**This paper is a pre-print. The updated peer-reviewed study is published in *Assessment*,
available at: <https://journals.sagepub.com/doi/full/10.1177/1073191119873718>**

**A replication study of the Five Facets of Mindfulness Questionnaire (FFMQ) latent
structure**

Oscar Lecuona ^a

Eduardo García-Garzón ^a

Carlos García-Rubio ^a

Raquel Rodríguez-Carvajal ^a

^a Faculty of Psychology, Universidad Autónoma de Madrid

Corresponding author: Oscar Lecuona, Department of Health & Biological Psychology,
Universidad Autónoma de Madrid, Iván Pavlov St., 6, 28049, Madrid.

e-mail: oscar.lecuona@hotmail.com

Note

The authors would like to clarify the lack of bias towards the findings of this study (e.g. cherry-picking data), as the original aim was to develop an IRT model for the FFMQ, but the failure to meet assumptions led to question the latent structure of the questionnaire, and therefore shaping this research publication.

Abstract

The Five Facets of Mindfulness Questionnaire (FFMQ) is a popular tool in mindfulness research. However, some controversies are being upheld regarding its psychometric qualities and its replicability. This study introduces novel evidence regarding two main issues: Firstly, questionable psychometric practices in validation studies confirmed previous concerns about the FFMQ's properties. Secondly, a replication study using frequentist and bayesian techniques tested the structure of the FFMQ. The original structure did not provide a good fit with both techniques, while a proposed alternative provided mixed results. Systematic fit improvements were found in both techniques regarding the exclusion of the Observe facet and the inclusion of method factors. Therefore, a failed replication of the structure of the FFMQ was concluded. Alternatively, a new provisional FFMQ model was proposed. Future research proposals on improving techniques and models towards mindfulness assessment are presented and discussed.

Keywords: mindfulness; FFMQ; confirmatory factor analysis; Bayesian structural equation modeling; replication

A psychometric review and replication study of the Five Facets of Mindfulness Questionnaire
(FFMQ) latent structure

Mindfulness research has experienced a rapid growth in interest and production (AMRA, 2017). As a recent review suggested (Creswell, 2016), mindfulness-based interventions (MBIs) seem to improve several health issues (e.g. depression relapse, addiction, chronic pain, etc.). In addition, increasing social and political attention is drawn to mindfulness to implement MBIs across educational, health, organizational and penitentiary contexts (The Mindfulness All-Party Parliamentary Group, 2015). This scientific landscape needs the essential core of a robust, reliable and viable measurement of mindfulness itself. In that sense, several self-report measures were proposed (for a review, see Bergomi, Tschacher, & Kupper, 2013). Among them, the Five Facets of Mindfulness Questionnaire (FFMQ, Baer et al., 2008; Baer, Smith, Hopkins, Krietemeyer, & Toney, 2006) stands out as one of the most popular ones in usage. This instrument defines mindfulness as an overall psychological construct from five main facets, named as: “Observe”, as the tendency to observe stimuli (e.g. sensations of the wind in the body), “Describe” as the tendency to verbally describe those stimuli, “Acting with Awareness” as the tendency to be presently aware of inner sensations, “Non-Judging of Inner Experience” as the tendency to not judge a particular inner experience as good or bad, and “Non-Reacting to Inner Experience” as the tendency to not immediately react to a particular inner experience and “take a step back” to gain perspective. FFMQ’s wide popularity among scientific community appears to have two main causes: (1) It was created by applying factor analysis techniques to several existing mindfulness measures, which somewhat synthesized the existing instruments into a more comprehensive one; and (2) it received several

international validations (see Table 1 for details), which could be seen as successful replications of the psychometric characteristics of the FFMQ.

The FFMQ was built by grouping pre-existing validated self-report instruments (see Baer et al., 2006 for details) to then collect data and select the best items and latent models that arose. This strategy of relying on purely empirical elements to generate scientific knowledge is supported by exploratory research (i.e. when there are not enough solid theoretical models available). However, the lack of theoretical models for guidance, selection or prediction can lead to generalization and validity issues. This could be the cause of the FFMQ's psychometric issues. Regarding the FFMQ's latent structure, method factors were proposed due to item wording and fit issue (Van Dam, Hobkrik, Danoff-Burg, & Earleywine, 2012), and a bifactor structure alongside method factors was also proposed (Aguado et al., 2015) (Figure 1). This model sets each item to measure two uncorrelated, independent dimensions: An overall mindfulness factor, and the correspondent specific facet to said item (Bonifay, Lane, & Reise, 2016). This boundary struggles to fit in mindfulness theoretical models, and also bifactor models held controversy for overestimating fit in hierarchical structures (Murray & Johnson, 2013). Therefore, a bifactor model for the FFMQ should be also supported by theory, and not be uniquely based on fit indexes. In other trends, item response theory (IRT) models of the FFMQ also found psychometric issues (Bowman, 2014; Medvedev, Tiktova, Siegert, Hwang, & Krägeloh, 2018), where a multidimensional IRT model was built (Van Dam, Bilgrami, & Eisenlohr-Moul, 2018) but not without item removal, response format issues and several modifications in latent structure. Moreover, issues also appeared regarding external and conceptual validity (Goldberg et al., 2015; Feng, Krägeloh, Billington, & Siegert, 2017). The "Observe" facet seems to be especially conflicting in literature: Although this facet is considered as an essential skill of

mindfulness (Lilja, Lundh, Josefsson, & Falkenström, 2013), it showed inconsistent associations with the other facets (for a review, see Rudkin, Medvedev, & Siegert, 2017) and psychological symptoms (Bergomi et al., 2013; Rudkin et al., 2017). Several authors proposed its removal from the questionnaire in non-trained populations (Baer et al., 2008; de Bruin, Topper, Muskens, Bögels, & Kamphuis, 2012). Therefore, it seems that the FFMQ shows psychometric limitations and an in-depth exploration is needed to clarify and map these issues.

