Mindful Machine Learning: Using Machine Learning Algorithms to Predict the Practice of Mindfulness

Article *in* European Journal of Psychological Assessment · April 2016

DOI: 10.1027/1015-5759/a000312

Mindful Machine Learning: Using Machine Learning Algorithms to Predict the Practice of Mindfulness

"This article does not exactly replicate the final version published in the journal at European Journal of Psychological Assessment. It is not a copy of the original published article and is not suitable for citation."

This manuscript is dated as May 2015. In its present form the manuscript has not yet been accepted for publication.

Sebastian Sauer[1,2], Ricardo Büttner[1], Thomas Heidenreich[3], Jana Lemke[4], Christoph Berg[1], Christoph Kurz[5]

1 FOM University of Applied Sciences

2 Samueli Institute

3 Esslingen University of Applied Sciences

4 Viadrina University Frankfurt/Oder

5 Helmholtz Zentrum München

Author Note

Correspondence concerning this paper should be directed to Sebastian Sauer, FOM University of Applied Sciences, Arnulfstr. 30, 80335 Munich, Germany; Email: Sebastian.sauer@fom.de, Phone +49 89 202 452-0, Fax +49 89 202 452-28

Abstract

Mindfulness refers to a stance of nonjudgmental awareness of present-moment experiences. A growing body of research suggests that mindfulness may alter the brain in a way that increases cognitive resources, thereby buffering stress. However, the measurement of mindfulness has been the subject of severe criticism. Existing models have essentially failed to achieve a consensus on how mindfulness should be operationalized. As the sound measurement of mindfulness is the foundation needed before substantial hypotheses can be examined, we propose a novel way of gauging the psychometric quality of a mindfulness measurement instrument. The Freiburg Mindfulness Inventory (FMI) was selected for this purpose. Specifically, we used predictive modeling from machine learning contexts. We employed 10 of the best performing algorithms to predict whether an individual from a sample of $N = 276$ was a mindfulness practitioner or not. A high predictive accuracy of class membership can be taken as an indicator of the psychometric quality of the instrument. In sum, our results indicate that the FMI was able to reliably predict class membership. However, many items appeared to be uninformative. From an applied methodological point of view, it appears that machine learning algorithms can valuable for predictive psychological research.

Keywords: machine learning, predictive modeling, mindfulness, measurement

Mindful Machine Learning: Using Predictive Algorithms to Predict the Practice of Mindfulness

The idea that mindfulness may be able to alleviate a variety of psychosomatic symptoms has gained widespread currency in recent years (Chiesa, Calati, & Serretti, 2011; Hofmann, Sawyer, Witt, & Oh, 2010). Mindfulness as a state of consciousness may be characterized by an attentional stance that focuses on the present moment ("presence") with a nonjudgmental attitude toward all perceptions and sensations ("acceptance") that may be part of the actual experience (Sauer, Walach, et al., 2013). It implies a sense of being in the present, underlining the vivid awareness of sensory and mental experience, and calls for a voluntarily disruption of automatic mental processes of cognitive evaluation such as "good," "bad," "wanted," or "unwanted" as well as the suspension of labeling experience with emotional tones.

Several meta-analytical reviews have investigated the relations between mindfulness, health, and well-being and have corroborated the health-relevant effects of mindfulness (Eberth & Sedlmeier, 2012; Sedlmeier et al., 2012; Zainal, Booth, & Huppert, 2013). Symptoms for which health-relevant effects of mindfulness have been found include depression, anxiety, stress, pain, substance abuse, personality disorders, and elevated blood pressure, among others.

Results from neuroscientific research (e.g., research applying brain imaging techniques) have begun to reveal the neurological basis of mindfulness and have linked mindfulness back to several brain regions that are known to be involved in attention and emotion-regulation processes such as the Anterior Cingulate Cortex or the Insula (Farb, Anderson, & Segal, 2012; Tang & Posner, 2013; Teper & Inzlicht, 2013). A recent review suggests that mindfulness may buffer age-related neural decay, thus preventing the brain from aging prematurely (Gard, Hölzel, & Lazar, 2014).

In the light of a promising new area of psychological research, rigorous examination is in place for scrutinizing the claims that mindfulness-based health interventions have the potential to improve health. It appears that the pivotal research question at the moment is whether or how mindfulness can be measured (Baer, 2011), and thus, it is noteworthy that more than a dozen instruments have been published (Sauer, Walach, et al., 2013).

Yet, no consensus on how mindfulness should be measured has emerged. On the contrary, severe criticism of the common approaches has been put forth (Grossman, 2008, 2011). In sum, it appears that the measurement of mindfulness remains unsatisfactory. In the light of this issue, it is the aim of the current research to further the measurement of mindfulness. To this end, we suggest that the methodology that is currently employed, as sophisticated as it may be, is insufficient and should be amended by novel methods. Machine learning may provide such a method that is novel to psychometrics. Machine learning is an umbrella term for a dynamic array of algorithms for predictive modelling that can supplement more “traditional” statistical approaches. Albeit not recognized much in psychology to date, in other fields such as gene expression, machine learning algorithms can now be considered a standard instrument (Dudoit, Fridlyand, & Speed, 2002; Shipp et al., 2002; Ye et al., 2003). Such fields are still debating which of the many algorithms can be considered superior to others, or, strictly speaking, which algorithms fit with which problems and data. A practical recommendation is, when in doubt, that several algorithms should be used and the results should be aggregated (Hastie, Tibshirani, & Friedman, 2009). This approach was taken in the present paper.

