Personality and Individual Differences 120 (2018) 288–298

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Facet Benchmarking: Advanced application of a new instrument refinement method

Alex B. Siegling *, Adrian Furnham, K.V. Petrides

Division of Psychology and Language Sciences, University College London, London, UK

ARTICLE INFO

Article history:
Received 27 August 2016
Received in revised form 2 November 2016
Accepted 9 December 2016
Available online 14 December 2016

Keywords:
Facet benchmarking
Scale construction and revision
Psychological measurement
Psychometrics
Dispositional mindfulness

ABSTRACT

This article presents an advanced application of Facet Benchmarking (FB), an instrument refinement method that sets out to identify *redundant* and *extraneous* facets (Siegling, Petrides, & Martskvishvili, 2015). FB uses external benchmarks to determine whether a measure's facets each occupy unique construct variance. In Study 1, three samples completed measures of dispositional mindfulness and an objectively derived set of construct-relevant criteria. A general factor extracted from these criteria was used to benchmark the measures' facets or subscales. Structural Equation Modelling, featuring a common latent (method) factor, was incorporated as an alternative statistical procedure, indicating that statistical or methodological artefacts were unlikely to account for the obtained results. Study 2 was conducted to cross-validate the results for a benchmark derived from a different set of criteria. The results support the method's robustness and efficacy.

1. Introduction

A core challenge with psychological constructs is to accurately determine their domain of measurable manifestations, or construct domain. This process is often facilitated by the explication of facets, especially where broader constructs are concerned. Determining construct domains involves considerable uncertainty (Costa & McCrae, 1998; Ziegler, Booth, & Bensch, 2013), since an individual and objective criterion against which measures can be evaluated does not normally exist (Epstein, 1984; John & Benet-Martinez, 2000). Various psychometric paradigms (e.g., construct validity vs. test validity theory; Borsboom, Mellenbergh, & van Heerden, 2004) and statistical procedures (e.g., Exploratory Structural Equation Modelling, Bifactor Modelling) have enriched psychometrics, but the process of operationalising constructs remains far from clear-cut (Ziegler & Bäckström, 2016). Consequently, one encounters a diversification of measures as well as an overall plethora of facets for many constructs (Pace & Brannick, 2010).

Relevant substantive approaches specifically concerned with the explication of facets and testing multi-faceted constructs have emerged within recent decades (Chen, Hayes, Carver, Laurenceau, & Zhang, 2012; Costa & McCrae, 1998; Hull, Lehn, & Tedlie, 1991). To various extents, available item-selection and -evaluation procedures can also be applied to the assessment of facets (see Smith, Fischer, & Fister, 2003). The problem is that the available approaches were not developed with the aim of identifying problem facets detrimental to validity, viz. redundant facets and, to a lesser extent, extraneous facets (see Siegling, Petrides, & Martskvishvili, 2015, for a more detailed conceptualisation of problem facets). The decisive characteristic of redundant and extraneous facets is that neither of them represent unique elements of the target construct; extraneous facets represent no elements whatsoever. It is this characteristic that (the authors contend) the conventional validation and scale development approaches cannot meticulously unveil.

Although much progress has been made to disentangle different sources of variance statistically (Morin et al., 2016; Raykov & Marcoulides, 2016; Schmid & Leiman, 1957), the reliable identification of redundant and extraneous facets, based on their inability to occupy unique construct variance, is not simply a matter of statistics. It depends heavily on the selection of variables and input data. Facets are typically evaluated against one another (i.e., along with variables characterised by a similar level of uncertainty). If a set of facets represents the target construct poorly, extraneous facets are more likely to load on the latent variable, and examining multicollinearity (e.g., in confirmatory factor analysis) is no trustworthy approach to detecting redundant facets. It is, thus, risky to assume that even advanced statistical procedures reliably distinguish the real target construct from other constructs as well as redundant construct variance from unique construct variance. Importantly, the concept of unique construct variance differs from specific variance; the former refers to a facet's unique part of the target construct and the latter to the part that is unrelated to the construct (see Fig. 1).

This article further examines the efficacy of Facet Benchmarking (FB), a recently proposed instrument refinement method that sets out to identify redundant and extraneous facets systematically (Siegling et

* Corresponding author at: London Psychometric Laboratory, University College London, WC1H 0AP, UK.
E-mail address: a.siegling@ucl.ac.uk (A.B. Siegling).

http://dx.doi.org/10.1016/j.paid.2016.12.014
0191-8869/

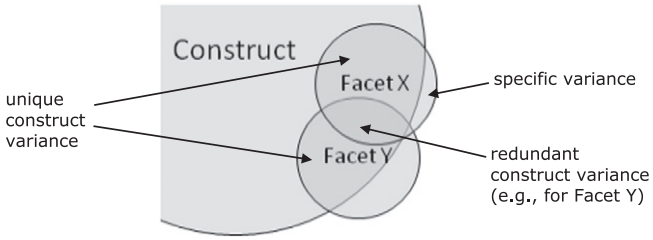

Fig. 1. Decomposition of construct variance into unique and redundant construct variance.

al., 2015). A concise description and advanced application guidelines are given next.

1.1. Facet Benchmarking (FB)

The concept of criterion validity has relevance in the identification of redundant and extraneous facets. Without unique construct variance, a facet is less likely to explain (unique) variance in construct-relevant criteria. Also, correlations of a scale composite that encompasses problem facets with construct-relevant criteria are systematically, although not necessarily always, lower than those of a composite comprised only of valid facets (see Smith et al., 2003, for a more detailed discussion of this effect). The construct-unrelated variance imposed on the composite by extraneous facets further compromises the composite's criterion validity (where construct-relevant criteria are concerned). Smith et al. have discussed how the external approach, or criterion keying, can be extended by means of incremental validity principles, with the aim of identifying and retaining facets with unique explanatory effects. However, the pivotal question not addressed in their seminal article concerns the criteria to be used for assessing the incremental value of individual facets, or whether each facet occupies unique construct variance.

One issue in leveraging criteria for the purpose of assessing facets is that, individually, they are unlikely to qualify as a comprehensive construct representation (Epstein, 1984; John & Benet-Martinez, 2000). Furthermore, like facets, individual criteria can comprise specific variance unrelated to the target construct; they are often multidimensional and cannot be expected to represent the construct variance exclusively (Smith & Zapolski, 2009). Due to sources of variance other than the target construct, there would be an increased chance of seeing explanatory effects of extraneous facets and, to a lesser extent, redundant facets. It also is realistic that some facets correlate positively with a particular criterion, whilst other facets correlate negatively with the same criterion (Ziegler, Danay, Schoelmreich, & Buehner, 2010).

As a remedy to the difficulties, individual benchmarks can be objectively derived from the shared variance of representative and balanced sets of construct-relevant criteria, selected with the construct as a reference point. Precisely, such a latent variable may be viewed as an approximation of the construct variance, with its accuracy depending on the method of derivation and knowledge about the construct already existing. In a five-stage process, FB examines whether a facet can occupy a unique portion of variance in these benchmarks.

1.1.1. Stage 1

The challenge is to select a set of construct-relevant criteria that represents the construct variance comprehensively (i.e., not missing any variance) and exclusively (i.e., not imposing variance unrelated to the construct). While both these requirements inevitably involve a theoretical process, exclusiveness is considerably facilitated by the statistical procedures described at Stage 2. Comprehensiveness is facilitated by incorporating varying, systematically selected sets of criteria, if necessary. Ideally, one would obtain a representative sample of all construct-relevant criteria without duplicating any elements, thus aiming for a balanced representation (it seems undesirable to use all conceivable criteria, since many of them are likely to overlap in their construct-related variance). If the benchmark is unbalanced with respect to the construct, the construct variance represented would shift towards individual facets, which can bias the FB results.[1]

Perhaps most straightforward is to rely on variables conceptualised as at least partial, direct psychological outcomes and, perhaps, known to correlate in the expected direction with the target construct. Indirectly-related outcomes increase the chances of significant explanatory effects of extraneous facets, since these are less likely to represent the target construct primarily. Although, prior empirical correlations may not be necessary, and other, more theory-driven approaches may be incorporated in making these decisions. Another consideration warranted during the criteria selection process are situational moderators, which can influence facet-criterion relationships. For instance, the central tenet of Trait Activation Theory is that situational factors (e.g., job demands, distractors) influence the expression of personality traits and their associations with relevant outcomes (Tett & Burnett, 2003). It is vital that the chosen criteria are either relevant across situations (i.e., general) or systematically sampled from all pertinent situations.

1.1.2. Stages 2 and 3

The basic idea is to extract the first latent factor, or benchmark, from the criteria administered to each sample, and then examine which facets occupy unique variance within this benchmark. There are two salient options for execution: (1) separately via factor analysis[2] and multiple regression; and (2) jointly via Structural Equation Modelling (SEM). Partially, the procedural choice depends on context considerations, such as sample size and number of facets and criteria involved.

