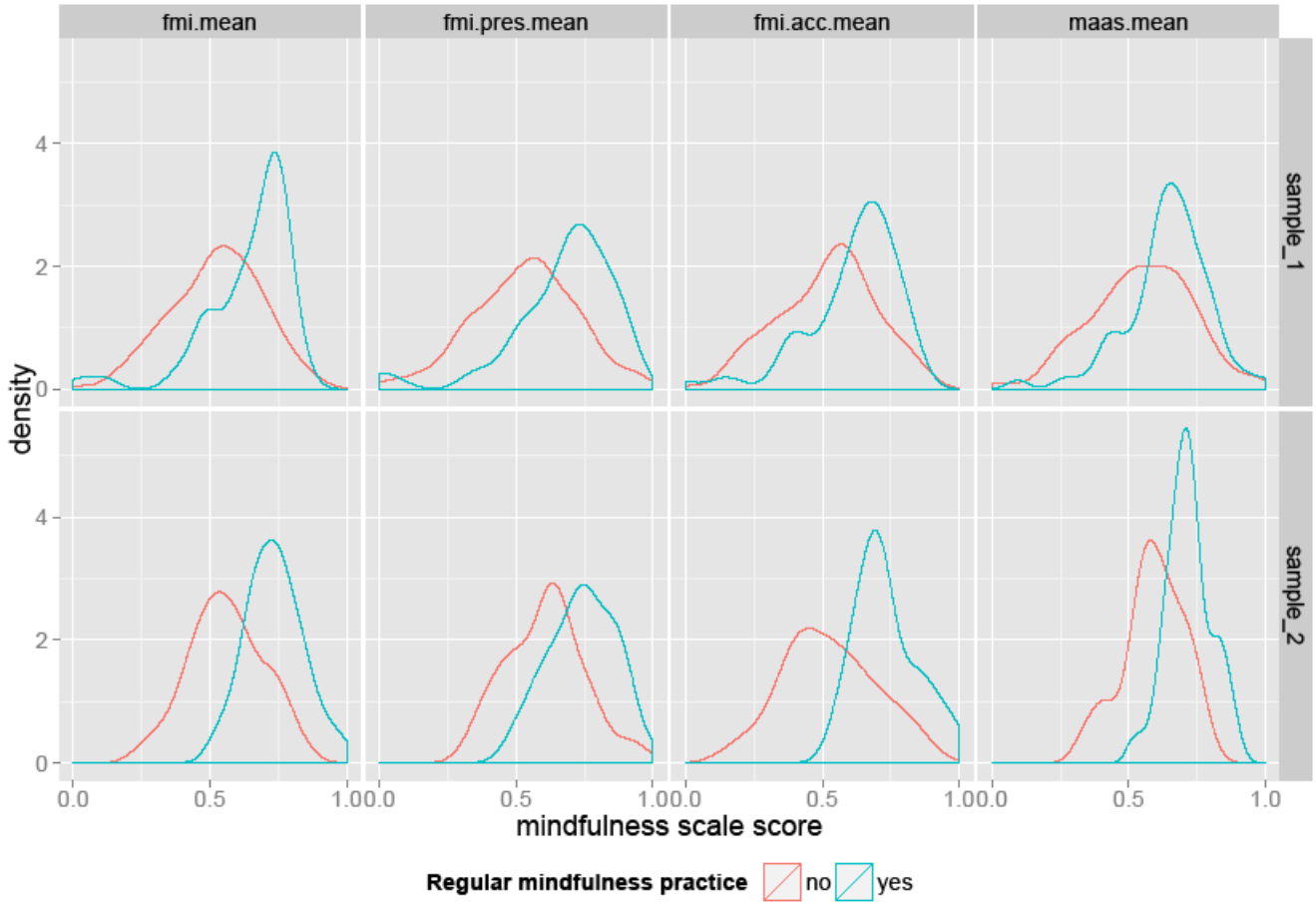

Figure 2. Density plots for the mindfulness scale and factor scores for subjects with and without regular mindfulness practice for the FMI and MAAS (Sample 1: upper part; Sample 2: lower part). fmi.mean: FMI total mean score. fmi.pres.mean: mean score for the subscale "presence." fmi.acc.mean: mean score for the subscale "acceptance." maas.mean: MAAS mean scale score.