More specifically, we used 10 different predictive algorithms to test whether the items from a well-known mindfulness inventory would be able to predict whether the participants were practicing mindfulness on a regular basis or not. Put bluntly, we wanted to test for a correspondence between the “saying” and “doing” of mindfulness. Assuming that the practice of mindfulness (i.e., whether someone regularly practices mindfulness or not) is measured reliably, mindfulness as measured via self-report can be seen as the parameter that drives the correspondence between the “saying” and “doing” of mindfulness. This idea is by no means new; however, we used a variety of machine learning algorithms, an approach that is novel to the area of mindfulness and to many other areas of psychology as well.

We hypothesized that some algorithms would be substantially more adequate for estimating the accuracy of class membership (mindfulness practitioners vs. nonpractitioners) than others. Following this line of reasoning, an averaged accuracy statistic should be more adequate than employing one single algorithm—given that it is not known in advance which algorithm is suitable. In our case, there was no prior research that had addressed the question of which algorithm was best suited for the problem and the data at hand. Our hypothesis would be falsified if the algorithms performed similarly. On the other hand, if the accuracy of the algorithms differed greatly, our hypothesis would be tentatively supported (although of course not proven).

In sum, our approach was built on predicting the class membership of the practice of mindfulness (i.e., practitioner vs. nonpractitioner) of each participant in our sample. The plan was to take the accuracy of the classification as an indicator of the validity of the scale. In order to arrive at more stable results, we compared 10 machine learning algorithms and used a cross-validation procedure.

Machine learning methods are aimed at identifying patterns in data in order to build and train algorithms to detect such patterns in data sets that were previously unseen (Hastie et al., 2009). Some of these algorithms are able to "learn" in the sense that the algorithms automatically adapt during learning phases such that they become better able to predict parameters or identify cases. A typical well-known machine learning algorithm is neural nets.

Current Mindfulness Measurement Practices

Utilizing psychometric scales is the standard approach for measuring mindfulness. Consequently, a substantial and growing number of scales have been published (Baer, 2011; Sauer, Walach, et al., 2013). The advantages of the psychometric approach are straightforward: convenient and quick application, well-known methodology, and empirical support (Mars & Abbey, 2010).

At the same time, there are shortcomings that may place considerable limits on the validity of the psychometric method (Grossman, 2008). One argument is that a wide array of conflicting definitions exists. For example, the concept of mindfulness put forward by Langer (Langer & Moldoveanu, 2000) differs substantially from Brown and Ryan (Brown & Ryan, 2003) or from Kabat-Zinn (Bishop et al., 2004), to name some of the most influential researchers. As a result of different conceptions and theoretical backgrounds, the operationalizations (i.e., the questionnaires) differ greatly, for example, in the number of supposed factors and the aspects deemed central to the construct (Grossman, 2011). The most widely used psychometric scales include the Five Factor Mindfulness Inventory (Bohlmeijer, Ten Klooster, Fledderus, Veehof, & Baer, 2011), the Mindfulness Attention and Awareness Scale (Brown & Ryan, 2003), and the Freiburg Mindfulness Inventory (Kohls, Sauer, & Walach, 2009).

Method

Sample

We included two samples in this study so that we could replicate the results and estimate the influence of data quality. The total sample size was $N = 276$. As Sample 2 was comparatively small, we refrained from analyzing each sample individually and report the combined results only. For the same reason, we used the *k*-fold cross-validation statistical procedure (see below) as the replication procedure rather than replicating the analysis on the second sample. Data can be accessed in Electronic Supplementary Material 5.

Sample 1. Sample 1 ($N = 201$) was collected as part of an unpublished online study investigating the relations between mindfulness, health, and emotion. About two thirds of the sample ($n = 129$) did not indicate that they engaged in any mindfulness training, whereas $n = 72$ individuals (36%) reported practicing mindfulness on a regular basis (1 missing value). Examples of types of mindfulness training included Buddhist meditation or Yoga. Their mean ages were 35 years for nonpractitioners ($SD = 13$) and 39 years for practitioners ($SD = 11$); $n = 141$ (70%) persons were female, and $n = 59$ (29%) persons were male (1 missing value). The inclusion criterion for the mindfulness group was "regular mindfulness meditation practice."

Sample 2. Sample 2 ($N = 75$) was comprised of one half ($n = 38$) expert mindfulness practitioners (21 female, 17 male) and an age- and sex-matched half of ($n = 37$) nonmindfulness practitioners (28 female, 10 male). Practitioners were trained in different Buddhist mindfulness traditions. The mean age was 51 years in both groups ($SD = 10$ years). Care was taken to obtain distinct groups (i.e., mindfulness practitioners vs. nonpractitioners). The inclusion criteria were at least 5 years of daily meditation practice for the mindfulness group and no meditation experience for the control group. See Sauer, Lemke, et al. for details (2012).

Analytical Procedure

The analytical procedure consisted of predicting class membership (mindfulness practitioners vs. nonpractitioners) using 10 different algorithms. The predictive accuracy (see below) was then taken as an indicator of the psychometric quality of the mindfulness measurement instrument.