1.1.2.1. The fragmented procedure: factor analysis and regression. The first latent factor is, in theory, the variable representing the target construct, since the criteria were selected using the construct as the reference point. As a proxy representation of the homogenous construct, the benchmark (an alternative latent variable of the construct variance) is most appropriately extracted via principal axis factoring (1 factor, no rotation), although we have previously used Principal Component Analysis for this purpose (the two procedures tend to yield virtually identical results for the first variable extracted). Any unrelated criteria (i.e., those that do not load well on the first factor) are identified and excluded in this process.

The question is at what threshold to drop or retain a criterion. Any divergent criteria may still co-vary due to sources other than the target construct, such as common method effects or chance. Consequently, they can introduce construct-unrelated variance on the benchmark. On the other hand, there is a danger of dropping valid criteria of the construct. For the time being, it makes sense to proceed with a generic minimum loading of 0.30, the common cut-off for scale items or facets in scale construction. A pre-specified value is intended to foster reliability and replicability of results, although it may be unwise to strictly advocate a specific cut-off. The important point is that adjustments are made a priori, guided by reason and theory.

Stage 3 of FB examines whether each of the facets occupies unique variance within the derived benchmark and if the variance explained is in the expected direction. A suitable statistical procedure for this purpose is statistical regression (also referred to as the stepwise method), with all facets entered at the initial step as explanatory variables of the benchmark. Stepwise regression both removes (criterion:

[1] Although a balanced representation of the construct is desirable in the context of FB, it is does not seem fatal if some criteria included within the benchmark are redundant with one another (if they do not share any specific or error variance with redundant or extraneous facets). Most redundant facets will be unable to account for unique construct variance, irrespective of whether their redundant variance is duplicated within an objectively derived benchmark.

[2] Note that the general limitations of factor analysis as a stand-alone statistical procedure in screening out redundant and extraneous facets are compensated within the context of FB.

$p \geq 0.05$) and possibly reenters (criterion: $p < 0.05$) predictors one-by-one, based on their ability to account for unique benchmark variance. In this process, redundant and extraneous facets may initially suppress (the significant effects of) valid facets. Yet, stepwise regression reenters facets removed at preceding steps if they gain significant explanatory effect at later steps. Of note, facets with betas in an unexpected direction are atheoretical and detrimental to validity. If present at the final step, the analysis is to be repeated without such facets.

To account for chance effects, facets that show significant betas on only rare occasions (e.g., <5% of the time) or of generally negligible magnitude (if one prefers to use effect size) may also be considered redundant or extraneous. To ascertain sufficient statistical power, the sample size should conform to accepted standards and best practices (Kelley & Maxwell, 2003; Tabachnick & Fidell, 2007), some of which revolve around the number of independent variables (predictors). As long as the minimum requirements are met, sample size variations tend to have relatively little influence on the number of significant predictors in automated subset selection algorithms (Derksen & Keselman, 2011).

The advantages of the fragmented procedure relate to the exploratory nature of FB, since it identifies problem facets based on specified statistical criteria; it requires relatively little manual labour, especially in the case of many facets. Whilst generally criticised, application of stepwise regression in the context of FB is justified in Appendix A. The criticisms of this procedure have little relevance with respect to the questions FB seeks to address or are compensated by the design and focus of FB on non-significant predictors.

1.1.2.2. The combined procedure: SEM. For the most part, the above procedure can be combined within an SEM framework, where one can draw on at least two different conditions for eliminating weak criteria: absolute loadings and significance of loadings (technically, this is possible too in the context of factor analysis if using the maximum-likelihood method of extraction). Once again, these decisions ought to be determined a priori, and the extent to which different selection criteria influence results has yet to be examined. Non-significant pathways to the benchmark can be dropped, but the corresponding facets should be retained within the model and modification indices continuously examined, in case these facets gain significance upon successive path deletions.

Although this procedure may involve more manual fiddling, its advantage is flexibility. For example, it facilitates the integration of additional variables, such as modelling common method factors. One can also directly examine the consequences of removing individual criteria on the explanatory effects of the facets.

1.1.3. Stage 4

Stage 4 serves to ascertain that no loss in validity occurs because of removing any problem facets as well as to get a basic idea of the magnitude of any improvement attained. The benchmark derived at Stage 2 is used as a gauge for this purpose: a modified scale composite, computed from facets showing significant explanatory effects in at least one of the samples used, is compared in its association with the benchmark to that of the original scale composite. As discussed above, the correlation of a composite containing non-explanatory facets with the benchmark should be systematically weaker than that of a composite comprised of explanatory facets only.

1.1.4. Stage 5

Stage 5 classifies the identified problem facets as redundant versus extraneous, based on associations with the modified scale composite. Consistently non-negligible correlations suggest that these facets are likely redundant, whereas non-significant correlations suggest that they are extraneous.

1.2. Initial findings and present investigation

A preliminary application of FB in the context of trait emotional intelligence showed promising results (Siegling et al., 2015). Data from six samples, which completed a broad, 15-facet measure of trait emotional intelligence and measures of construct-relevant criteria, exposed four facets that did not explain unique benchmark variance in any sample (an additional facet explained variance in a direction opposite to that explained by the other facets in some of the samples). Consequently, a composite of the 10 remaining facets converged significantly better with the benchmark than the original 15-facet composite in each sample. Although the criteria used to derive the benchmark implied some degree of consistency of facet effects, a limitation of the study is that it relied on pre-existing datasets. Despite involving a broad and diverse set of criteria, there is no guarantee that all elements of the construct variance were represented.

The present investigation is an advanced application of FB, based on data collected specifically for applying the method. Aimed at increasing certainty in the method's robustness and efficacy, the criteria were selected systematically with aim of obtaining an accurate representation of the target construct. The same set of criteria used to derive the benchmark was assessed in multiple samples (Study 1), and an entirely different set of criteria was used to cross-validate the results on another sample (Study 2). Moreover, the combined SEM procedure was applied to validate the results obtained via the fragmented procedure of factor analysis and statistical regression, also offering the opportunity of examining common method variance as a competing explanation.

A narrow construct comprised of relatively few facets was used to facilitate a systematic application of FB: dispositional mindfulness (broadly conceptualised as the practise of living in, and accepting, the present moment as it is, rather than being preoccupied). The mindfulness literature in the past 10 to 15 years has seen a proliferation of measures (e.g., Bergomi, Tschacher, & Kupper, 2013), most of which seem to tap into the same dimension (Baer, Smith, & Allen, 2004; Baer, Smith, Hopkins, Krietemeyer, & Toney, 2006; Siegling & Petrides, 2014). In the interest of space, we refer the reader to recent publications that provide a concise account of construct theory and measurement, also featuring a theoretical comparison of facets (e.g., Bergomi et al., 2013; Siegling & Petrides, 2016).

2. Study 1

FB was applied to the Five Facet Mindfulness Questionnaire (FFMQ; Baer et al., 2006) and Kentucky Inventory of Mindfulness Skills (KIMS; Baer et al., 2004), which have facet scores suitable for use in research and of satisfactory reliability. A predecessor of the FFMQ, the KIMS was used to increase certainty that the results are not influenced by measurement error of an individual measure. The additional four measures involved in the development of the FFMQ were used to examine the validity of the benchmark.

An additional purpose of Study 1 was to illustrate how the method's utility extends to examining the validity of facets or subscales across multiple measures. For example, FB can reveal whether conceptually identical or similar facets from different measures are redundant with one another, or if each encompasses unique manifestations of the same construct element (indicating that both are too narrow). FB was reapplied to the FFMQ facets and the four subscales of two bidimensional measures: the Philadelphia Mindfulness Scale (PHLMS; Cardaciotto, Herbert, Forman, Moitra, & Farrow, 2008) and Toronto Mindfulness Scale (TMS; Davis, Lau, & Cairns, 2009). Although three subscales from these additional measures (PHLMS Awareness, TMS Decenter, and TMS Curiosity) seem to diverge psychometrically from a superordinate mindfulness dimension (Siegling & Petrides, 2016), they were nonetheless included here to further demonstrate the efficacy of FB in distinguishing problem and valid facets. FB should be able to cross-validate factor analytical results.

Criterion variables used were those most frequently employed in previous development studies of mindfulness scales. This procedure resulted in a manageable number of variables but was also conducive to a comprehensive, consensus-based representation of the construct variance. The focus was on relatively narrow psychological expressions directly linked to mindfulness, rather than broader clinical and mental health criteria (e.g., alexithymia, depression, and anxiety) or other personality constructs (e.g., emotional intelligence). Table 1 shows the selected criteria, along with their occurrence in mindfulness scale development studies.

The combined SEM procedure was incorporated at Stages 2 and 3 (for the FFMQ) as a way of exploring statistical artefacts linked to an individual analytical procedure (in this case, the fragmented procedure). A second aim of this component was to examine common method variance as a competing explanation of the results. The marker-variable technique described, for instance, in Johnson, Rosen, and Djurdjevic (2011) was applied for this purpose. Extraversion was used as the marker variable, being largely distinct from mindfulness (Giluk, 2009; Siegling & Petrides, 2014) and showing relatively high associations with social desirability indices (Bäckström & Björklund, 2013).