(Figure 1 goes here)

(Table 1 goes here)

Apart from structural ones, another set of issues comes regarding questionable psychometric practices among FFMQ research literature. As it is exposed in Table 1, although the majority implemented Confirmatory Factor Analysis (CFA), all validations showed questionable practices (e.g. using Pearson's correlations instead of polychoric correlations to build the input matrix for the CFA). Moreover, while some validations did not explicit the estimation method (which impoverishes the review), only three of them implemented updated estimation methods (i.e. unweighted least squares or maximum likelihood robust). Nevertheless, these methods are considered acceptable but not optimal to Likert-type of data, being Weighted Least Squares Mean-and-Variance Adjusted (WLSMV, or Robust WLS) the most unbiased available method in this context (Li, 2016). The rest of the validations implemented questionable estimation methods, mainly Maximum Likelihood (ML) for CFAs, Principal Axis (PA) for Exploratory Factor Analysis (EFA) or even unrotated Principal Components.

Regarding fit measures, most of the validations used adequate fit measures according to available literature (CFI, TLI, RMSEA, and SRMR, and also the Chi-

square statistic). Only two studies reported questionable fit measures (i.e. GFI, AGFI, and PGFI). Regarding reliability, all the studies except one (Meda et al., 2015) implemented Cronbach's α , while less biased indices (e.g. McDonald's ω , Guttman's λ_6 or Greatest Lower Bound) are available (for details, see Revelle & Zinbarg, 2009; Sitjima, 2009; Trizano-Hermosilla & Alvarado, 2016). Regarding other practices, only three validations considered raw item data, while six of them applied item parcels which is considered inappropriate due to its tendency to overestimations (Bandalos, 2008). Finally, other questionable practices are present in some studies (e.g. scree plots to decide the number of dimensions instead of parallel analysis or inclusion of modification indices without theoretical specification). Although many of those questionable practices come directly from limitations of software (e.g. AMOS or EQS), one conclusion can be drawn from this landscape: FFMQ's psychometric properties are not free from mixed results and systematic bias due to methodological issues. Thus, its structure and reliability need more research to unravel its psychometric properties with up-to-date techniques and models.

Those needs complement with the more general issue of the replication crisis (Open Science Collaboration, 2015), in which a great deal is explained from questionable research practices and non-transparent policies of journals and researchers (e.g. Fiedler & Schwarz, 2016). To correct these tendencies, replication studies and other initiatives are proposed to ensure reproducible research (Munafò et al., 2017). Therefore, accounting for all the mentioned limitations of the FFMQ, a new validation is proposed here exploring all the models available with updated techniques. The most used psychometric technique to this endeavor is CFA, although it can be too restrictive due to generally not allowing cross-loadings. However, contemporary techniques such as Exploratory Structural Equation Modeling (ESEM) or Bayesian Structural Equation

Modeling (BSEM) aim to use the advantages of CFA but accounting for cross-loadings. Therefore, it allows researchers to correctly estimate models including several cross-loadings (Asparouhov, Muthén, & Morin, 2015), and it has shown to be successful when applied to scales with several cross-loadings and residual covariances (De Bondt & Van Petegem, 2015). Therefore, BSEM stands as a new technique to account and explore latent structures with cross-loadings.

This study aims for a replication of the structure and psychometric properties of the FFMQ using CFA and BSEM. This replication study will be centered in fitting and estimating latent models for the FFMQ, rather than estimating relationships with other constructs or conceptual validity. Reasons for aiming for such scope are (1) model comparison will be comprehensive and numerous, (2) employed techniques need extensive specifications for interpretation, and (3) a large sample size is required to estimate parameters confidently, which needs to ensure success in recruitment, and therefore, a minimal participants' workload (i.e. number of items).

Method

Design

Examining the two predominant latent models of the FFMQ (Figure 1) and literature leads to three distinct independent variables: (1) An overall mindfulness latent variable as a hierarchical (second-order) or as a bifactor one; (2) The inclusion or exclusion of the "Observe" facet, and (3) the exclusion, inclusion or double inclusion of method factors (with inclusion meaning an all-item loading to a single method factor, and double inclusion as two method factors representing positive or negative item wording). These three variables, when crossed, form 12 models (Table 2), with both previously explored models included among them.

(Table 2 goes here)

In order to assess the influence of those variables in the models, differences between models took place accounting for isolating the effect of a single variable. For example, to observe the influence of the Observe facet, models 2 – 1, 4 – 3, 6 – 5, 8 – 7, 10 – 9 and 12 – 11 were performed and a mean and statistical deviation (SD) was obtained. If the mean was larger than the SD, the influence was considered greater than zero. This process was applied to each variable and each model fit index.

Sample

A total of 800 persons participated in the study. Gender was predominantly female (73.75%), young (Mode = 18, SD = 13.35), college educated or above (79.94%, IQV = .779¹) and coupled (45.55%, IQV = .684) closely followed by singles (44.18%). Additionally, a majority of the sample (84.73%) had previous experience with meditation (measured as a minimum contact with meditative practice).

Instruments

The Five Facets of Mindfulness Questionnaire (FFMQ): As mentioned before, this self-report consists of 39 items with a Likert response format from 1 (totally disagree) to 5 (totally agree). It has five facets with a hierarchical overall mindfulness score or a bifactor structure with two method factors (Cebolla et al., 2012).

Procedure

All participants signed an informed consent of participation and were all informed about the privacy, anonymity, and confidentiality of the treatment of their data, following the Helsinki protocol. Then, a copy of the FFMQ was provided and participants responded to it. They were introduced to the rules of responding of the

questionnaire (Likert response format, the meaning of each value of the scale, they were not right or wrong answers but personal preferences or attitudes, not to overthink the response, and don't leave any item unanswered). They also were told that the researcher in charge of that collection could answer any question or issue they had, and they can withdraw the questionnaire at any time. Then, questionnaires were delivered to the researcher, digitalized and stored in a database (available at osf.io/m64wp/). Participants were given no retribution for their participation.