In summary, these results corroborate previous findings by suggesting that the total scale scores of both the FMI and the MAAS can be used to distinguish between mindfulness practitioners and nonpractitioners.

3.2 What is the (positive and negative) classification accuracy for each instrument and in each sample in predicting class membership?

Average classification accuracy was computed with the following algorithm:

Classification accuracy = (correct positives + correct negatives)/(all positives + all negatives)

The percentage of practitioners in the two samples served as a baseline for gauging the classification accuracy (Sample 1 = 72/202 = 36%; Sample 2: 38/76 = 50%). Precision gain was computed as follows:

Gain = classification accuracy/percentage of practitioners.

A comparison of the classification error rates (see Table 2) revealed that in Sample 2, average classification accuracy was higher than in Sample 1 for all cases. However, further inspection of sensitivity and specificity revealed that for both instruments the poorer accuracy in Sample 1 was due to the low sensitivity (i.e., a considerable number of practitioners were not correctly identified as practitioners but were falsely identified as nonpractitioners). With regard to the FMI, the acceptance factor of the FMI exhibited marginally weaker sensitivity and specificity indices than the presence factor. However, each subfacet was found to show only slightly weaker parameters than the total FMI scale. In direct comparison, classification accuracy was found to be higher for the MAAS.

Table 2. *Predictive accuracy*

Instrument	sample	sensitivity	specificity	accuracy	gain
FMI	1	0.51	0.91	0.77	2.15
FMI	2	0.92	0.73	0.83	1.63
FMI Presence	1	0.53	0.91	0.77	2.15
FMI Presence	2	0.92	0.70	0.81	1.61
FMI Acceptance	1	0.47	0.91	0.75	2.10
FMI Acceptance	2	0.89	0.70	0.80	1.58
MAAS	1	0.46	0.94	0.77	2.14
MAAS	2	0.95	0.78	0.87	1.71
FMI-MAAS combined	1	0.54	0.94	0.80	2.22
FMI-MAAS combined	2	0.92	0.81	0.87	1.71

Note. Accuracy reflects the total proportion of correctly identified persons. Gain shows the ratio of mean classification accuracy in accordance with the proportion of practitioners.

3.3 How important is each item for predictive accuracy?

Item importance measures can be used to identify the most important variables for gauging the predictive accuracy of a scale. This indicator may thus conceptually be compared to *B* in

traditional regression terminology. As can be seen in Figure 3, the importance of each item—as measured by the MDA—differed considerably for each instrument between the two samples.