2.1. Method

2.1.1. Samples and procedure

FB was applied to three community and convenience samples. None of the samples were recruited from mindfulness populations, such as meditators or recipients of mindfulness-based therapy. Thus, it should be assumed that the samples are mostly inexperienced in mindfulness. Table 2 presents a summary of their characteristics.[3]

Samples 1 and 2 were recruited via the institutional subject pool of a British university over approximately two years (February 2012–March 2014). The combined sample size was 397, which was split randomly in half, with an equal proportion of male and female students assigned to each subsample. The subject pool includes undergraduate and Master's students, predominantly from psychology or linguistics backgrounds, as well as some non-student affiliates of the university community. Most students received course credit and all participants were entered into a draw for gift cards.

Sample 3 was recruited online. A recruitment notice was posted on participant recruitment platforms for psychological research (e.g., http://www.onlinepsychresearch.co.uk/). As an additional means to attracting (interested) participants, links to these posts were disseminated via the social media pages of two promoters of mindfulness. Many of the participants had earned a Bachelor (35.7%) or Master's (25.1%) degree; most of the remaining cases had not gone beyond secondary education (32.2%) at the time of the study. Over half the sample was working full- or part-time (55.0%), while approximately a third of the sample was studying full- or part-time (33.9%). A price draw of gift cards was offered to participants as a token of appreciation.

Participation occurred via an anonymous online survey system. Participants provided demographic information and completed the measures below in randomised order. They were automatically notified of any missing responses and given the opportunity to add them.

Table 1
Study 1: commonly used validation criteria in the development of mindfulness scales.

Criteria	Mindfulness scales validated
Experiential avoidance	FFMQ, KIMS, CAMS–R, PHLMS
Rumination and reflection	CAMS–R, LMS, MAAS, TMS, PHLMS
Thought suppression	FFMQ, CAMS–R, LMS, PHLMS
Worry	CAMS–R, LMS
Absent-mindedness	FFMQ, TMI
Dissociative activities	FFMQ, KIMS, FMI, TMS
Absorption	KIMS, MAAS
Self-consciousness	MAAS, FMI, TMI
Positive and negative affect	SMQ, LMS
Emotion regulation	FFMQ, LMS

2.1.2. Measures

The measures were based on self-report and multiple-point response scales. Across samples, all variables achieved alpha coefficients above 0.70.

2.1.2.1. Mindfulness

2.1.2.1.1. FFMQ (Baer et al., 2006). The FFMQ is a relatively comprehensive operationalisation of the construct, constructed by factor-analysing the items of five other measures. It consists of five facets (Observe, Describe, Act with Awareness, Accept without Judgement, and Nonreact) that can be combined to produce a global mindfulness score. The items (7–8 per facet) are rated on a 5-point Likert scale, ranging from 1 (*never or very rarely true*) to 5 (*very often or always true*).

2.1.2.1.2. KIMS (Baer et al., 2004). The KIMS has four facets (8–12 item per facet): Observe, Describe, Act with Awareness, and Accept without Judgement. These are now contained within the FMMQ. It is based on the same 5-point response scale as the FFMQ.

2.1.2.1.3. Cognitive and Affective Mindfulness Scale–Revised (CAMS–R; Feldman, Hayes, Kumar, Greeson, & Laurenceau, 2006). The CAMS–R yields a global mindfulness score, based on four facets (attention, present focus, awareness, and acceptance; the scale developers advise against the use of facet scores). Its 12 items are rated on a 4-point Likert scale from 1 (*Rarely/Not at all*) to 4 (*Almost Always*).

2.1.2.1.4. Southampton Mindfulness Questionnaire (SMQ, Chadwick et al., 2008). The 16-item scale composite consists of four facets: mindful observation, letting go of reacting, opening awareness to difficult experience, and acceptance (the scale developers also advise against the use of facet scores). The response scale ranges from 0 (*Disagree Totally*) to 6 (*Agree Totally*).

2.1.2.1.5. Mindful Attention Awareness Scale (MAAS; Brown & Ryan, 2003). The MAAS is restricted to attentional aspects of mindfulness. Its 15 items are rated on a 6-point Likert scale, ranging from 1 (*Almost Always*) to 6 (*Almost Never*).

2.1.2.1.6. Freiburg Mindfulness Inventory (FMI; Walach, Buchheld, Buttenmüller, Kleinknecht, & Schmidt, 2006). Although designed to assess global mindfulness, two highly interrelated factors have been derived from the FMI items (Kohls, Sauer, & Walach, 2009): attention to present moment (presence) and non-judgemental attitude (acceptance). The 14 FMI items are based on a response scale of 1 (*rarely*) to 4 (*almost always*).

2.1.2.1.7. PHLMS (Cardaciotto et al., 2008). The PHLMS was explicitly designed to operationalise two orthogonal, 10-item subscales: Acceptance and Awareness. Items are rated on a 5-point response scale, ranging from 1 (*Never*) to 5 (*Very Often*).

2.1.2.1.8. TMS (trait version; Davis et al., 2009). The TMS was created to permit oblique factors, but correlations of its two subscales, Curiosity (6 items) and Decenter (7 items), are not particularly indicative of a shared higher-order factor, ranging from $r = 0.26$ to 0.42 (Lau et al., 2006). The subscales were, therefore, interpreted as measuring distinct but related latent constructs. The items are responded to on a 4-point Likert scale, ranging from 1 (*Not at all*) to 5 (*Very much*).

2.1.2.2. Criteria. The measures used in previous validation studies of mindfulness scales were used here. Newer versions of the original measures were utilised for experiential avoidance and absorption. The number of items per criterion is given in Table 3.

2.1.2.2.1. Experiential avoidance. This criterion was measured with the Acceptance and Action Questionnaire II (Bond et al., 2011). Higher scores indicate greater experiential avoidance (alteration of the form, frequency, and situational sensitivity of experiences). The scale items

[3] Treatment of missing data is described in a related article (Siegling & Petrides, 2016). The same procedures were followed in Study 2.

Table 2
Study 1: demographic characteristics of study samples.

Sample (*N*)	Age (years)			Gender (*n*)		Ethnicity (%)				
	M	*SD*	Range	Male	Female	Caucasian	East Asian	South Asian[a]	African	Other/mixed
1 (199)	21.9	4.3	18.0–57.2	46	153	55.3	27.1	9.0	3.5	5.0
2 (198)	21.9	5.4	18.2–55.0	46	152	55.0	29.8	7.1	1.0	7.1
3 (171)	37.3	14.2	18.0–76.2	35	136	84.8	2.3	1.8	4.7	6.4

Note. Samples 1 and 2 two are split-halves of a university student sample.
[a] Includes Pakistani, Bangladeshi, Indian, and Sri Lankan backgrounds.

are responded to on a 7-point Likert scale from 1 (*never true*) to 7 (*always true*).

2.1.2.2.2. Rumination and reflection. The Rumination-Reflection Questionnaire (Trapnell & Campbell, 1999) is based on a bidimensional model of private self-consciousness and has two subscales: rumination and reflection. The items have a 5-point Likert scale, ranging from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*).

2.1.2.2.3. Thought suppression. The White Bear Suppression Inventory (Wegner & Zanakos, 1994) measures a person's attempts to suppress particular thoughts. The items have a 5-point Likert scale, ranging from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*).

2.1.2.2.4. Worry. The Penn State Worry Questionnaire (Meyer, Miller, Metzger, & Borkovec, 1990) is a measure of worry, a dominant feature of generalised anxiety disorder. The items have a 5-point Likert scale, ranging from 1 (*not at all typical of me*) to 5 (*very typical of me*).

2.1.2.2.5. Absent-mindedness. The Cognitive Failures Questionnaire (Broadbent, Cooper, Fitzgerald, & Parkes, 1982) measures the frequency of mistakes people make in perception, memory, and motor function, but it was also conceptualised as a measure of absent-mindedness. The items have a 5-point Likert scale, ranging from 0 (*Never*) to 4 (*Very often*).

2.1.2.2.6. Dissociative activities. The Scale of Dissociative Activities (Mayer & Farmer, 2010) measures dissociative behaviours. Responses are given on a 5-point Likert scale and range from *Never* to *Very Frequently*.

2.1.2.2.7. Absorption. The Modified Tellegen Absorption Scale (Jamieson, 2005) was used to measure the disposition of getting absorbed in mental imagery. It differs from the original scale in using a multi-point response scale, ranging from 0 (*never*) to 4 (*very often*).