One of the dangers of predictive data analysis is over-fitting. Over-fitting refers to a situation in which a statistical model is excessively complex and thereby describes random noise as a systematic pattern. A possible solution to this problem is to split the data into a training set and a test set (Hastie et al., 2009). The model is then built on the training set only; for model results, only the test set is used. We used a well-known and more stringent variant known as the repeated k-fold cross-validation procedure. Here, the data are split into k distinct blocks of similar size. In our case, we split the data into $k = 10$ blocks and repeated this procedure for $n = 10$ repetitions. Then, in n runs, each of the k blocks was left out, and the model was fit to the remaining 9/10 of the sample. The model results, built on the "left out" 1/10, were then averaged. In n-repeated k-fold cross-validation, n versions of the folds are created and aggregated (Kuhn, 2008).

All algorithms were taken from the R package *caret* (Kuhn, 2008). The statistical software *R* 3.0.1 on a Windows 7 computer was used as the statistical environment. Item scores were standardized to fall within the interval of 0 to 1. We used the default values of the statistical procedures (see R-syntax for details, Electronic Supplementary Material 4).

To gauge the predictive quality of the mindfulness instrument, we computed the accuracy of predicting the class membership (mindfulness practitioner vs. nonpractitioners). More specifically, we split the predictive accuracy into the two aspects of sensitivity and specificity. Specificity is defined as the rate of mindfulness practitioners correctly identified as such ("true positives"). On the other hand, specificity refers to the rate of correctly identified nonpractitioners ("true negatives"). In addition, we computed Cohen's Kappa (Fleiss & Cohen, 1973) as this statistic corrects for chance concordance. For this statistic, 0 indicates no concordance between predicted and actual values. According to Landis and Koch (1977), $.41 < \kappa < .60$ can be seen as moderate concordance, with smaller values indicating only slight or poor concordance and higher values indicating substantial concordance.

Instruments

The psychometric instrument under investigation was the Freiburg Mindfulness Inventory-short version (FMI-14). The instrument was designed to measure mindfulness as a stable trait. Whereas the scale was first developed to be a unidimensional scale, a recent study indicated a two-factor solution (Kohls et al., 2009). It has been validated not only with classical psychometric methods such as exploratory and confirmatory factor analyses (Kohls et al., 2009) but also with item response theory (Sauer, Walach, Offenbächer, Lynch, & Kohls, 2011). The instrument in its current form consists of 14 items with four answer options. The instrument has been employed in a number of studies (Eisendrath et al., 2008; Leigh, Bowen, & Marlatt, 2005; Sauer, Walach, & Kohls, 2010).

Predictive Models

We chose algorithms mainly according to their predictive track records as well as on the basis of (Hastie et al., 2009) recommendations. As each employed algorithm has been well-documented, we will provide only a brief introduction to each method in Electronic Supplementary Material 1 (see Hastie et al., 2009 for details). The following models were employed: Generalized linear model (*glm*), Neural networks (*nn*), Boosted classification trees (*ada*), Extreme learning machines (elm), Stochastic Gradient Boosting machine (*gbm*), k nearest neighbors (*knn*), Quadratic discriminant analysis (*qda*), Random Forests (*rfs*), Support vector machine with linear kernel (*SvmLinear*), Support vector machine with polynomial kernel function (*svmPoly*).

Results

As expected, the overall mindfulness level of the mindfulness practitioner group was substantially higher than that of the nonpractitioners (the average across all items; see Figure C1 in Electronic Supplementary Material 3). This was true for both the training and test samples (average in the training sample: .54 vs. .68 for mindfulness practitioners vs. nonpractitioners [scale from 0 to 1], test sample: .50 vs. 67). In other words, mindfulness exhibited a statistically significant influence on the overall mindfulness level in a linear regression (b = .17; SE = .04, t = 3.87, p < .0001). The picture was similar at the item level (Figure 2). The mean difference between the two groups was a Cohen's d of 0.47 (averaged across all 14 items). The items that separated the mindfulness practitioners most strongly from the nonpractitioners in terms of Cohen's d were items 1, 5, and 14 (Figure 3).

Descriptive statistics for all items (Table B1) as well as a visualization of the correlation matrix (Figure C2) can be found in the Electronic Supplementary Material 2 and 3, respectively.