Data Analysis

Confirmatory Factor Analysis (CFA) was implemented as the first approach. As the FFMQ's Likert response format, all variables were considered ordinal, and therefore, polychoric correlations were selected with WLSMV estimation (Brown, 2006; Flora & Curran, 2004; Proitsi, 2009), which are designed for categorical data. In addition, Satorra-Bentler chi-square correction was selected (Bryant & Satorra, 2012; Satorra & Bentler, 2001). Chosen fit measures were the χ^2 statistic, the chi-square ratio (χ^2/df), the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), the Root Mean Square Error Approximation (RMSEA), and the Weighted Root Mean Square Residual (WRMR). For the CFI and TLI indices, cutoff criteria for good fit is with values $\geq .95$, while for the RMSEA is $\leq .05$, and for the WRMR is ≈ 1 . In addition, modification indices (MIs) were examined.

Estimation in BSEM consists of a Markov Chain Monte Carlo (MCMC) method using the Gibbs sampler (Merkle & Rosseel, 2016). A default of 5,000 (burn-in, adaptive and sampler) iterations were set for the models, although were increased when convergence issues appeared to a top of 50,000 (due to computation time). Convergence was assessed with the Gelman-Rubin Potential Scale Reduction Factor (PSRF) for each parameter, where values ≤ 1.2 is considered as proper convergence. Model fit was

assessed with the log-likelihood (Log-Lik) and the Posterior Predictive bayesian P-value (PPP), from which a well-fitting model is expected to have a $PPP \geq .05$. However, the PPP seems to behave similarly to the χ^2 p-value with complex models, so zero or close-to-zero values are expected. In addition, model comparison was based on the Bayesian Information Criterion (BIC), the Discrepancy Information Criterion (DIC), the Widely Applicable Information Criterion (WAIC) and the Leave-One-Out Cross-Validation statistic (LOO-CV). In any case, smaller values suggest a better fit. However, LOO-CV is often preferred due to its robustness (Raftery, 1993). We decided to include minimally informative priors in our models: Factor loading's priors were specified as $N(.50, 10)$ truncated -1 to 1. Item intercept priors were specified following the Likert structure as $N(3, .10)$ truncated from 1 to 5. Also, item variance priors were specified as $\log-N(1, .10)$. Factor intercept priors were specified as $N(0, 1)$ and factor SD priors as $\text{gamma}(1, .10)$. Regarding cross-loadings, two sets of BSEMs were introduced. First, they were constrained to zero. Second, small-variance priors were specified as $N(0, .10)$ truncated from -1 to 1, thus allowing to take values near zero. These selected priors were previously applied in other empirical contributions (Merkle & Wang, 2018).

In order to assess scale reliability, as mentioned before, McDonald's ω , Guttman's λ_6 and Greatest Lower Bound were implemented instead of Cronbach's α . In addition, ω 's and λ_6 's if item dropped were examined.

All models were computed with the R statistical software, with packages *lavaan* (Rosseel, 2012) for CFAs, and *blavaan* (Merkle & Rosseel, 2016) for BSEMs. All other analyses were computed with JASP (<http://jasp-stats.org/>).

Results

Confirmatory Factor Analysis

Most models provided unsatisfactory fit measures (Table 2), including the original hierarchical model (Baer et al., 2008). Moreover, models 5, 7 and 8 showed convergence issues. Assessing models with robust fit measures leads to select models 10 (hierarchical structure with both method factors and without the Observed facet), 11 (Aguado et al., 2015) and 12 (bifactor structure with both method factors and without the Observe facet). Models 10 and 11 show similar borderline-good fit measures, while model 12 shows good fit measures. Therefore, models proposed by literature provided unsatisfactory fit.

(Table 3 goes here)

In addition, the variables showed that the distinction between hierarchical and bifactor models did not provide considerable improvements of fit. However, eliminating the Observe facet and including method factors seem to improve considerably model fit in all measures, only improving RMSEA and χ^2 ratio comparing one to two method factors. As a more detailed analysis is needed, unexpected loadings were examined. Overall, they had a low prevalence (around 2.54%), but second-order factors (i.e. overall mindfulness and method factors) had a higher prevalence (around 6.46%) than facets (around .87%). Thus, facets seem to display a more robust latent structure than overall mindfulness or method factors. Lastly, modification indices were explored for each model. In all models, the maximum scaled value of the modification indices was lower than the 10% of the χ^2 statistic, making quite unlikely to produce a considerable improvement.

Bayesian Structure Equation Modelling

Coherently with CFA, some models showed issues of convergence. PRSF diagnostics were found in models 8, 9, 10 and 12 although showing a low prevalence

(around 8.36%) and all were loading parameters, which discards issues like Heywood cases. In models with liberated cross-loadings, all models showed issues in the same direction (i.e., low prevalence – 10.41% - and occurring in factor loading parameters). Second order factors (i.e. overall mindfulness or method factors) showed frequent convergence issues, while the Actaware and Non-Judging factors were the most frequent among facets. Regarding model fit (Table 4), all models showed bad fit (all PPPs<.5). Assessing fit, models 12, 8 and 10 were selected. This is convergent with CFA except for model 8. This model ranking is consistent across constrained and liberated cross-loadings, although in that last condition model comparison seems clearer.

(Table 4 goes here)

Regarding structure variables, all seem to systematically improve fit (this is, lower levels of log-likelihood, BIC, DIC, WAIC, and LOO-CV). The inclusion of cross-loadings seems to alter lightly the effects, without changing general patterns. All indices tend to behave in the same manner, with little discrepancies between them. The exclusion of the Observe facet seems to heavily improve model fit, while the inclusion of two method factors improves considerably model fit (even better than the inclusion of one method factor), and bifactor structure also seems to improve mildly model fit.