2.1.2.2.8. Self-consciousness. The Self-Consciousness Scale (Fenigstein, Scheier, & Buss, 1975) has three subscales: private self-consciousness, public self-consciousness, and social anxiety. Items are responded to on a 5-point scale, ranging from 0 (*extremely uncharacteristic* [*not at all like me*]) to 4 (*extremely characteristic* [*very much like me*]).

2.1.2.2.9. Positive and negative affect. These two mood dimensions were measured with the Positive and Negative Affect Schedule (Watson, Clark, & Tellegen, 1988), which consists of adjectives of positive or negative valence. Each adjective is rated on a 5-point scale, ranging from 1 (*very slightly or not at all*) to 5 (*extremely*).

2.1.2.2.10. Emotion regulation. The Emotion Regulation Questionnaire (Gross & John, 2003) measures two distinct aspects of emotion regulation: emotion reappraisal (people's inner experience of emotions) and emotion suppression (the behaviour linked to people's feelings). Scale items are based on a 7-point response scale, ranging from 1 (*strongly disagree*) to 7 (*strongly agree*).

2.1.2.3. Extraversion. The Big Five Inventory (John & Srivastava, 1999) was used to measure Extraversion. Respondents indicate the degree to which brief descriptive items apply to them on a 5-point scale, ranging from 1 (*disagree strongly*) to 5 (*agree strongly*).

2.1.2.4. Statistical analysis. At Stage 2 of FB, the first factor was extracted via principal axis factoring, and any criteria showing consistently weak loadings (<0.30) on this factor were identified and excluded from the benchmark. The benchmark was then regressed separately on the FFMQ and KIMS facets in each sample, as described in the Introduction (Stage 3). The procedure was then repeated for a joint analysis of the FFMQ facets with the PHLMS and TMS subscales. The data from most Sample 1 and 2 participants (those who completed the measure of Extraversion) were then merged to apply the combined SEM procedure (Stages 2 and 3) to the FFMQ facets, while partialling out common method variance. This analysis was carried out using AMOS (Arbuckle, 2014), using maximum-likelihood estimates.

Steiger's *Z* test was computed separately for the FFMQ and KIMS (Stage 4). The modified scale composites were compared against the respective original scale composites in their associations with the benchmark. At Stage 5, any identified problem facets were correlated with the modified scale composite to distinguish them as redundant versus extraneous. A follow-up analysis examined the extent to which the benchmark yields an accurate representation of the construct variance: average bivariate correlations of the benchmark with mindfulness scales were compared to the average intercorrelations among the

Table 3
Study 1: principal axis factor analysis results for validation criteria.

Criteria (no. of items)	Sample 1 ($N = 199$)		Sample 2 ($N = 198$)		Sample 3 ($N = 171$)	
	Factor loading	Communality	Factor loading	Communality	Factor loading	Communality
Absent-mindedness (25)	0.57	0.32	0.63	0.40	0.62	0.39
Experiential avoidance (10)	0.78	0.61	0.75	0.56	0.85	0.73
Rumination (12)	0.77	0.60	0.74	0.54	0.84	0.70
Reflection (12)	0.24	0.06	0.32	0.10	0.12	0.01
Dissociative activities (35)	0.66	0.44	0.69	0.47	0.69	0.48
Thought suppression (15)	0.73	0.54	0.78	0.61	0.81	0.66
Positive affect (10)	−0.14	0.02	−0.09	0.01	−0.42	0.18
Negative affect (10)	0.63	0.40	0.70	0.49	0.72	0.52
Private self-consciousness (10)	0.50	0.25	0.53	0.28	0.46	0.21
Public self-consciousness (7)	0.64	0.41	0.58	0.34	0.64	0.41
Social anxiety (6)	0.53	0.28	0.46	0.21	0.64	0.40
Absorption (34)	0.39	0.15	0.47	0.22	0.34	0.12
Worry (16)	0.70	0.50	0.62	0.39	0.78	0.61
% of variance	35.12		35.47		41.70	

Note. Emotion reappraisal (6 items) and emotion suppression (4 items) were excluded, due to weak loadings on the first latent factor (<0.30).

mindfulness scales across samples. The analysis was conducted both with and without the FFMQ, which derives from the other scales and, thus, might introduce bias.

2.2. Results and discussion

2.2.1. Factor analysis of criteria

Two criteria (emotional reappraisal and suppression) were not assessed in Sample 3 and only in 111 and 90 participants in Samples 1 and 2, respectively. These variables were removed, since they did not load on the same factor as the other criteria in a preliminary analysis of the data. Table 3 shows the principal axis factoring results for all other criteria, which loaded on the first latent factor in at least one, but in most cases all, of the samples. The benchmark was derived from these criteria, omitting emotional reappraisal and reflection only.

2.2.2. Regression of benchmark on facets

The results shown in Table 4 indicate at least some consistency across samples. Ignoring the FFMQ's additional facet (Nonreact), which reached significant betas, results were also similar between measures. One facet (Observe) showed an atheoretical explanatory effect, opposite to that of the other facets. This effect appeared for both measures in almost all instances (the facet did not reach significance for the KIMS in Sample 2). The Describe facet showed a significant effect in one sample and for only one measure (KIMS), while the other two facets consistently occupied theory-consistent benchmark variance, indicated by negative beta weights.

Extension of the FFMQ analysis by inclusion of the PHLMS and TMS subscales exposed the same three FFMQ facets as significant explanatory variables. Additionally, FFMQ Describe reached significance in Sample 3, as already seen for the equivalent KIMS facet in that sample. Of the PHLMS and TMS subscales, PHLMS Acceptance consistently reached a significant and negative beta weight. Since the conceptually similar FFMQ Accept without Judgement facet remained a significant predictor, each variable seems to encompass unique content of the mindfulness element in question. Also, the bivariate correlation between these two variables is not as high to suggest equivalence (Cardaciotto et al., 2008; Siegling & Petrides, 2016).

2.2.3. SEM

The results summarised in Fig. 2 present the initial model without any removal of facets. Consistent with preceding results, Observe maintains its atheoretical effect, while the other four FFMQ facets show significant negative regression weights. A common latent factor shares negligible variance with the criteria and none with the FFMQ facets, meaning that there is no evidence that common method variance accounts for the results. In sum, neither a different computation method (Maximum Likelihood) nor a modelled common latent (method) factor changed the results as regards FFMQ Observe.

2.2.4. Composite correlations with benchmark

Correlations of the modified scale composites (omitting the Observe items) with the benchmark were compared to the respective original composites in their associations with the benchmark. These correlations were significantly higher for the modified scale composites in all three samples ($p < 0.01$) and for both measures (range of r difference: 0.07 to 0.15 for FFMQ, 0.11 to 0.26 for KIMS). Hence, the Observe items compromise the strength of association with the benchmark.

2.2.5. Correlations of observe facet with modified scale composites

FFMQ/KIMS Observe facet correlated either non-significantly (Samples 1 and 2; $r = -0.08$ to 0.06) or weakly (Sample 3; $r = 0.28$ and 0.21) with the modified composite. The facet seems marginally related to mindfulness in these non-meditating samples, a pattern characteristic of that expected for an extraneous facet.

2.2.6. Follow-up analysis

Average scale intercorrelations for Samples 1 to 3 were 0.53, 0.51, and 0.65, respectively (of the Sample 3 participants, only 115 completed the MAAS and FMI). Average correlations of the benchmark with these scales were similar to these scale intercorrelations ($r = -0.56$, -0.54, and -0.63). Upon excluding the FFMQ, average correlations of the benchmark with mindfulness scales became somewhat larger ($r = -0.56$, -0.53, and -0.61) than the average scale intercorrelations ($r = 0.48$, 0.44, and 0.60). These results speak favourably to the validity of the benchmark.

Table 4
Study 1: stepwise regression analysis summaries for (a) FFMQ facets, (b) KIMS facets, and (c) FFMQ facets and PHLMS and TMS subscales as predictors of the benchmark.