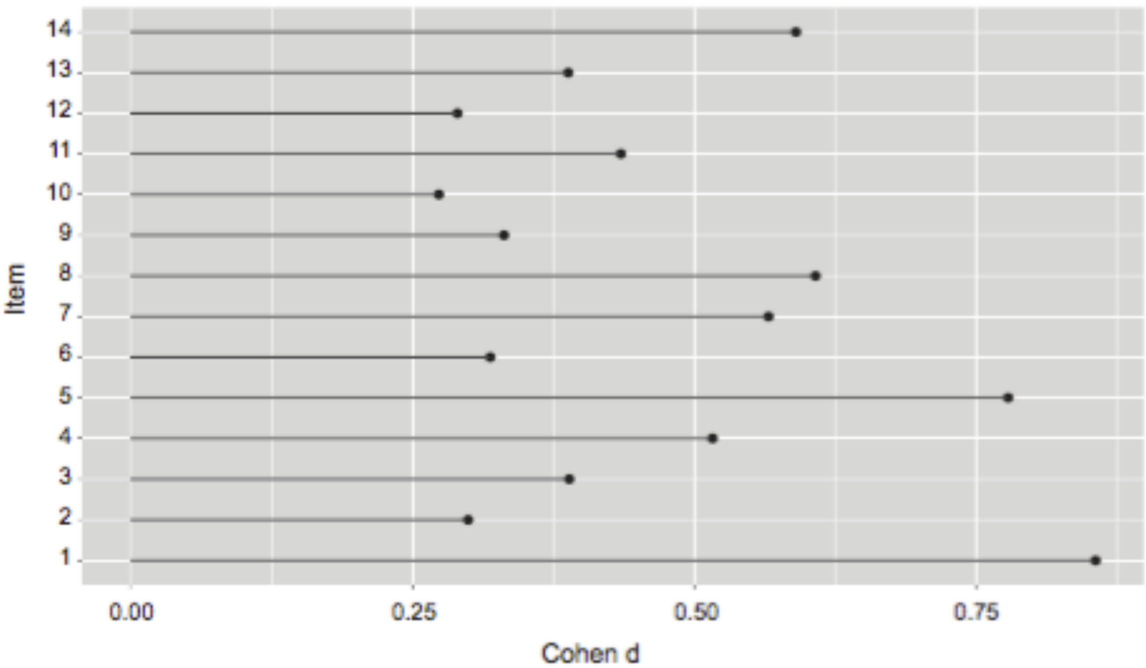

Figure 1. Cohen's d (mindfulness trait level differences) for each item. This figure shows the effect size difference (Cohen's d) in individual mindfulness levels between mindfulness practitioners and nonpractitioners.

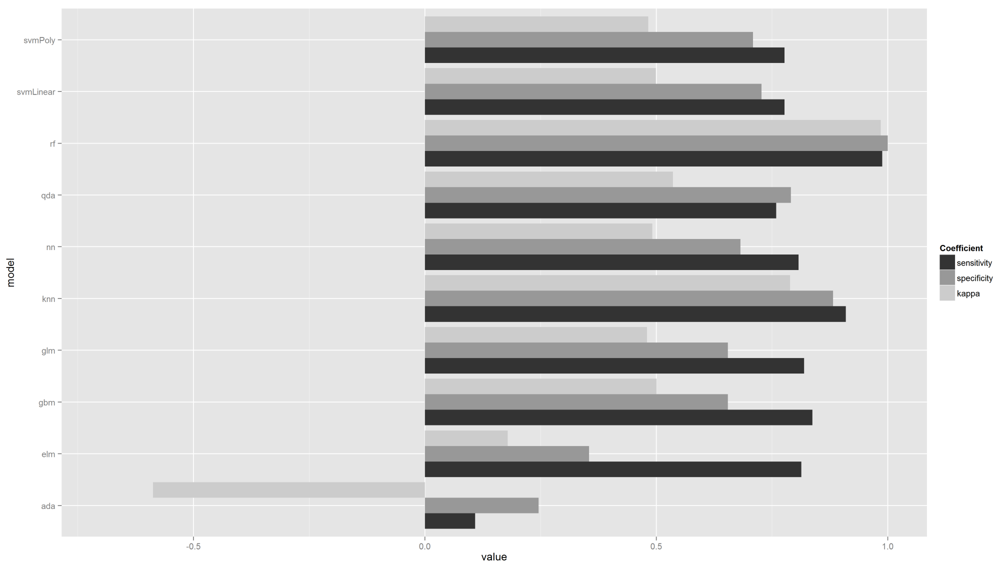

Figure 2. Results of predictive models. This figure shows sensi- tivity, specificity, and the kappa coefficient for each of the 10 predictive models.

For the predictive models, the (arithmetic) average sensitivity across all 10 models was .76 (*Mdn* = .81; *SD* = .24). The average specificity was .67 (*Mdn* = .70; *SD* = .23). These values must be judged against the baseline rate of nonpractitioners of .60. Thus, the ratios were 1.27 (average sensitivity/baseline) and 1.12 (average specificity/baseline). All models except *ada* had sensitivity values above the baseline. All models except *ada* and *elm* had specificity values above the baseline (see Figure 4). The best algorithm in terms of accuracy was, by far, the *rfs* algorithm (sensitivity = .99; specificity = 1). For Cohen's kappa, the picture was similar. The average kappa was moderate in size (*M* = .44; *Mdn* = .50).