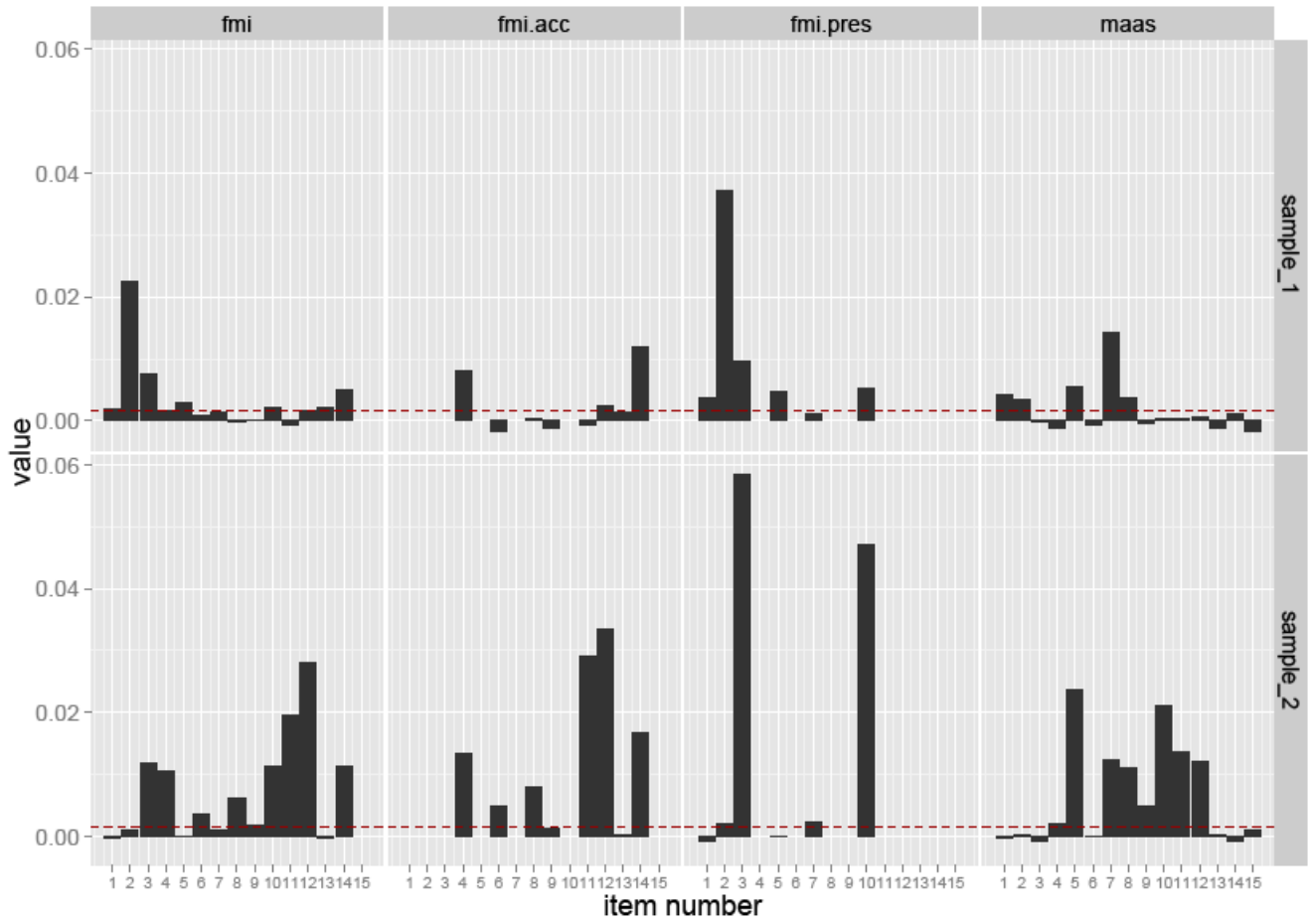

Figure 3. Item importance for predictive accuracy computed on the basis of the mean decrease in accuracy after permutation (MDA). (Sample 1: upper part; Sample 2: lower part). The length of the bar indicates a higher importance for total classification accuracy. The red dashed horizontal line indicates the relevance threshold (suggested by Strobl et al., 2009, as a rule of thumb with the rationale that positive scores up to the same magnitude as negative scores are due to chance alone; this is not a test of statistical significance). fmi.mean: FMI mean score across all items. fmi.pres.mean: mean score of the FMI presence items. fmi.acc.mean: mean score of the FMI

acceptance items. maas.mean: MAAS mean score across all items. The item numbers (14 for the FMI and 15 for the MAAS) are shown on the x-axis. The importance (for predictive accuracy; MDA) of the items is depicted on the y-axis.