Mindfulness scale and facets	Sample 1 ($N = 199$)			Sample 2 ($N = 198$)			Sample 3 ($N = 171$)		
	β	F	R^2_{Adj}	β	F	R^2_{Adj}	β	F	R^2_{Adj}
FFMQ (all facets)		39.51***	0.49		63.08***	0.61		61.12***	0.64
FFMQ (final model)			0.45		97.10***	0.59			0.61
Observe	–	54.19***		–			–	89.76***	
Describe	–			–			–		
Act with Awareness	−0.29***			−0.34***			−0.25***		
Accept without Judgement	−0.40***			−0.50***			−0.45***		
Nonreact	−0.22***			−0.26***			−0.28***		
KIMS (all facets)		50.00***	0.50		52.28***	0.50		56.68***	0.57
KIMS (final model)		81.11***	0.45		99.10***	0.50		75.80***	0.57
Observe	–			–			–		
Describe	–			–			−0.16**		
Act with Awareness	−0.27***			−0.57***			−0.19**		
Accept without Judgement	−0.53***			−0.29***			−0.61***		
FFMQ, PHLMS, and TMS		30.37***	0.57		38.45***	0.63		52.09***	0.73
Final model		51.24***	0.50		75.72***	0.60		70.32***	0.67
FFMQ describe	–			–			−0.15**		
FFMQ Act with Awareness	−0.30***			−0.34***			−0.21***		
FFMQ Accept without Judgement	−0.34***			−0.48***			−0.36***		
FFMQ Nonreact	−0.22***			−0.25***			−0.22***		
PHLMS Acceptance	−0.25***			−0.11*			−0.27***		

* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

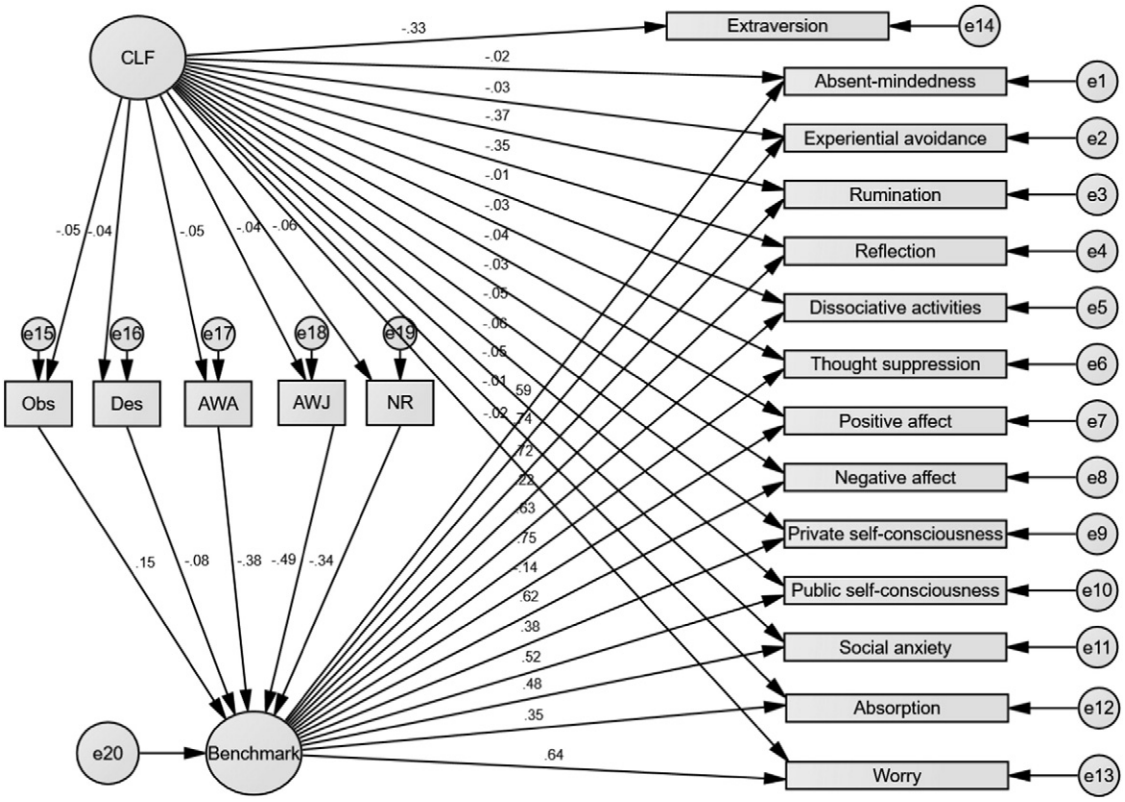

Fig. 2. $N = 358$. Explanatory effects of FFMQ facets on benchmark, while partialling out a common latent (method) factor (CLF) using Extraversion as a marker variable.

2.2.7. Strengths and limitations

Consistency of results, coupled with findings of other studies that cast doubt on the validity of the Observe facet, speak to the method's robustness and efficacy in identifying problem facets. The benchmark derived and used here seems to represent the mindfulness construct variance quite well, given the present conceptualisation and available measures. Yet, while a systematically derived set of criteria is a step forward in demonstrating the efficacy of FB, there is no guarantee that it represents the common variance of every valid facet. The study's emphasis on reliability necessitated the same set of criteria across samples, which made it less practical to include all the criteria considered previously used in mindfulness research. Study 2 was conducted to ascertain that the results generalise to a different set of criteria.

3. Study 2

To exhaust the list of previously used criteria, this study relied on the criteria previously used in mindfulness scale validation studies and not already included in Study 1. Using only the least frequently used criteria for this purpose is a dangerous approach, as they are less likely to represent mindfulness as accurately as the more common criteria used in Study 1. On the other hand, confirming evidence from those criteria would yield strong support for the results obtained in Study 1 and, more generally, for the efficacy of FB.

The following criteria were utilised here (the respective mindfulness scales for which they were previously used are shown in parentheses): curiosity (Langer Mindfulness Scale [LMS]; Bodner & Langer, 2001; Pirson, Langer, Bodner, & Zilcha, 2012), need for cognition (MAAS), self-monitoring (MAAS), overgeneralisation (CAMS–R), self-compassion (FFMQ), psychological mindedness (TMS), anticipatory mental coping (4 subscales; CAMS–R), and subjective happiness (PHLMS). One criterion, hopelessness (PHLMS), was somewhat ambiguous in terms of its suitability. Hopelessness has somewhat of a clinical nature and represents thought content, rather than cognitive or affective processes or style. The decision was therefore made to proceed without this criterion.

3.1. Method

3.1.1. Participants and procedure

Participants were recruited in the same way as Samples 1 and 2 of Study 1. The sample consisted of 188 undergraduate and Master's students (87.2% female), mostly from Psychology and Linguistics. Participant ages averaged to 21.1 years ($SD = 5.7$) and ranged from 17.9 to 74.5 years. Ethnic backgrounds were mostly Caucasian (56.4%) and Asian (31.4%); the remaining were South Asian (India, Pakistan, Sri Lanka, Bangladesh; 4.3%), African (2.1%), or multi-ethnic (5.9%).

3.1.2. Measures

3.1.2.1. Mindfulness. FB was reapplied to the FFMQ and KIMS (see Study 1). All facets reached satisfactory alpha coefficients ($\alpha = 0.78$ to 0.90).

3.1.2.2. Criteria. The criteria also reached satisfactory alpha coefficients ($\alpha = 0.78$ to 0.88), except for one instance, where McDonalds omega is reported additionally. The number of items per criterion is shown in Table 5.

3.1.2.3. Curiosity. The Curiosity and Exploration Inventory-II (Kashdan et al., 2009) composite score was used. The scale items are responded to on a 4-point Likert scale, ranging from 1 (*Very Slightly or Not At All*) to 5 (*Extremely*).

3.1.2.4. Need for cognition. The Need for Cognition Scale (Cacioppo, Petty, & Kao, 1984) is responded to on a 5-point Likert scale (1 = *extremely uncharacteristic of me*, 5 = *extremely characteristic of me*).

3.1.2.5. Self-monitoring. The Self-Monitoring Scale-Revised (Lennox & Wolfe, 1984) measures people's "ability to modify self-presentation" and "sensitivity to expressive behaviour of others". Its items are rated

on a 6-point scale, ranging from 0 (*certainly, always false*) to 5 (*certainly, always true*).

3.1.2.6. Overgeneralisation. Overgeneralisation, the disposition to generalise from individual failures to one's overall self-worth, was assessed using a subscale of the Attitudes Towards Self Scale (Carver, Voie, Kuhl, & Ganellen, 1988). The items are based on a 5-point Likert scale (1 = *I agree a lot*, 5 = *I disagree a lot*).

3.1.2.7. Self-compassion. The Self-Compassion Scale–Short Form (Raes, Pommier, Neff, & Van Gucht, 2011) is suitable for assessing the global construct and shows near-perfect correlations with the full form. The scale items are responded to on a 5-point scale, ranging from 1 (*Almost never*) to 5 (*Almost always*).

3.1.2.8. Psychological mindedness. The Psychological Mindedness Scale (Conte, Ratto, & Karasu, 1996) has been conceived of as measuring a person's capacity for tolerating psychological distress (Shill & Lumley, 2002). On a 4-point Likert scale (*strongly agree* to *strongly disagree*), respondents indicate the extent to which descriptive items represent them.

3.1.2.9. Anticipatory mental coping. The Measure of Anticipatory Mental Processes (Feldman & Hayes, 2005) assesses two productive and two unproductive strategies for coping with future stressful events. The four subscales are (Feldman & Hayes, 2005, pp. 490–491): problem analysis, plan rehearsal ($\alpha = 0.65$, $\omega = 0.81$), stagnant deliberation, and outcome fantasy. Respondents are asked to imagine a problem and then to indicate how often various items reflect their typical response in this kind of situation on a 5-point scale (1 = *Never true for me*; 5 = *Always true for me*).

3.1.2.10. Subjective happiness (Lyubomirsky & Lepper, 1999). The Subjective Happiness Scale items are responded to on a 7-point Likert scale from 1 to 7. The scale anchors vary across items.