Regarding unexpected factor loadings, a detailed examination also threw issues regarding unexpected loadings. Overall, little proportion of loadings were unexpected ($M=3.86\%$, $SD=2.90\%$), where second order factors report more ($M=8.16\%$, $SD=6.54\%$) than facets ($M=2.84\%$, $SD=3.34\%$). Models allowing cross-loadings reported similar results ($M=10.52\%$, $SD=6.74\%$) although with more accused differences between second order factors ($M=35.17\%$, $SD=23.49\%$) and facets

($M=.85\%$, $SD=1.41\%$). Models 8, 10 and 12 also showed unexpected loadings in both BSEMs.

Scale Reliability

Regarding dimensions' reliability measures, all show insufficient fit (Table 5), which is compatible with CFA and BSEM issues. The highest value across the three measures was the positive bias factor, followed by the Non-Judge facet and the general mindfulness factor, while the lowest value was by far the Non-React facet. No items presented a major reliability improvement if dropped, with only item 19 in the Negative bias factor ($\Delta\omega=.003$, $\Delta\lambda_6=.005$) and item 4 in Non-React facet ($\Delta\lambda_6=.013$).

(Table 5 goes here)

Discussion

This study aimed to replicate the psychometric structure and reliability of the FFMQ. To do so, updated classical techniques (i.e. CFA) and more contemporary techniques (i.e. BSEM) were employed. In that sense, replication came to be with multiple issues and problems regarding model fit. Both techniques (CFA and BSEM) did not provide a good fit for the original model of the FFMQ. Even more, it was the model with lesser fit among the proposed models of the study when robust measures were available. This leads to a preliminary conclusion of the need for rethinking and reconstructing the FFMQ's dimensional structure, along with some inquiry about if it is a viable instrument to assess mindfulness. In that sense, more models were proposed, like the bifactor structure with two method factors. Nevertheless, that model did provide borderline fit with the data in CFA and low comparative fit in BSEM, while two alternative models showed better fit in both techniques: One with a hierarchical

structure, two method factors and omitting the Observe facet (model 10), and another equivalent but with a bifactor structure instead (model 12).

Regarding variable effects, mixed evidence arose: Firstly, the inclusion of method factors showed systematic improvements in fit. More concretely, in CFA leads to increases in relative fit (CFI, TLI, RMSEA and χ^2 ratio), and in BSEM, general fit improvements when constrained cross-loadings but more mixed results when liberated cross-loadings, although LOO-CV shows stable improvements. These findings suggest the need for including method factors in FFMQ revisions. Secondly, bifactor structures compared to hierarchical ones provided uncertain fit improvements in CFA but stable (although mild) improvements in BSEM. Therefore, when cross-loadings are taken into account, the bifactor structure seems to improve fit. However, given literature warnings of overfitting in bifactor structures and difficult theoretical compatibilities with mindfulness models (e.g. Hölzel et al., 2011; Lindsay & Creswell, 2017), it seems difficult to match the existence of said structure. Alternatively, a hierarchical model seems more appealing, including possible covariance between items acknowledging for linguistic bias. Thirdly, the exclusion of the Observe facet provided systematic improvements in both techniques, although CFA only showed improvements in direct fit (i.e. χ^2 values) and the WRMR (thus, with little influence in fit interpretation by researchers). However, BSEMs showed systematic strong improvements across all fit measures (thus, heavily influencing interpretation). These results lead to consider a possible technical bias, with two possible causes: (1) one of them (CFA or BSEM) is more accurate than the other, or (2) both fail to picture the psychometric influence of the Observe facet. Literature and evidence suggest issues and mixed evidence regarding said facet as a part of mindfulness, therefore the first option seems more plausible, favoring BSEM as a more accurate technique. Regarding reliability, results were

convergent with previous CFA and BSEM models. One possible explanation to why the Non-React facet was the less reliable could be of the lack of participant's understanding of its items, as previously suggested in literature (Belzer et al., 2013).

As a preliminary conclusion, the original factor structure of the FFMQ seems insufficient in this sample, and also alternative proposed structures are problematic. Searching for ways to overcome this situation, it seems that excluding the Observe facet and including two method factors systematically improves fit, while changing to a bifactor structure throws mixed evidence. But even more, when models are selected based on those corrections, issues with factor loadings appear, showing several non-significant loadings or negative, significant ones. This raises concerns about the psychometric appropriateness of the FFMQ, expanding contributions questioning its properties (e.g. Aguado et al., 2015; Goldberg et al., 2015). Therefore, a preliminary conclusion is that the structure of the FFMQ does not replicate, even when major corrections take place.

Limitations

Some limitations in the study are needed to be considered. Sample size, although high, had some demographic biases (e.g. female-oriented) which undermine its reproducibility. No multigroup contrasts were performed between meditators and non-meditators to discard differences caused by meditation practice. Also, no criterion validity analysis was performed, thus relationships of FFMQ dimensions with other constructs remain without replication. Lastly, no longitudinal analysis were performed to see if the FFMQ remains sensitive to mindfulness training, which paired with multigroup analysis could explain if the FFMQ has different latent structures dependent on meditation expertise.

Future directions

This study sets the first steps to an updated, more robust exploration of the instrument. Therefore, more research is needed on behalf of finding the reproducibility of the FFMQ's psychometric properties, aiming to produce enough studies to perform a psychometric meta-analysis that provides more robust conclusions.

In addition, new, more robust and reproducible instruments are needed to measure mindfulness. As one main critique of the FFMQ is its lack of a theoretical model due to its inexistence at the time the instrument was developed, future research needs a preliminary phase of synthesizing literature and proposing models in order to construct reliable and valid measures (as suggested by Rau & Williams, 2016). In that direction, some models provide helpful insights and conceptual background necessary to develop instruments. Therefore, future research could aim to build new self-report measures based on conceptual models.

Further, a critique to self-reports could be held. To measure experiential phenomena (e.g. mindfulness) with verbal assessments (e.g. self-reports) could be seen as non-optimal. This proposal could be backed with studies finding a struggle in western participants to understand mindfulness self-report items (Belzer et al., 2013) due to lack of (or primitive state of) cultural backgrounds to properly self-acknowledge mindfulness phenomena (and maybe, any other non-Eastern cultures as well). If future research assumes this critique, then current mindfulness research needs to shift from verbal assessment paradigms to experiential or behavioral ones.