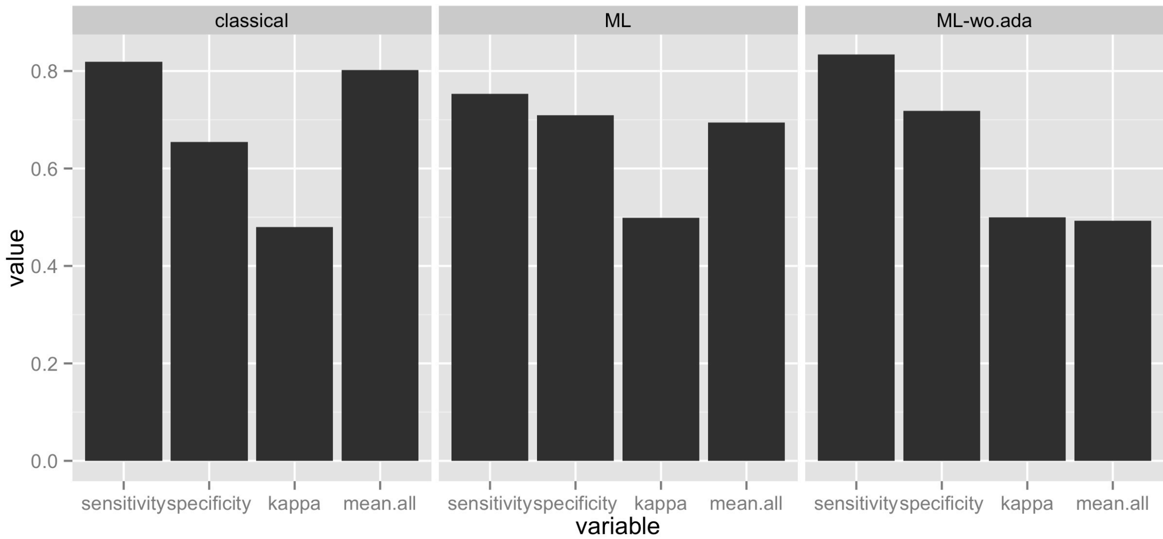

Figure 3. Comparison of predictive strengths of different types of models. The figure shows the predictive strength of (logistic) regression (glm) as a "clas- sical" model (a), machine learning (ML) models (b), and ML models without ada (ML-wo.ada), a model which performed exceptionally weak (c). The values (i.e., sensitivity, specificity, kappa) of the ML models are based on the average per- formance of the rest of the models presented in this research.

Interestingly, the accuracy of the 10 algorithms differed substantially (range [max – min] for sensitivity: .88; range for specificity: .76). It was clear that relying on a single algorithm could lead to poor predictive performance. However, note that the range for accuracy was primarily due to one algorithm (*ada*). In sum, the data provided support for our hypothesis that using an aggregated accuracy statistic would be more adequate than relying on one algorithm alone.

As the *rf* algorithm allowed us to gauge the relative "importance" of the predictor variables (i.e., the items) and because this model was the most accurate, it is interesting to know the importance of the items. The importance of the items can be seen in their contribution to predictive accuracy. More precisely, the *rfs* procedure uses the following rationale. Step 1: compute the overall accuracy of the model as usual. Step 2: randomly permutate the values of a given predictor. Step 3: rebuild the model with the "destroyed" predictor. Step 4: compute the predictive accuracy with the destroyed, (i.e., randomly permuted) predictor. The mean decrease in accuracy across all trees in the forest can be taken as an estimation of the overall importance of the item. Several coefficients exist; we used the Gini coefficient for this purpose. In short, the Gini coefficient indicates the equality or evenness in a distribution. A low Gini coefficient is indicative of a more equal distribution, whereas higher Gini coefficients indicate a more unequal distribution. As the goal of decision trees and random forests consists of creating a distribution that is as equal (homogeneous) as possible, the Gini coefficient is frequently used to assess the importance or contribution of a variable toward achieving equal distributions of the criterion variable (Hastie et al., 2009).

Overall, the mean decrease in the Gini coefficient value was 9.06. The five items with the highest values were 3, 2, 10, 12, and 8. Interestingly, the items with the lowest values (1, 6, 7, 9, and 11) did not include Item 13, which had been flagged as problematic in previous research (Sauer, Ziegler, Danay, Ives, & Kohls, 2013).

In order to assess the relative impact of the two samples, we ran a logistic regression with all of the items and the samples as independent variables, again to predict class membership. Interestingly, the *B* coefficient for the sample was weak (B = .33; SE = .33; z = 1.02; p = .31).

Discussion

In this paper we tested the predictive quality of the FMI in an attempt to gauge the validity of the scale. This novel approach was applied to contribute to the measurement of mindfulness. In sum, our results indicate that the FMI is able to distinguish mindfulness practitioners from nonpractitioners. This result can be taken as support for the FMI's criterion validity. However, substantial room for improvement exists. The contribution of the items was not homogeneous, and the importance of many items was weak.

The analytical method employed in this research relied on three aspects. First, we used well-acknowledged predictive algorithms novel to psychometrics. Second, we used a sophisticated cross-validation approach (repeated k-fold cross-validation). Third, our results did not rely on only one algorithm but rather on the average of several (i.e., 10) algorithms, and thus can probably be regarded as stable in this sense.

Interestingly, predictive algorithms differ in the accuracy with which they are able to separate mindfulness practitioners from nonpractitioners. Due to the particularities of the analytical problem and its data, some algorithms may obviously be severely misguided. In our case, *ada*, generally known as a high-performing algorithm, performed poorly. Whereas we can only speculate about the statistical reasons behind the performance problems, this result supports the hypothesis that aggregating algorithms is generally superior to relying on one standard algorithm alone, at least if the algorithm has no prior experience with the specific situation. On the other hand, *rfs* performed extremely well, and the "classic psychological" algorithm, the (logistic) regression (*glm*), yielded a somewhat intermediate performance. This is particularly interesting as psychological research tends to rely on this algorithm (i.e., algorithmic pluralism appears superior to unquestioned algorithmic autocracy). To date, we can only speculate as to why *rfs* performed so well, whereas *ada* performed poorly. Data particularities may be the cause. More research is clearly needed. We recommend that future research similarly compute several models and compare the results.

Our results speak clearly with regard to the predictive quality of the items: Substantial room for improvement exists. Only a few items substantially contributed toward separating the two groups (mindfulness practitioners vs. nonpractitioners). This result substantiates a similar notion regarding items from a different mindfulness questionnaire (Hoefling, Moosbrugger, Schermelleh-Engel, Heidenreich, & Höfling, 2011; Van Dam, Earleywine, & Borders, 2010). Item 13, which was flagged as problematic in previous research (Sauer, Ziegler, Danay, Ives, & Kohls, 2012) was not amongst the most problematic items in the current study. However, given the argument that negatively coded items do not appear to simply convey the "same but reversed" effect, we propose that Item 13 still be treated as a revision candidate (Hoefling et al., 2011).