3.2. Statistical analysis

The statistical stages of FB (2–5) were executed as in Study 1, but without the additional SEM procedure: (2) the criteria were submitted to a principal axis factor analysis to extract the first latent factor and identify any divergent criteria; (3) the derived benchmark was regressed separately on the FFMQ and KIMS facets, starting with all facets in the initial model; (4) associations of modified scale and original scale composites with the benchmark were compared; (5) associations of any problem facets with the modified scale composites were examined.

3.3. Results and discussion

3.3.1. Factor analysis of criteria

Results for the first latent factor underlying the criteria are shown in Table 5. Five criteria (psychological mindedness, planned rehearsal, need for cognition, self-monitoring, and problem analysis) did not load adequately on this factor ($\lambda < 0.30$). The benchmark was derived from the remaining six criteria.

Table 5
Study 2: principal axis factor analysis results for validation criteria.

Criteria (no. of items)	Factor loading	Communality	% of variance
Self-compassion (12)	−0.76 (−0.85)	0.06 (0.69)	20.82 (37.02)
Overgeneralisation (4)	0.72 (0.83)	0.52 (0.73)	
Subjective happiness (4)	0.69 (0.65)	0.58 (0.11)	
Stagnant deliberation (5)	0.48 (0.43)	0.19 (0.42)	
Curiosity (10)	0.43 (0.33)	0.05 (0.19)	
Outcome fantasy (2)	0.29 (0.31)	0.48 (0.09)	
Psychological mindedness (45)	0.24	0.23	
Planned rehearsal (3)	0.22	0.05	
Need for cognition (18)	0.22	0.08	
Self-monitoring (13)	0.20	0.04	
Problem analysis (5)	0.13	0.02	

Note. $N = 188$. Values in parentheses derive from an analysis excluding facets that did not satisfy the specified criteria for inclusion.

3.3.2. Regression of benchmark on facets

Stepwise regression analysis results are shown in Table 6. For both measures, the Observe facet did not reach the final step due to non-significant betas. The Describe facet occupied benchmark variance for the KIMS but not for the FFMQ. The remaining facets all reached significance for both measures. Despite using a completely different set of criteria, these results closely resemble those obtained in Study 1.

3.3.3. Composite correlations with benchmark

Zero-order correlations of the original FFMQ and KIMS composites with the benchmark were 0.60 and 0.45, respectively ($p < 0.001$). As expected, correlations involving the modified composites were slightly higher at 0.63 (FFMQ, $p < 0.001$) and 0.50 (KIMS, $p < 0.001$). These differences were not significant for either the FFMQ, $Z(185) = 1.40$, $p > 0.05$, or the KIMS, $Z(185) = 1.42$, $p > 0.05$, but the key outcome is that correlations for the modified composites were not weaker.

3.3.4. Correlations of observe facet with modified scale composites

FFMQ/KIMS Observe did not correlate significantly with the respective modified scale composite, (FFMQ: $r = 0.02$, $p = 0.74$; KIMS: $r = -0.06$, $p = 0.44$). These results agree with those in Study 1, where this facet correlated non-significantly or modestly with the modified scale composites. They suggest that the Observe facet is extraneous, and the atheoretical beta weights seen in some of the Study 1 analyses provide further support for this inference. Since a comprehensive set of proxy mindfulness criteria was used across the two studies, there is reason to assume that no important criteria were missing (and linked to Observe in way that would have led to different results).

4. General discussion

This article presented an advanced application of FB to a different construct. A key step forward was that the criteria were selected systematically, which increases the likelihood that they were representative of the construct. In fact, a hitherto comprehensive sample of scale validation criteria, representing direct outcomes of mindfulness, were included here. The current application of FB, thus, reflected current thinking on the focal construct (at the time of data collection, to be precise), minimising the chance of any valid facets not being represented by the derived benchmarks.

An advantage with respect to examining the reliability of results is that the same set of criteria was assessed in multiple samples (Study 1), and that generalisability of the results to a different set of

Table 6
Study 2: stepwise regression analysis summaries for FFMQ and KIMS facets as predictors of the benchmark.

Mindfulness scale and facets	β	*F*	R^2_{Adj}
FFMQ (all facets)		33.66***	0.47
FFMQ (final model)		55.52***	0.47
Observe	–		
Describe	–		
Act with Awareness	0.20***		
Accept without Judgement	0.40***		
Nonreact	0.43***		
KIMS (all facets)		18.98***	0.28
KIMS (final model)		24.07***	0.27
Observe	–		
Describe	0.12*		
Act with Awareness	0.17*		
Accept without Judgement	0.44***		

Note. $N = 188$.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

criteria was examined subsequently (Study 2). Additionally, FB was applied separately to two multi-faceted measures of the same construct, based on almost identical models (a study feature that alleviates concerns linked to random measurement effects). What is more, the method was advanced and scrutinised by integrating additional statistical and methodological components (SEM and modelling common method bias). These advances provide a foundation for ascertaining the method's robustness and efficacy.

4.1. Summary and interpretation of results

In both studies, and across samples, the same facet (Observe) was unable to occupy theoretically expected construct variance, as captured by the respective benchmarks. This pattern held up when applying SEM as a statistical alternative and including an unrelated marker variable for partialling out a general method factor. Removal of the Observe facet from the scale composites led to an improvement in validity, as evidenced by the relative magnitude of associations between the modified scale composites and the benchmark. Two additional measures were used to examine how FB performs when applied to multiple measures, and to illustrate its potential in examining facets across multiple measures. Consistent with factor analytical results based on the same data (Siegling & Petrides, 2016), PHLMS Awareness as well as TMS Curiosity and Decenter did not occupy benchmark variance.

The results build on the initial application of FB (Siegling et al., 2015), which exposed five redundant trait emotional intelligence facets that, like the FFMQ/KIMS Observe facet in the present investigation, compromised the measure's validity with respect to the benchmark. Yet, the current results also depart from those obtained previously, because the identified problem facet can be best classified as an extraneous facet. The results, therefore, suggest that FB is also able to spot extraneous problem facets not assertively identified as such by other methods.

The results fit into an increasing pattern of standard scale validation findings attesting to the distinctiveness of the FFMQ/KIMS Observe facet (in non-meditating samples). While evidence supports the originally envisaged five-factor model in meditators (e.g., Aguado et al., 2015; Baer et al., 2008), a four-factor hierarchical model without the Observe facet often results in better model fit for the FFMQ in non-meditating samples (e.g., Baer et al., 2008; Curtiss & Klemanski, 2014; Gu et al., 2016; Siegling & Petrides, 2016; Williams, Dalgleish, Karl, & Kuyken, 2014). The facet has also shown negligible incremental explanatory effects over the other facets, including some detrimental effects (Cash & Whittingham, 2010; Christopher & Gilbert, 2009; Consedine & Butler, 2013; Vujanovic, Bonn-Miller, Bernstein, McKee, & Zvolensky, 2010). Building on prior findings for this facet, the present results speak to the method's efficacy in identifying problem facets in a particular context (here, non-meditators).

4.2. Implications

There is now quite promising evidence supporting the robustness and efficacy of FB as an instrument refinement method. Although existing psychometric approaches have value in identifying extraneous facets, the current investigation shows that FB can provide specific, reliable, and efficiently gathered evidence that a given facet is extraneous. Still, FB is intended to supplement the existing psychometric approaches and best viewed as an ongoing process. For any given construct, the use of multiple samples and benchmarks will increase certainty in the FB results.

Beyond demonstrating the efficacy of FB in identifying redundant and extraneous facets, the findings suggest that FB can also identify individual measures (or facets) as incomprehensive, relative to other measures (or conceptually similar facets in other measures). Aside from offering a general comparison of measures and their constituent facets, the information can indicate whether individual measures would benefit from additional scale content in the form of items or entire facets. Otherwise, FB may indicate that some measures lack unique content.

A construct-specific implication of the findings concerns the representation and operationalisation of mindfulness. In conjunction with previous findings, the current investigation provides good evidence that FFMQ/KIMS Observe either is not a valid and useful facet in the general population or remains inadequately operationalised. The results give good indication that this facet is extraneous, but somehow survived initial factor-analytic work. Furthermore, the mindfulness facet represented in FFMQ Accept without Judgement and PHLMS Acceptance does not seem fully captured by either of these subscales, which could be expanded or integrated. To the contrary, the other PHLMS subscale and both TMS subscales seem to be either redundant or to tap into a different construct than the one underlying most mindfulness scales.

4.3. Limitations and future directions

Both the current and initial investigation of FB are limited in that that all measures involved were based on a self-report response format. Although an effort was made to control for common method variance, method effects cannot be ruled out with complete certainty. Such a scenario is arguably unlikely, because the results are consistent with factor-analytical results and there is no prima facie evidence that the Observe facet differs from the other facets in terms of method variance. Future research utilising different forms, or at least sources, of measurement for construct and criteria is needed to rule out method effects definitively (e.g., observer ratings, behavioural observations, and electronic diaries).