Conclusion

The FFMQ shows psychometric issues regarding replicability and latent structure, with evidence suggesting better fit for models with method factors, and the

exclusion of the Observe facet, across frequentist and bayesian techniques. However, even after applying said corrections, issues regarding factor loadings and convergence of models arise. Future research should focus on (1) explore the replicability of these findings, and (2) select and apply theoretical models towards more robust and stable instruments.

Endnotes

¹IQV measures dispersion in nominal variables, and is interpretable as a proportion (for details, see Mueller, Schuessler, & Costner, 1977).

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. <http://doi.org/10.3389/fpsyg.2015.00404>
- AMRA, A. M. R. A. (2017). Mindfulness Journal Publications by Year, 1980-2016. Retrieved October 10, 2017, from <https://goamra.org/resources/>
- Asparouhov, T., Muthén, B., & Morin, A. J. S. (2015). Bayesian Structural Equation Modeling With Cross-Loadings and Residual Covariances: Comments on Stromeier et al. *Journal of Management*, 41(6), 1561–1577. <http://doi.org/10.1177/0149206315591075>
- 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. <http://doi.org/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), 329–42. <http://doi.org/10.1177/1073191107313003>
- Bandalos, D. L. (2008). Is Parceling Really Necessary? A Comparison of Results From Item Parceling and Categorical Variable Methodology. *Structural Equation Modeling: A Multidisciplinary Journal*, 15(2), 211–240. <http://doi.org/10.1080/10705510801922340>
- Barros, V. V., Kozasa, E. H., Souza, I. C. W., & Ronzani, T. M. (2014). Validity

- evidence of the Brazilian version of the Five Facet Mindfulness. *Psicologia: Teoria e Pesquisa*, 30(3), 317–327. <http://doi.org/10.1590/S0102-37722014000300009>
- Belzer, F., Schmidt, S., Lucius-Hoene, G., Schneider, J. F., Orellana-Rios, C. L., & Sauer, S. (2013). Challenging the Construct Validity of Mindfulness Assessment—a Cognitive Interview Study of the Freiburg Mindfulness Inventory. *Mindfulness*, 4(1), 33–44. <http://doi.org/10.1007/s12671-012-0165-7>
- Bergomi, C., Tschacher, W., & Kupper, Z. (2013). The Assessment of Mindfulness with Self-Report Measures : Existing Scales and Open Issues, 191–202. <http://doi.org/10.1007/s12671-012-0110-9>
- Bonifay, W., Lane, S. P., & Reise, S. P. (2016). Three Concerns With Applying a Bifactor Model as a Structure of Psychopathology. *Clinical Psychological Science*, 5(1), 184–186. <http://doi.org/10.1177/2167702616657069>
- Bowman, A. (2014). *Evaluating the Psychometric properties of the Five-Facet Mindfulness Questionnaire: An Item Response Theory Analysis*. Middle Tennessee State University.
- Brown, T. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: Guildford Press.
- Bryant, F. B., & Satorra, A. (2012). *Principles and Practice of Scaled Difference Chi-Square Testing. Structural Equation Modeling: A Multidisciplinary Journal* (Vol. 19). <http://doi.org/10.1080/10705511.2012.687671>
- Cebolla, A., García-Palacios, A., Soler, J., Guillen, V., Baños, R., & Botella, C. (2012). Psychometric properties of the Spanish validation of the Five Facets of

- Mindfulness Questionnaire (FFMQ). *The European Journal of Psychiatry*, 26(2), 118–126. <http://doi.org/10.4321/S0213-61632012000200005>
- Creswell, J. D. (2016). Mindfulness Interventions. *Annual Review of Psychology*, (September), 1–26. <http://doi.org/10.1146/annurev-psych-042716-051139>
- De Bondt, N., & Van Petegem, P. (2015). Psychometric evaluation of the Overexcitability Questionnaire-Two applying Bayesian Structural Equation Modeling (BSEM) and multiple-group BSEM-based alignment with approximate measurement invariance. *Frontiers in Psychology*, 6(1963). <http://doi.org/10.3389/fpsyg.2015.01963>
- de Bruin, E. I., Topper, M., Muskens, J. G. A. M., Bögels, S. M., & Kamphuis, J. H. (2012). Psychometric Properties of the Five Facets Mindfulness Questionnaire (FFMQ) in a Meditating and a Non-meditating Sample. *Assessment*, 19(2), 187–197. <http://doi.org/10.1177/1073191112446654>
- Dundas, I., Vøllestad, J., Binder, P. E., & Sivertsen, B. (2013). The Five Factor Mindfulness Questionnaire in Norway. *Scandinavian Journal of Psychology*, 54(3), 250–260. <http://doi.org/10.1111/sjop.12044>
- Feng, X. J., Krägeloh, C. U., Billington, D. R., & Siegert, R. J. (2017). To What Extent is Mindfulness as Presented in Commonly Used Mindfulness Questionnaires Different from How it is Conceptualized by Senior Ordained Buddhists? *Mindfulness*. <http://doi.org/10.1007/s12671-017-0788-9>
- Fiedler, K., & Schwarz, N. (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, 7(1), 45–52. <http://doi.org/10.1177/1948550615612150>