Some words of caution need to be presented. First, one limitation of this study is that the quality of the online sample was probably moderate at best. Limited sample quality may be likely to attenuate existing associations and the power of statistical tests. In addition, the allocation to groups in the online sample relied on self-report and was not controlled. Third, the predictive criterion (mindfulness practice or not) is only one of many plausible criteria. We would like to emphasize that this criterion is probably unable to produce a pure splitting of the groups by their mindfulness levels. The reason is that mindfulness level is less than perfectly correlated with (amount of mindfulness) practice.

The results and the limitations of the present research provide some ideas for new avenues of research. We urge researchers to make use of the predictive models that are widely used in other high-quality research branches such as gene expression or molecular biology in general. These results are not only of importance for the research on mindfulness but may help researchers in different psychological areas as well. It should be mentioned that the *R* package *caret* provides a unified syntax for all of its approximately 160 models, thus providing comfortable and low-barrier access to researchers who do not have a background in machine learning (Kuhn, 2008). In a similar vein, we propose that different criteria be used as the target to be predicted by actual machine learning algorithms. Although mindfulness practice is perhaps one of the most straightforward criteria to choose from, other criteria such as drug intake or health incidents may provide empirically founded criteria (Leigh et al., 2005). In addition, it seems strongly warranted that future research strive to improve the instrument's items. Many instruments for measuring mindfulness exist; their psychometric property may also be investigated using machine learning procedures. This would help us better understand the benefits and potential problems of employing this method in psychometric research.

For the time being, we conclude that a multimodel predictive approach is a high-quality approach that can be applied to test certain aspects of the validity of measurement instruments in psychology and in mindfulness research in particular. Given the widespread interest in and the promising preventative and curative health benefits of mindfulness, an improved understanding of how it is measured is of high importance.

Electronic Supplementary Material.

No	*File name*	*Description*
1	ESM1_predictive_models.docx	Description of the statistical rationale of the predictive models employed in the present study
2	ESM2_Table_B1.docx	Table reporting detailed descriptive statistics
3	ESM3_Figures.docx	Additional figures (C1 and C2)
4	ESM4_syntax.R	R Syntax
5	ESM5_data.csv	Data (csv format)

References

Baer, R. A. (2011). Measuring mindfulness. *Contemporary Buddhism*, *12*(01), 241–261.

Bishop, S. R., Lau, M., Shapiro, S., Carlson, L., Anderson, N. D., Carmody, J., … Speca, M. (2004). Mindfulness: A proposed operational definition. *Clinical Psychology Science and Practice*, *11*(3), 230–241.

Bohlmeijer, E., Ten Klooster, P. M., Fledderus, M., Veehof, M., & Baer, R. (2011). Psychometric properties of the five facet mindfulness questionnaire in depressed adults and development of a short form. *Assessment*, *18*, 308–320. doi:10.1177/1073191111408231

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. doi:10.1037/0022-3514.84.4.822

Chiesa, A., Calati, R., & Serretti, A. (2011). Does mindfulness training improve cognitive abilities? A systematic review of neuropsychological findings. *Clinical Psychology Review*, *31*(3), 449–464. doi:10.1016/j.cpr.2010.11.003

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, *97*, 77–87. doi:10.1198/016214502753479248

Eberth, J., & Sedlmeier, P. (2012). The Effects of Mindfulness Meditation: A Meta-Analysis. *Mindfulness*, *3*(3), 174–189. doi:10.1007/s12671-012-0101-x

Eisendrath, S. J., Delucchi, K., Bitner, R., Fenimore, P., Smit, M., & McLane, M. (2008). Mindfulness-Based Cognitive Therapy for Treatment-Resistant Depression: A Pilot Study. *Psychotherapy and Psychosomatics*, *77*(5), 319–320.

Farb, N. A. S., Anderson, A. K., & Segal, Z. V. (2012). The mindful brain and emotion regulation in mood disorders. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie*, *57*(2), 70–7.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*.

Gard, T., Hölzel, B. K., & Lazar, S. W. (2014). The potential effects of meditation on age-related cognitive decline: A systematic review. *Annals of the New York Academy of Sciences*, *1307*, 89–103. doi:10.1111/nyas.12348

Grossman, P. (2008). On measuring mindfulness in psychosomatic and psychological research. *Journal of Psychosomatic Research*, *64*(4), 405–408.

Grossman, P. (2011). Defining Mindfulness by How Poorly I Think I Pay Attention During Everyday Awareness and Other Intractable Problems for Psychology's (Re)Invention of Mindfulness: Comment on Brown et al. (2011). *Psychological Assessment*, *23*(4), 1034–1040.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. *Elements* (Vol. 1). New York City: Springer. doi:10.1007/b94608

Hoefling, V., Moosbrugger, H., Schermelleh-Engel, K., Heidenreich, T., & Höfling, V. (2011). Mindfulness or Mindlessness? *European Journal of Psychological Assessment*, *27*(1), 59–64. doi:10.1027/1015-5759/a000045

Hofmann, S. G., Sawyer, A. T., Witt, A. A., & Oh, D. (2010). The Effect of Mindfulness-Based Therapy on Anxiety and Depression: A Meta-Analytic Review. *Journal of Consulting and Clinical Psychology*, *78*(2), 169–183.