Another limitation is the use of convenience samples, with uneven distributions of demographic factors (e.g., gender) possibly impinging on the pattern of results obtained. Relatedly, the results may not generalise to meditating samples, in which the Observe facet has shown better psychometric results.

A general limitation of the research conducted on FB thus far is the focus on personality constructs. While relatively practical for scrutinising FB, the reliance on personality constructs restricts the evidence for the method's utility. As is the case with other psychometric approaches, FB should have utility in refining constructs other than personality traits, such as abilities, motives, or even psychological states. The hitherto promising findings certainly warrant the application of FB to other types of constructs.

Appendix A. Justification for utilising stepwise regression at Stage 3 of Facet Benchmarking

A general concern is that various automated selection algorithms lack theoretical basis, operating purely on a pre-specified empirical criterion. Where the predictors of an important criterion of interest are examined, it is relatively unsophisticated to rely on some automated selection procedure, especially when using different types of predictors. In the context of FB, however, the predictors are all the same type (i.e., facets of the target construct) and the question to be answered is a statistical one: whether the facets explain unique benchmark variance. There is no theoretical order among the facets, and the focal question of unique benchmark variance can hardly be answered on theoretical grounds.

A second criticism is the removal of predictors based on their ability to explain criterion variance. Multicollinearity among predictors is generally considered problematic, because it can compromise the explanatory effects of individual predictors (Pedhazur, 1997). Somewhat paradoxically, FB capitalises on this principle in identifying redundant facets. Essentially, high intercorrelations mean that the predictors concerned are likely to explain (much of) the same criterion variance, rendering some of them redundant. Regardless of their intercorrelations

with other facets, extraneous facets should not explain any notable variance of a benchmark.

Another concern is that the procedure is unduly influenced by chance features of the data and that the ensuing models are therefore difficult to replicate. In testing multiple models, stepwise regression is prone to overfitting the data. FB partly accounts for this limitation by means of built-in replication, conducted across the same and different benchmarks. More fundamentally, however, it does not require the same solution to be obtained across samples. The crux are the predictors that never find their way into (the final step of) the regression models.

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*, 1–14. http://dx.doi.org/10.3389/fpsyg.2015.00404.

Arbuckle, J. L. (2014). *Amos 23.0 user's guide*. Chicago: IBM SPSS.

Bäckström, M., & Björklund, F. (2013). Social desirability in personality inventories: Symptoms, diagnosis and prescribed cure. *Scandinavian Journal of Psychology*, *54*(2), 152–159. http://dx.doi.org/10.1111/sjop.12015.

Baer, R. A., Smith, G. T., & Allen, K. B. (2004). Assessment of mindfulness by self-report: The Kentucky inventory of mindfulness skills. *Assessment*, *11*(3), 191–206. http://dx.doi.org/10.1177/1073191104268029.

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://dx.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), 329–342. http://dx.doi.org/10.1177/1073191107313003.

Bergomi, C., Tschacher, W., & Kupper, Z. (2013). The assessment of mindfulness with self-report measures: Existing scales and open issues. *Mindfulness*, *4*(3), 191–202. http://dx.doi.org/10.1007/s12671-012-0110-9.

Bodner, T. E., & Langer, E. J. (2001). Individual differences in mindfulness: The Mindfulness/Mindlessness Scale. *Poster presented at the 13th Annual American Psychological Society Conference. Toronto, Canada*.

Bond, F. W., Hayes, S. C., Baer, R. A., Carpenter, K. M., Guenole, N., Orcutt, H. K., et al. , & Zettle, R. D.(2011). Preliminary psychometric properties of the Acceptance and Action Questionnaire-II: A revised measure of psychological inflexibility and experiential avoidance. *Behavior Therapy*, *42*(4), 676–688. http://dx.doi.org/10.1016/j.beth.2011.03.007.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. http://dx.doi.org/10.1037/0033-295×.111.4.1061.

Broadbent, D. E., Cooper, P. F., Fitzgerald, P., & Parkes, K. R. (1982). The Cognitive Failures Questionnaire (CFQ) and its correlates. *The British Journal of Clinical Psychology*, *21*(Pt 1), 1–16. http://dx.doi.org/10.1111/j.2044-8260.1982.tb01421.x.

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. http://dx.doi.org/10.1037/0022–3514.84.4.822.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306–307. http://dx.doi.org/10.1207/s15327752jpa4803_13.

Cardaciotto, L., Herbert, J. D., Forman, E. M., Moitra, E., & Farrow, V. (2008). The assessment of present-moment awareness and acceptance: The Philadelphia Mindfulness Scale. *Assessment*, *15*(2), 204–223. http://dx.doi.org/10.1177/1073191107311467.

Carver, C. S., Voie, L. L., Kuhl, J., & Ganellen, R. J. (1988). Cognitive concomitants of depression: A further examination of the roles of generalization, high standards, and self-criticism. *Journal of Social and Clinical Psychology*, *7*, 350–365. http://dx.doi.org/10.1521/jscp.1988.7.4.350.

Cash, M., & Whittingham, K. (2010). What facets of mindfulness contribute to psychological well-being and depressive, anxious, and stress-related symptomatology? *Mindfulness*, *1*(3), 177–182. http://dx.doi.org/10.1007/s12671-010-0023-4.

Chadwick, P., Hember, M., Symes, J., Peters, E., Kuipers, E., & Dagnan, D. (2008). Responding mindfully to unpleasant thoughts and images: Reliability and validity of the Southampton mindfulness questionnaire (SMQ). *British Journal of Clinical Psychology*, *47*(Pt 4), 451–455. http://dx.doi.org/10.1348/014466508x314891.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J. -P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, *80*(1), 219–251. http://dx.doi.org/10.1111/j.1467-6494.2011.00739.x.

Christopher, M. S., & Gilbert, B. D. (2009). Incremental validity of components of mindfulness in the prediction of satisfaction with life and depression. *Current Psychology*, *29*(1), 10–23. http://dx.doi.org/10.1007/s12144-009-9067-9.

Consedine, N. S., & Butler, H. F. (2013). Mindfulness, health symptoms and healthcare utilization: Active facets and possible affective mediators. *Psychology, Health & Medicine*, *19*(4), 392–401. http://dx.doi.org/10.1080/13548506.2013.824596.

Conte, H. R., Ratto, R., & Karasu, T. B. (1996). The Psychological Mindedness Scale: Factor structure and relationship to outcome of psychotherapy. *Journal of Psychotherapy Practice and Research*, *5*(3), 250–259.

Costa, P. T., & McCrae, R. R. (1998). Six approaches to the explication of facet-level traits: Examples from conscientiousness. *European Journal of Personality*, *12*(2), 117–134. http://dx.doi.org/10.1002/(SICI)1099-0984.

Curtiss, J., & Klemanski, D. H. (2014). Factor analysis of the Five Facet Mindfulness Questionnaire in a heterogeneous clinical sample. *Journal of Psychopathology and Behavioral Assessment*, *36*, 683–694. http://dx.doi.org/10.1007/s10862-014-9429-y.

Davis, K. M., Lau, M. A., & Cairns, D. R. (2009). Development and preliminary validation of a trait version of the Toronto Mindfulness Scale. *Journal of Cognitive Psychotherapy*, *23*(3), 185–197. http://dx.doi.org/10.1891/0889–8391.23.3.185.

Derksen, S., & Keselman, H. J. (2011). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, *45*(2), 265–282. http://dx.doi.org/10.1111/j.2044-8317.1992.tb00992.x.

Epstein, S. (1984). A procedural note on the measurement of broad dispositions. *Journal of Personality*, *52*(4), 318–325. http://dx.doi.org/10.1111/j.1467-6494.1984.tb00354.x.

Feldman, G., & Hayes, A. (2005). Preparing for problems: A measure of mental anticipatory processes. *Journal of Research in Personality*, *39*(5), 487–516. http://dx.doi.org/10.1016/j.jrp.2004.05.005.

Feldman, G., Hayes, A., Kumar, S., Greeson, J., & Laurenceau, J. -P. P. (2006). Mindfulness and emotion regulation: The development and initial validation of the Cognitive and Affective Mindfulness Scale-Revised (CAMS-R). *Journal of Psychopathology and Behavioral Assessment*, *29*(3), 177–190. http://dx.doi.org/10.1007/s10862-006-9035-8.

Fenigstein, A., Scheier, M. F., & Buss, A. H. (1975). Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology*, *43*(4), 522–527. http://dx.doi.org/10.1037/h0076760.

Giluk, T. L. (2009). Mindfulness, Big Five personality, and affect: A meta-analysis. *Personality and Individual Differences*, *47*(8), 805–811. http://dx.doi.org/10.1016/j.paid.2009.06.026.

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, *85*(2), 348–362. http://dx.doi.org/10.1037/0022–3514.85.2.348.