- Flora, D., & Curran, P. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466–491.
- Giovannini, C., Giromini, L., Bonalume, L., Tagini, A., Lang, M., & Amadei, G. (2014). The Italian five facet mindfulness questionnaire: A contribution to its validity and reliability. *Journal of Psychopathology and Behavioral Assessment*, 36(3), 415–423. <http://doi.org/10.1007/s10862-013-9403-0>
- Goldberg, S. B., Wielgosz, J., Dahl, C., Schuyler, B., Maccoon, D. S., Rosenkranz, M., ... Davidson, R. J. (2015). Does the Five Facet Mindfulness Questionnaire Measure What We Think It Does? Construct Validity Evidence From an Active Controlled Randomized Clinical Trial. *Psychological Assessment, Advance on*. <http://doi.org/http://dx.doi.org/10.1037/pas0000233>
- Heeren, A., Douilliez, C., Peschard, V., Debrauwere, L., & Philippot, P. (2011). Cross-cultural validity of the Five Facets Mindfulness Questionnaire: Adaptation and validation in a French-speaking sample. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, 61(3), 147–151. <http://doi.org/10.1016/j.erap.2011.02.001>
- Hölzel, B. K., Lazar, S. W., Gard, T., Schuman-Olivier, Z., Vago, D. R., & Ott, U. (2011). How Does Mindfulness Meditation Work? Proposing Mechanisms of Action From a Conceptual and Neural Perspective. *Perspectives on Psychological Science*, 6(6), 537–559. <http://doi.org/10.1177/1745691611419671>
- Hou, J., Wong, S. Y.-S., Lo, H. H.-M., Mak, W. W.-S., & Ma, H. S.-W. (2014). Validation of a Chinese Version of the Five Facet Mindfulness Questionnaire in Hong Kong and Development of a Short Form. *Assessment*, 21(3), 363–371.

<http://doi.org/10.1177/1073191113485121>

- Kruschke, J. K. (2011). Doing Bayesian Data Analysis: A Tutorial with R and BUGS. *Europe's Journal of Psychology*, 7(4), 1–187. <http://doi.org/10.5964/ejop.v7i4.163>
- Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, 48(3), 936–949. <http://doi.org/10.3758/s13428-015-0619-7>
- Lilja, J. L., Frodi-Lundgren, A., Hanse, J. J., Josefsson, T., Lundh, L.-G., Sköld, C., ... Broberg, A. G. (2011). Five Facets Mindfulness Questionnaire—Reliability and Factor Structure: A Swedish Version. *Cognitive Behaviour Therapy*, 40(November 2012), 291–303. <http://doi.org/10.1080/16506073.2011.580367>
- Lilja, J. L., Lundh, L. G., Josefsson, T., & Falkenström, F. (2013). Observing as an Essential Facet of Mindfulness: A Comparison of FFMQ Patterns in Meditating and Non-Meditating Individuals. *Mindfulness*, 4(3), 203–212. <http://doi.org/10.1007/s12671-012-0111-8>
- Lindsay, E. K., & Creswell, J. D. (2017). Mechanisms of mindfulness training: Monitor and Acceptance Theory (MAT). *Clinical Psychology Review*, 51, 48–59. <http://doi.org/10.1016/j.cpr.2016.10.011>
- Meda, R. M., Herrero, M., Blanco-Donoso, L. M., Moreno-Jiménez, B., & Chávez, A. P. (2015). Propiedades psicométricas del “cuestionario de cinco facetas de la conciencia plena” (five facet mindfulness questionnaire, FFMQ-M) en México. *Behavioral Psychology/ Psicología Conductual*, 23(3), 467–487.
- Medvedev, O. N., Tiktova, E. A., Siegert, R. J., Hwang, Y.-S., & Krägeloh, C. U. (2018). Evaluating Short Versions of the Five Facet Mindfulness Questionnaire

- Using Rasch Analysis. *Mindfulness*. <http://doi.org/10.1007/s12671-017-0881-0>
- Merkle, E. C., & Rosseel, Y. (2016). blavaan: Bayesian structural equation models via parameter expansion. *ArXiv 1511.0560*. Retrieved from <https://arxiv.org/abs/1511.05604>
- Merkle, E. C., & Wang, T. (2018). Bayesian latent variable models for the analysis of experimental psychology data. *Psychonomic Bulletin & Review*, 25(1), 256–270.
- Michalak, J., Zarbock, G., Drews, M., Otto, D., Mertens, D., Ströhle, G., ... Heidenreich, T. (2016). Erfassung von Achtsamkeit mit der deutschen Version des Five Facet Mindfulness Questionnaires (FFMQ-D). *Zeitschrift Fur Gesundheitspsychologie*, 24(1), 1–12. <http://doi.org/10.1026/0943-8149/a000149>
- Morey, R. D., Romeijn, J. W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72, 6–18. <http://doi.org/10.1016/j.jmp.2015.11.001>
- Mueller, J. H., Schuessler, K. F., & Costner, H. L. (1977). *Statistical reasoning in sociology*. Boston, MA: Houghton Mifflin.
- Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., ... Ioannidis, P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1, 0021.
- Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bifactor versus higher-order models of human cognitive ability structure. *Intelligence*, 41, 407–422. <http://doi.org/10.1016/j.intell.2013.06.004>
- Open Science Collaboration, O. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

- Proitsi, P. (2009). A Multiple Indicators Multiple Causes (MIMIC) model of Behavioural and Psychological Symptoms in Dementia (BPSD). *Neurobiology of Aging*. <http://doi.org/10.1016/j.neurobiolaging.2009.03.005>
- Raftery, A. E. (1993). Bayesian model selection in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 163–180). Beverly Hills, CA: Sage.
- Rau, H. K., & Williams, P. G. (2016). Dispositional mindfulness: A critical review of construct validation research. *Personality and Individual Differences*, 93, 32–43. <http://doi.org/10.1016/j.paid.2015.09.035>
- Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega and the glb: comments on Sijtsma. *Psychometrika*, 74(1), 145.
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1–36. Retrieved from <http://www.jstatsoft.org/v48/i02/>.
- Rudkin, E., Medvedev, O. N., & Siegert, R. J. (2017). The Five-Facet Mindfulness Questionnaire: Why the Observing Subscale Does Not Predict Psychological Symptoms. *Mindfulness*. <http://doi.org/10.1007/s12671-017-0766-2>
- Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66(4), 507–514.
- Sitjsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74, 107–120. <http://doi.org/10.1007/s11336-008-9101-0>
- Sugiura, Y., Sato, A., Ito, Y., & Murakami, H. (2012). Development and Validation of the Japanese Version of the Five Facet Mindfulness Questionnaire. *Mindfulness*,