Regarding the MAAS, a pronounced difference in the MDA scores was found in the two samples. In Sample 2, the MDA of the variables was found to be considerably higher than in Sample 1, possibly as a consequence of the higher sample quality reported previously. It is noteworthy that in Sample 1, only five of the 15 items exhibited a substantial contribution to MDA, indicating that 10 items did not contribute to the classification accuracy of the MAAS. By contrast, in Sample 2, eight items exhibited a substantial contribution to the MDA scores. Focusing the analyses on the items with the highest contribution to MDA, only two items (Item 7: "It seems I am ‘running on automatic’ without much awareness of what I’m doing"; Item 8 "I rush through activities without being really attentive to them") succeeded in both samples. In sum, in the high-quality sample, about half of the items contributed to classification ability (Items 5 and 7-12). In the reduced-quality sample, only a few items were able to differentiate class membership (mainly Item 7). Summing across the items from both samples, 45% of the items fell above the relevance threshold (i.e., at least 55% of the items were unlikely to have contributed to predictive accuracy).

Regarding the FMI, results were comparable. In the lower quality sample, only one item exhibited a high contribution to MDA (Item 2: "I sense my body, whether eating, cooking, cleaning, or talking"). Similar to the MAAS, in the higher quality sample, about half of the items contributed to the classification ability of the total scale (Items 3, 4, 5, 7, 8, 9, 10, 12) including

both presence and acceptance items. The strongest item regarding MDA contribution was Item 12 ("I experience moments of inner peace and ease, even when things get hectic and stressful"). Interestingly, Item 13 failed to contribute to predictive accuracy; especially in the high-quality sample, the contribution of Item 13 was negligible. For the presence subfacet, 75% of the items from both samples fell above the relevance threshold. For the acceptance subfacet, 56% of the items from both samples fell above the relevance threshold.

In total, both scales included "weak" items in the sense that many items had little ability to differentiate between mindfulness practitioners and nonpractitioners. In this analysis, we found little evidence for a stable sample-independent predictive accuracy for the majority of the items. It appears that the MAAS includes a larger number of uninformative items than the FMI.

3.4 Can incremental accuracy be gained by combining the instruments?

When all items of both instruments were entered as predictors of class membership, the overall predictive gain in Sample 1 increased to 2.22 and to 1.71 in Sample 2 (see Table 2). Thus, the overall classification rate was not much higher than for each of the individual instruments. It may be inferred that predictive accuracy can be only marginally improved by combining the item sets of the FMI and MAAS.

4 Discussion

In this study, we examined the psychometric quality of instruments assessing mindfulness as a trait variable by employing a novel analytical framework called random forests.

As expected, predictive power was better in the high-quality Sample 2, which consisted of carefully selected mindfulness experts and age- and gender-matched controls with no mindfulness practice, compared to the lower quality Sample 1, which was collected using an uncontrolled online-sampling procedure. The predictive accuracy of the MAAS was better than the FMI in both samples.

In addition, the predictive quality of the two FMI factors did not show considerable differences between the samples. These findings shed new light on the ongoing debate of whether mindfulness should be conceived as a unidimensional or two-dimensional construct. Our results show that predictive accuracy improves if mindfulness is divided into the two factors of presence and acceptance. A further finding was that items that exhibited high predictive accuracy when only one facet alone was taken into account were not necessarily powerful when the two facets were combined (i.e., when all items of the FMI were combined). This may occur for at least two reasons. First there may be spurious correlations between some items and the dependent variables, and these spurious correlations are eliminated when more items are included, thereby controlling for such spurious associations (Kohls et al., 2009). Second, there may be other types of interactions that occur between the items, thus blurring the differential effects. Previous research speaks in favor of the second interpretation (Sauer, Walach, Schmidt, et al., 2011).

In a similar vein, collapsing all items from the FMI and the MAAS into one set of items did not improve the accuracy of class membership prediction. This fact warrants further elaboration as the scales are conceptually different. Whereas the FMI was designed to measure "mindfulness" on the basis of both an attention and an emotion facet, the MAAS taps into "mindlessness," thereby solely building upon a lack of attention as a unidimensional construct. However, further research is warranted to determine whether this result can be seen as supporting the notion that the acceptance/emotion facet is functionally redundant.