Kohls, N., Sauer, S., & Walach, H. (2009). Facets of mindfulness--Results of an online study investigating the Freiburg mindfulness inventory. *Personality and Individual Differences*, *46*(2), 224–230.

Kuhn, M. (2008). caret Package. *Journal Of Statistical Software*, *28*, 1–26.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, *33*, 363–374. doi:10.2307/2529786

Langer, E. J., & Moldoveanu, M. (2000). The Construct of Mindfulness. *Journal of Social Issues*, *56*(1), 1–9.

Leigh, J., Bowen, S., & Marlatt, G. A. (2005). Spirituality, mindfulness and substance abuse. *Addictive Behaviors*, *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. *International Journal of Osteopathic Medicine*, *13*(2), 56–66. doi:10.1016/j.ijosm.2009.07.005

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

Sauer, S., Walach, H., & Kohls, N. (2010). Gray's Behavioural Inhibition System as a mediator of mindfulness towards well-being. *Personality and Individual Differences*, *50*(4), 506–551.

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

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.

Sauer, S., Ziegler, M., Danay, E., Ives, J., & Kohls, N. (2012). Specific Objectivity of Mindfulness—A Rasch Analysis of the Freiburg Mindfulness Inventory. *Mindfulness*, *September*, 1–10. doi:10.1007/s12671-012-0145-y

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.

Sedlmeier, P., Eberth, J., Schwarz, M., Zimmermann, D., Haarig, F., Jaeger, S., & Kunze, S. (2012). The psychological effects of meditation: a meta-analysis. *Psychological Bulletin*, *138*(6), 1139–71. doi:10.1037/a0028168

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C. T., … Golub, T. R. (2002). *Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature medicine* (Vol. 8, pp. 68–74). doi:10.1038/nm0102-68

Tang, Y.-Y., & Posner, M. I. (2013). Tools of the trade: theory and method in mindfulness neuroscience. *Social Cognitive and Affective Neuroscience*, *8*(1), 118–20. doi:10.1093/scan/nss112

Teper, R., & Inzlicht, M. (2013). Meditation, mindfulness and executive control: the importance of emotional acceptance and brain-based performance monitoring. *Social Cognitive and Affective Neuroscience*, *8*(1), 85–92. doi:10.1093/scan/nss045

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

Ye, Q.-H., Qin, L.-X., Forgues, M., He, P., Kim, J. W., Peng, A. C., … Wang, X. W. (2003). Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. *Nature Medicine*, *9*, 416–423. doi:10.1038/nm843

Zainal, N. Z., Booth, S., & Huppert, F. A. (2013). The efficacy of mindfulness-based stress reduction on mental health of breast cancer patients: a meta-analysis. *Psycho-Oncology*, *22*(7), 1457–65. doi:10.1002/pon.3171

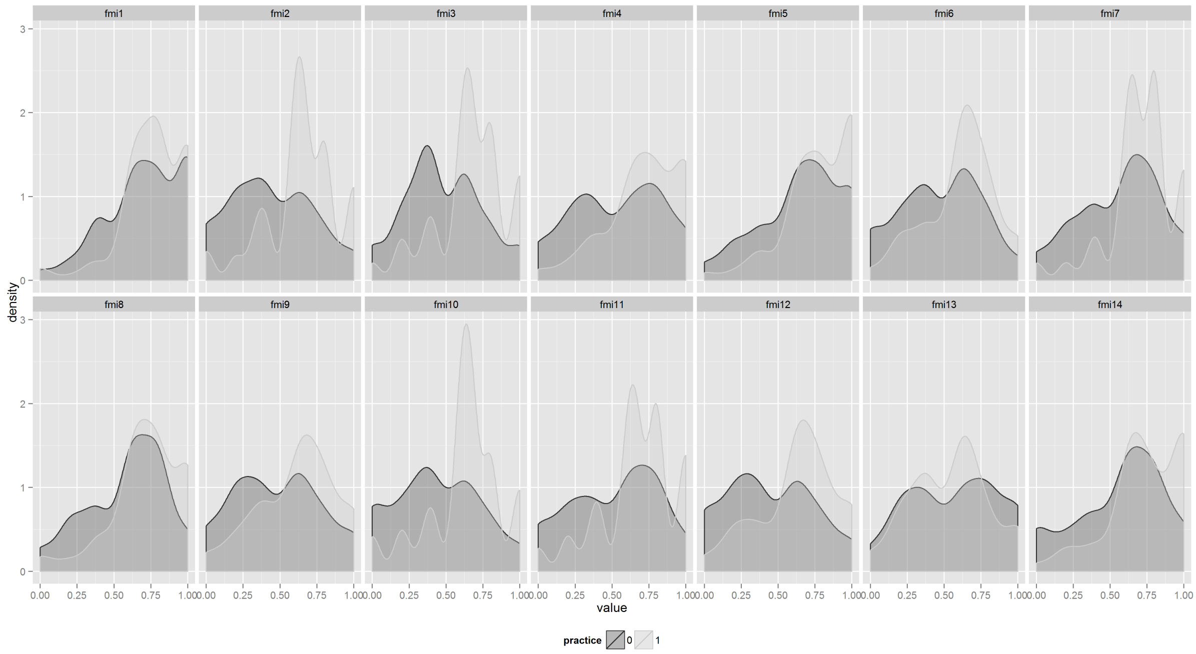

Figure 1. Density diagrams of mindfulness levels for each item. The figure shows the probability densities for each item comparing individuals with and without mindfulness practice.