Gu, J., Strauss, C., Crane, C., Barnhofer, T., Karl, A., Cavanagh, K., & Kuyken, W. (2016). Examining the factor structure of the 39-item and 15-item versions of the Five Facet Mindfulness Questionnaire before and after mindfulness-based cognitive therapy for people with recurrent depression. *Psychological Assessment*, *28*(7), 791–802.

Hull, J. G., Lehn, D. A., & Tedlie, J. C. (1991). A general approach to testing multifaceted personality constructs. *Journal of Personality and Social Psychology*, *61*(6), 932–945. http://dx.doi.org/10.1037//0022-3514.61.6.932.

Jamieson, G. A. (2005). The Modified Tellegen Absorption Scale: A clearer window on the structure and meaning of absorption. *Australian Journal of Clinical and Experimental Hypnosis*, *33*(2), 119–139.

John, O. P., & Benet-Martinez, V. (2000). Measurement: Reliability, construct validation, and scale construction. *Handbook of research methods in social and personality psychology* (pp. 339–369). http://dx.doi.org/10.1017/CBO9780511996481.

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin, & O. P. John (Eds.), *Handbook of personality: Theory and research, vol. 2*. (pp. 102–138). New York, NY: Guilford Press.

Johnson, R. E., Rosen, C. C., & Djurdjevic, E. (2011). Assessing the impact of common method variance on higher order multidimensional constructs. *Journal of Applied Psychology*, *96*(4), 744–761. http://dx.doi.org/10.1037/a0021504.

Kashdan, T. B., Gallagher, M. W., Silvia, P. J., Winterstein, B. P., Breen, W. E., Terhar, D., & Steger, M. F. (2009). The Curiosity and Exploration Inventory-II: Development, factor structure, and psychometrics. *Journal of Research in Personality*, *43*(6), 987–998. http://dx.doi.org/10.1016/j.jrp.2009.04.011.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, *8*(3), 305–321. http://dx.doi.org/10.1037/1082-989×.8.3.305.

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. http://dx.doi.org/10.1016/j.paid.2008.10.009.

Lau, M. A., Bishop, S. R. S., Buis, T., Anderson, N. D., Carlson, L., Carmody, J., & Segal, Z. (2006). The Toronto mindfulness scale: Development and validation. *Journal of Clinical Psychology*, *62*(12), 1445–1467 (http://doi.org/10.1002/jclp).

Lennox, R. D., & Wolfe, R. N. (1984). Revision of the Self-Monitoring Scale. *Journal of Personality and Social Psychology*, *46*, 1349–1364. http://dx.doi.org/10.1037//0022-3514.46.6.1349.

Lyubomirsky, S., & Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research*, *46*(2), 137–155. http://dx.doi.org/10.1023/A:1,006,824,100,041.

Mayer, J. L., & Farmer, R. F. (2010). The development and psychometric evaluation of a new measure of dissociative activities. *Journal of Personality Assessment*, *80*(2), 185–196. http://dx.doi.org/10.1207/s15327752jpa8002_07.

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the Penn State Worry Questionnaire. *Behaviour Research and Therapy*, *28*(6), 487–495. http://dx.doi.org/10.1016/0005–7967(90)90135-6.

Morin, A. J. S., Boudrias, J. -S., Marsh, H. W., McInerney, D. M., Dagenais-Desmarais, V., Madore, I., & Litalien, D. (2016). Complementary variable- and person-centered approaches to the dimensionality of psychometric constructs: Application to psychological wellbeing at work. *Journal of Business and Psychology*. http://dx.doi.org/10.1007/s10869-016-9448-7.

Pace, V. L., & Brannick, M. T. (2010). How similar are personality scales of the "same" construct? A meta-analytic investigation. *Personality and Individual Differences*. http://dx.doi.org/10.1016/j.paid.2010.06.014.

Pedhazur, E. J. (1997). *Multiple-regression in behavioral research: Explanation and prediction* (3rd ed.). New York, NY: Harcourt Brace (http://doi.org/10.2307/2285468).

Pirson, M., Langer, E. J., Bodner, T., & Zilcha, S. (2012). The development and validation of the Langer Mindfulness Scale — Enabling a socio-cognitive perspective of mindfulness in organizational contexts. *SSRN Electronic Journal*, *1–54*. http://dx.doi.org/10.2139/ssrn.2158921.

Raes, F., Pommier, E., Neff, K. D., & Van Gucht, D. (2011). Construction and factorial validation of a short form of the Self-Compassion Scale. *Clinical Psychology & Psychotherapy*, *18*(3), 250–255. http://dx.doi.org/10.1002/cpp.702.

Raykov, T., & Marcoulides, G. A. (2016). On examining specificity in latent construct indicators. *Structural Equation Modeling: A Multidisciplinary Journal*, *00*, 1–11. http://dx.doi.org/10.1080/10705511.2016.1175947.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 53–61. http://dx.doi.org/10.1007/BF02289209.

Shill, M. A., & Lumley, M. A. (2002). The Psychological Mindedness Scale: Factor structure, convergent validity and gender in a non-psychiatric sample. *Psychology and Psychotherapy*, *75*(Pt 2), 131–150. http://dx.doi.org/10.1348/147608302169607.

Siegling, A. B., & Petrides, K. V. (2014). Measures of trait mindfulness: Convergent validity, shared dimensionality, and linkages to the Five-Factor Model. *Frontiers in Psychology*, *5*(October), 1–8. http://dx.doi.org/10.3389/fpsyg.2014.01164.

Siegling, A. B., & Petrides, K. V. (2016). Zeroing in on mindfulness facets: Similarities, validity, and dimensionality across three independent measures. *PLOS One*, *11*(4), e0153073.

Siegling, A. B., Petrides, K. V., & Martskvishvili, K. (2015). An examination of a new psychometric method for optimizing multi-faceted assessment instruments in the context of trait emotional intelligence. *European Journal of Personality*, *29*(1), 42–54. http://dx.doi.org/10.1002/per.1976.

Smith, G. T., & Zapolski, T. C. B. (2009). Construct validation of personality measures. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 81–88). Oxford, UK: Oxford University Press (http://doi.org/http://dx.doi.org/10.1093/oxfordhb/9780195366877.001.0001).

Smith, G. T., Fischer, S., & Fister, S. M. (2003). Incremental validity principles in test construction. *Psychological Assessment*, *15*(4), 467–477. http://dx.doi.org/10.1037/1040–3590.15.4.467.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson (http://doi.org/10.1037/022267).

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, *88*(3), 500–517. http://dx.doi.org/10.1037/0021–9010.88.3.500.

Trapnell, P. D., & Campbell, J. D. (1999). Private self-consciousness and the five-factor model of personality: Distinguishing rumination from reflection. *Journal of Personality and Social Psychology*, *76*(2), 284–304. http://dx.doi.org/10.1037//0022-3514.76.2.284.

Vujanovic, A. A., Bonn-Miller, M. O., Bernstein, A., McKee, L. G., & Zvolensky, M. J. (2010). Incremental validity of mindfulness skills in relation to emotional dysregulation among a young adult community sample. *Cognitive Behaviour Therapy*, *39*(3), 203–213. http://dx.doi.org/10.1080/16506070903441630.

Walach, H., Buchheld, N., Buttenmüller, V., Kleinknecht, N., & Schmidt, S. (2006). Measuring mindfulness—The Freiburg Mindfulness Inventory (FMI). *Personality and Individual Differences*, *40*(8), 1543–1555. http://dx.doi.org/10.1016/j.paid.2005.11.025.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect — The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. http://dx.doi.org/10.1037/0022–3514.54.6.1063.

Wegner, D. M., & Zanakos, S. (1994). Chronic thought suppression. *Journal of Personality*, *62*(4), 616–640. http://dx.doi.org/10.1111/j.1467-6494.1994.tb00311.x.

Williams, M. J., Dalgleish, T., Karl, A., & Kuyken, W. (2014). Examining the factor structures of the Five Facet Mindfulness Questionnaire and the Self-Compassion Scale. *Psychological Assessment*, *26*(2), 407–418. http://dx.doi.org/10.1037/a0035566.

Ziegler, M., & Bäckström, M. (2016). 50 facets of a trait — 50 ways to mess up? *European Journal of Psychological Assessment*, *32*(2), 105–110. http://dx.doi.org/10.1027/1015–5759/a000372.

Ziegler, M., Booth, T., & Bensch, D. (2013). Getting entangled in the nomological net: Thoughts on validity and conceptual overlap. *European Journal of Psychological Assessment*, *29*(3), 157–161. http://dx.doi.org/10.1027/1015–5759/a000173.

Ziegler, M., Danay, E., Schoelmreich, F., & Buehner, M. (2010). Predicting academic success with the Big 5 rated from different points of view: Self-rated, other rated and faked. *European Journal of Personality*, *24*(4), 341–355. http://dx.doi.org/10.1002/per.