Our results also indicate that one particular aspect of the predictive accuracy of mindfulness instruments, namely sensitivity, may be particularly vulnerable to a lower sample quality. As sensitivity figures in the low-quality sample were barely above the base rates for both the MAAS and the FMI, it can be concluded that for both scales, the algorithm had trouble correctly identifying mindfulness practitioners. We must remember that the prediction scores from the RF analysis are based on OOB data (i.e., they are from a cross-validation part of the sample and not from the part of the sample that was used to build the model). However, as the algorithm achieved high specificity in Sample 1 and high overall accuracy in Sample 2, we suggest for the time being that this may be a specific problem that is associated with sensitivity rates.

Focusing on a lower level of measurement units, our results are also somewhat discouraging: Only about half of the items of both scales were able to add to the predictive accuracy in the high-quality sample, whereas in the low-quality sample, only one or two items contributed to subgroup differentiation. This finding is disappointing as it suggests that many of the items that we investigated are in need of revision. For example, FMI Item 13, an item found to be plagued

by psychometric problems in previous analyses (Sauer, Walach, Offenbächer, Lynch, & Kohls, 2011b; Sauer, Ziegler, et al., 2013), did not contribute to predictive accuracy. Nevertheless, as several other items that were not identified as problematic by previous analyses showed similar results, no strong conclusions can be drawn. With regard to the MAAS, there were also problems with several mindfulness items. This interpretation is corroborated by Van Dam, Earleywine, and Borders' (2010) findings, which suggested that only five of the 15 items were informative. Interestingly, four of these five items were also identified by our analysis.

In sum, whereas our study seems to have revealed several shortcomings of two established mindfulness instruments—one of which we had the pleasure to contribute to substantially—a word of caution is needed. First, we used a novel approach, with which there is naturally a lack of experience and comparability. Second, our dependent variable—mindfulness practice or lack thereof—is *ipso facto* not the only aspect that mindfulness instruments should be able to differentiate. Although we decided to employ this binary selection criterion for pragmatic reasons, it may well be the case that, for other dependent variables (e.g., psychological health, well-being), the predictive accuracy may be different. Nevertheless, even when applying a seemingly straightforward binary criterion, problems naturally associated with the self-attribution of inner states can arise: Although response shift and other form of biases may actually impair sensitivity indices, it may also be argued that not all mindfulness practitioners have higher (self-reported) mindfulness levels than nonpractitioners. We opine that it would therefore appear naïve to assume clear-cut, homogeneous results—even when a powerful procedure such as random forests is applied. Nonetheless, it may well be the case that response shift and other forms of biases relevant for assessing self-attributed mindfulness levels may

occur as a consequence of a regular mindfulness practice. Our data may suggest that some mindfulness practitioners may actually tend to report lower self-attributed mindfulness scores, probably as a consequence of their sensitivity to states of mindlessness. We believe that there are presently only two remedies to that problem: (a) Researchers should strive to implement well-controlled experimental settings, and (b) researchers should endeavor to employ supplementary measurement approaches in addition to the quantitative measurement paradigm such as bistable images (Sauer et al., 2012).

In summary, the results shed new light on the quality of the instruments. It appears that many of the MAAS and FMI items are unable to reliably predict class membership. A substantial need for revision may be deduced. However, both instruments showed good predictive accuracy overall—the probability of correctly classifying an individual as a nonpractitioner was twice as high as chance would suggest.

References

Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). A comparison of machine learning techniques for phishing detection. Proc eCrime Res Summit 2007.

Bohlmeijer, E., Prenger, R., Taal, E., & Cuijpers, P. (2010). The effects of mindfulness-based stress reduction therapy on mental health of adults with a chronic medical disease: A meta-analysis. J Psychosom Res., *68*(6), 539–544.

Bosch, A., Zisserman, A., & Muoz, X. (2007). Image classification using random forests and ferns. In Computer Vision, 2007 ICCV . Rio de Janeiro.