3(2), 85–94. <http://doi.org/10.1007/s12671-011-0082-1>

The Mindfulness All-Party Parliamentary Group. (2015). *Mindful Nation UK*. London (UK). Retrieved from http://themindfulnessinitiative.org.uk/images/reports/Mindfulness-APPG-Report_Mindful-Nation-UK_Oct2015.pdf

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology*, 7(MAY), 1–8. <http://doi.org/10.3389/fpsyg.2016.00769>

Van Dam, N. T., Bilgrami, Z. R., & Eisenlohr-Moul, T. (2018). Development and Validation of the Balanced Inventory of Mindfulness-Related Skills (BIMS). Retrieved from mindrxiv.org/8qdzg/

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. <http://doi.org/10.1177/1073191112438743>

Veehof, M. M., Ten Klooster, P. M., Taal, E., Westerhof, G. J., & Bohlmeijer, E. T. (2011). Psychometric properties of the Dutch Five Facet Mindfulness Questionnaire (FFMQ) in patients with fibromyalgia. *Clinical Rheumatology*, 30(8), 1045–1054. <http://doi.org/10.1007/s10067-011-1690-9>

Wagenmakers, E., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). *Journal of Personality and Social Psychology*, 100(3), 426–32. <http://doi.org/10.1037/a0022790>

Watson-Stingleton, N. N., Walker, J. H., LoParo, D., Mack, S. A., & Kaslow, N. J.

(2017). Psychometric Evaluation of the Five Facet Mindfulness Questionnaire in a Clinical Sample of African Americans. *Mindfulness*.

<http://doi.org/10.1007/s12671-017-0776-0>

Figure 1: Current latent models for the FFMQ

Table 1: Psychometric techniques and specifications of the FFMQ's published validations

Publication	Population	Model	Corrs	Estimation	χ^2 corr	Fit indices	Reliability	Parcels
Baer et al., 2006	USA	EFA + CFA	Pearson	PA, ML	No	χ^2 , CFI, TLI, RMSEA	α	Yes
Veehof, Ten Kster, Taal, Westerhof, & Bohlmeijer, 2011	Dutch	CFA	Pearson	ML	No	χ^2 , TLI, CFI, SRMR, RMSEA	α	-
Lilja et al., 2011	Swedish	CFA	Pearson	-	No	χ^2 , CFI, RMSEA	α	Yes
Heeren, Douilliez, Peschard, Debrauwere, & Philippot, 2011	French	CFA	Pearson	ULS	No	GFI, AGFI, PGFI	α	No
Cebolla et al., 2012	Spanish	CFA	Pearson	ML	No	χ^2 , CFI, GFI, AGFI, SRMR, RMSEA	α	Yes
Sugiura, Sato, Ito, & Murakami, 2012	Japanese	EFA + CFA	Pearson	PA, -	No	CFI, TLI, RMSEA, AIC	α	Yes
Dundas, Vøllestad, Binder, & Sivertsen, 2013	Norwegian	CFA	Pearson	-	No	χ^2 , χ^2/df , CFI, RMSEA	α	-

Giovannini et al., 2014	Italian	CFA	Pearson	ML	No	χ^2/df , CFI, TLI, RMSEA	α	-
de Barros, Kozasa, Souza, & Ronzani, 2014	Brazilian	EFA	Pearson	PCA	No	-	α	-
Hou, Wong, Lo, Mak, & Ma, 2014	Chinese	CFA	Pearson	ML	No	CFI, TLI, RMSEA, SRMR	α	Yes
Meda, Herrero, Blanco-Donoso, Moreno-Jiménez, & Chávez, 2015	Mexican	CFA	Pearson	MLR	Yes	χ^2 , χ^2/df , CFI, TLI, SRMR, RMSEA	ω	No
Michalak et al., 2016	German	CFA	Pearson	-	Yes	CFI, SRMR, RMSEA	α	No
Watson-Stingleton, Walker, LoParo, Mack, & Kaslow, 2017	African-american	EFA + CFA	Pearson	PA, MLR	No	χ^2 , CFI, RMSEA, SRMR	α	No
de Bruin, Topper, Muskens, Bögels, & Kamphuis, 2012	Meditators/lays	CFA	Pearson	?	No	χ^2 , CFI, TLI, RMSEA	α	Yes

Cells with a dash indicate a not available data in the article. Corrs = type of correlations employed. χ^2 corr = Chi-square correction. EFA =

Exploratory Factor Analysis; CFA = Confirmatory Factor Analysis; PA = Principal Axis; ML = Maximum Likelihood; ULS = Unweighted Least

Squares; MLR = Maximum Likelihood-Robust; PCA = Principal Components Analysis; χ^2 = Chi-square statistic; χ^2/df = Chi-square under degrees of freedom ratio; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error Approximation; SRMR = Standardized Root Mean Residual; GFI = Goodness of Fit Index; AGFI = Adjusted Goodness of Fit Index; PGFI = p-value of the GFI; α = Cronbah's alpha; ω = Mcdonald's Omega.

Table 2: Replication study models resulting from crossing three variables.

	Hierarchical (second-order)		Bifactor	
	“Observe”	“Observe”	“Observe”	“Observe”
	included	excluded	included	excluded
No method factors	Model 1 ^a	Model 2	Model 3	Model 4
One overall method factor	Model 5	Model 6	Model 7	Model 8
Two method factors	Model 9	Model 10	Model 11 ^b	Model 12

^aProposed by Baer et al. (2008). ^bProposed by Aguado et al. (2015).

Table 3: CFA Fit measures for the models proposed in the study.