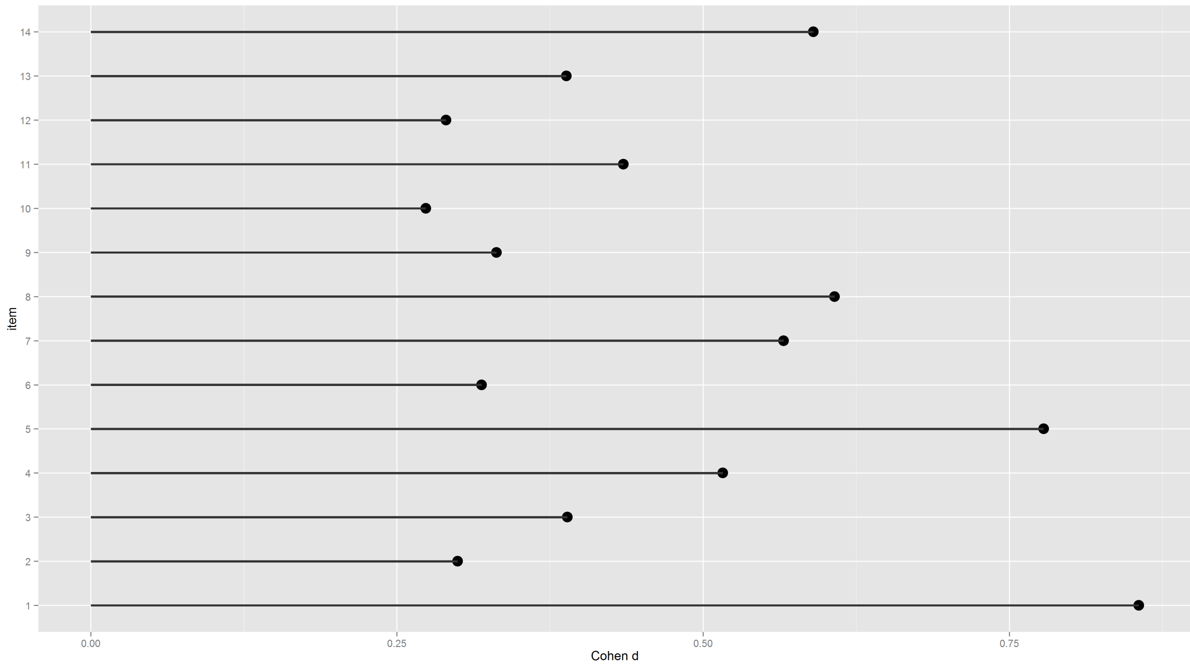

Figure 2. Cohen's d for each item. This figures shows the effect size difference (Cohen's d) between mindfulness practitioners and nonpractitioners.

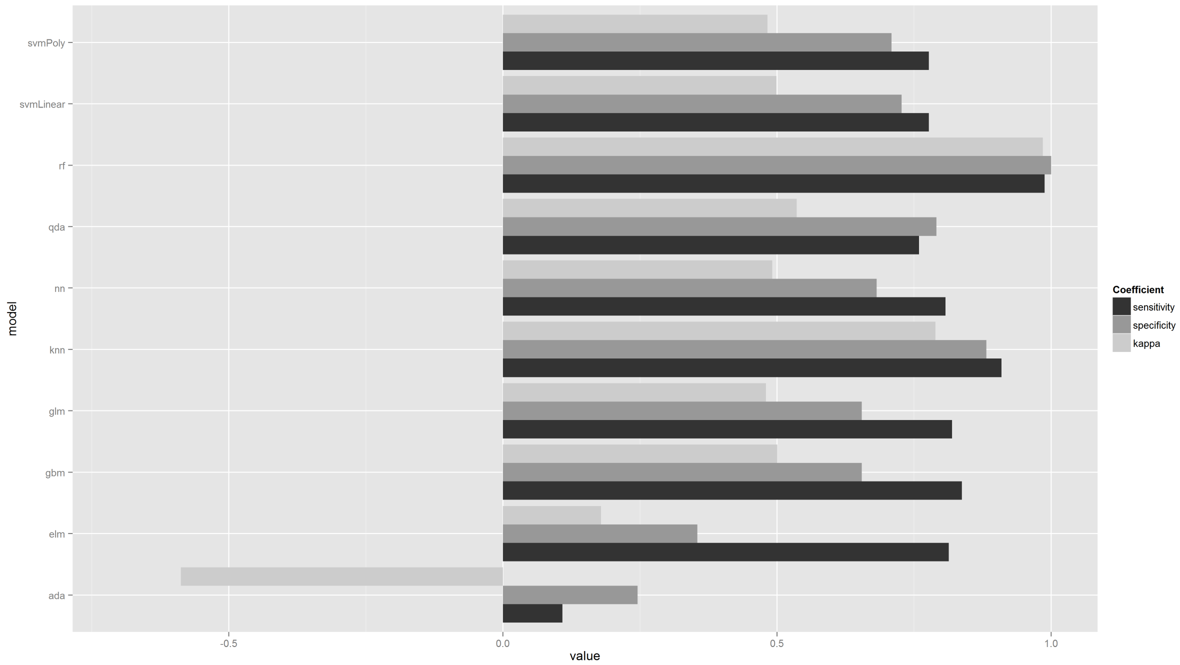

Figure 3. Results of predictive models. This figure shows sensitivity, specificity, and the kappa coefficient for each of the 10 predictive models.