Breiman, L. (2001). Random forests. Mach Learn., *45*(1), 5–32.

Chadwick, P., Hughes, S., Russell, D., Russell, I., & Dagnan, D. (2009). Mindfulness groups for distressing voices and paranoia: a replication and randomized feasibility trial. Behav Cogn Psychother., *37*, 403–412. doi:10.1017/S1352465809990166

Gaylord, S. A., Palsson, O. S., Garland, E. L., Faurot, K. R., Coble, R. S., Mann, J. D., … Whitehead, W. E. (2011). Mindfulness Training Reduces the Severity of Irritable Bowel Syndrome in Women: Results of a Randomized Controlled Trial. Am J Gastroenterol., *106*(9), 1678–1688.

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). Psychol Assess., *23*(4), 1034–1040.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. J Comput Graph Stat., *15*(3), 651–674.

Kohls, N., Sauer, S., & Walach, H. (2009). Facets of mindfulness–Results of an online study investigating the Freiburg Mindfulness Inventory. Pers Individ Dif., *46*(2), 224–230.

Leigh, J., Bowen, S., & Marlatt, G. A. (2005). Spirituality, mindfulness and substance abuse. Addict Behav., *30*(7), 1335–1341. doi:10.1016/j.addbeh.2005.01.010

Mars, T. S., & Abbey, H. (2010). Mindfulness meditation practise as a healthcare intervention: A systematic review. Int J Osteopath Med., *13*(2), 56–66. doi:10.1016/j.ijosm.2009.07.005

R-Core-Team. (2008). R: A language and environment for statistical computing. R Found Stat Comput., *3*(10).

Sauer, S., Lemke, J., Wittmann, M., Kohls, N., Mochty, U., & Walach, H. (2012). How long is now for mindfulness meditators? Pers Individ Dif., *52*(6), 750–754.

Sauer, S., Walach, H., Offenbächer, M., Lynch, S., & Kohls, N. (2011a). Measuring Mindfulness: A Rasch Analysis of the Freiburg Mindfulness Inventory. Religions., *2*(4), 693–706.

Sauer, S., Walach, H., Offenbächer, M., Lynch, S., & Kohls, N. (2011b). Measuring mindfulness: a Rasch analysis of the Freiburg mindfulness inventory. Religions., *2*(4), 693–706.

Sauer, S., Walach, H., Schmidt, S., Hinterberger, T., Horan, M., & Kohls, N. (2011). Implicit and explicit emotional behavior and mindfulness. Conscious Cogn., *20*(4), 1558–1569.

Sauer, S., Walach, H., Schmidt, S., Hinterberger, T., Lynch, S., Büssing, A., & Kohls, N. (2013). Assessment of mindfulness: review on state of the art. Mindfuln., *4*(1), 3–17.

Sauer, S., Ziegler, M., Danay, E., Ives, J., & Kohls, N. (2013). Specific Objectivity of Mindfulness—A Rasch Analysis of the Freiburg Mindfulness Inventory. Mindfuln., *4*(1), 45–54.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics., *8*(1), 25. doi:10.1186/1471-2105-8-25

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. Psychol Methods., *14*(4), 323–348.

Van Dam, N. T., Earleywine, M., & Borders, A. (2010). Measuring mindfulness? An Item Response Theory analysis of the Mindful Attention Awareness Scale. Pers Individ Dif., *49*(7), 805–810. doi:10.1016/j.paid.2010.07.020

Walach, H., Buchheld, N., Buttenmüller, V., Kleinknecht, N., & Schmidt, S. (2006). Measuring mindfulness—the Freiburg Mindfulness Inventory (FMI). Pers Individ Dif., *40*(8), 1543–1555.

Walach, H., Ferrari, M.-L. G., Sauer, S., & Kohls, N. (2012). Mind-Body Practices in Integrative Medicine. Religions., *3*(1), 50–81. doi:10.3390/rel3010050