<i>Model</i>	$\chi^2{}^a$	<i>df</i>	χ^2/df	<i>CFI</i>	<i>TLI</i>	<i>RMSEA</i>	<i>WRMR</i>
1	3886.05	697	5.75	.879	.871	.076	2.581
2	2635.77	430	6.13	.913	.905	.080	2.401
3	3406.05	663	5.13	.896	.884	.072	2.225
4	1964.57	403	4.87	.938	.929	.070	1.854
5 ^b	2983.35	658	4.53	-	-	-	1.824
6	1743.31	399	4.36	.947	.938	.065	1.733
7 ^b	4439.93	624	7.11	-	-	-	2.225
8 ^b	2024.07	372	5.44	-	-	-	1.854
9	2479.31	658	3.76	.931	.922	.059	1.804
10	1694.08	399	4.24	.949	.940	.064	1.769
11	1947.04	624	3.12	.950	.940	.052	1.461
12	1196.70	372	3.21	.967	.959	.053	1.325
<i>Comparison</i>	$\Delta\chi^2(**)$	Δdf	$\Delta\chi^2/df$	ΔCFI	ΔTLI	$\Delta RMSEA$	$\Delta WRMR$
Observe in vs out	-1348.23	-258	-.19	.018	.018	-.001	-.205
SD	606.17	6	.78	.024	.027	.006	.148
Hierarchical vs							
Bifactor	-73.92	-31	.02	.016	.014	.000	-.195
SD	822.10	4	1.50	.017	.020	.013	.363
0 vs 1 method factors	-175.45	-35	-.11	.057	.060	-.001	-.360
SD	923.88	5	1.71	.027	.029	.014	.400
1 vs 2 method factors	-968.38	0	-1.78	-.015	-.017	-.017	-.323
SD	1065.28	0	1.72	.019	.022	.015	.385
0 vs 2 method factors	-1143.83	-35	-1.89	.043	.043	-.018	-.683

SD	341.89	5	.16	.012	.012	.002	.125
----	--------	---	-----	------	------	------	------

^aAll χ^2 statistics had a p-value <.01 except “1 vs 2 method factors” due to zero degrees of freedom. ^bModels with convergence issues, thus CFI, TLI, and RMSEA were not computed. Shades follow a ranking in each column with darker greys as better fit measures.

Table 4: BSEM Fit measures for models proposed in the study.

	<i>Models without cross-loadings</i>						<i>Models with cross-loadings</i>					
<i>Model</i>	<i>-(Log-Lik)</i>	<i>PPP</i>	<i>BIC</i>	<i>DIC</i>	<i>WAIC</i>	<i>LOO-CV</i>	<i>-(Log-Lik)</i>	<i>PPP</i>	<i>BIC</i>	<i>DIC</i>	<i>WAIC</i>	<i>LOO-CV</i>
1 ^b	41531	0	83877	83306	83319	83319	41744	0	85347	80476	82295	82295
2 ^b	32390	0	65430	64975	64990	64990	32289	0	65849	64306	65056	64916
3 ^b	41165	0	83374	82644	82686	82687	41336	0	84757	80298	81843	81844
4 ^b	32066	0	64962	64381	64404	64404	32275	0	66002	63099	64071	64071
5 ^{a,b}	41415	0	83907	81654	82436	82436	40816	0	83751	81337	81842	81843
6 ^{a,b}	32026	0	64909	64308	64341	64341	31887	0	65244	63876	64070	64070
7 ^{a,b}	41187	0	83679	81118	81979	81979	41088	0	84516	79898	81418	81418
8 ^{a,b}	32040	0	65117	63278	63884	63884	31944	0	65539	62846	63641	63641
9 ^{a,b}	41112	0	83300	81414	82051	82046	41025	0	84422	80020	81416	81417
10 ^{a,b}	32107	0	65069	63367	63952	63951	32012	0	65709	62706	63639	63639
11 ^b	40637	0	82579	81665	81702	81702	-41142	0	84892	79477	81314	81315
12 ^{a,b}	31829	0	64696	63416	63812	63800	32161	0	66181	62082	63523	63524

<i>Comparison</i>	Δ -(Log-Lik)	Δ PPP	Δ BIC	Δ DIC	Δ WAIC	ΔLOO-CV	Δ -(Log-Lik)	Δ PPP	Δ BIC	Δ DIC	Δ WAIC	ΔLOO-CV
Observe in vs out	9014	0	18271	18099	18091	18094	9049	0	18789	17239	17778	17779
SD	190	0	368	372	157	154	190	0	346	477	219	163
Hierarchical vs Bifactor	275	0	347	420	436	437	28	0	260	836	418	394
SD	165	0	317	478	177	173	234	0	464	469	320	272
0 vs 1 method factors	120	0	7	1236	689	690	477	0	726	-55	573	538
SD	171	0	360	446	151	150	307	0	599	615	275	205
1 vs 2 method factors	246	0	492	123	280	285	151	0	538	917	269	269
SD	260	0	520	632	148	144	77	0	141	405	183	183
0 vs 2 method factors	366	0	499	1360	970	975	326	0	187	973	843	807
SD	132	0	236	464	280	277	270	0	511	477	415	351

^aModels without cross-loadings that did not converge after 100,000 iterations. ^bModels with cross-loadings that did not converge after 100,000

iterations. Δ = difference between two models (2nd model – 1st model) in a given fit measure. LOO-CV is bolded to highlight its preference

towards interpretation. Shades follow a ranking in each column with darker greys as better fit measures.

Table 5: Reliability indices for factors and facets of the FFMQ.

<i>Model</i>	<i>McDonald's ω</i>	<i>Gutmann's λ_6</i>	<i>Greatest lower bound</i>
Mindfulness	.896	.933	.902
Bias ^a	-	-	-
Positive	.899	.931	.947
Negative	.814	.847	.894
Observe	.779	.769	.817
Describe	.880	.882	.897
Actaware	.889	.896	.927
Non-Judge	.907	.902	.904
Non-React	.698	.682	.714

^aNot reported due being mathematically equivalent to the general mindfulness factor (all items).