Introduction

Work on mindfulness and its applications is booming, in considerable part due to the success of Mindfulness Based Stress Reduction (Kabat-Zinn, 2013). In this context, the term mindfulness refers to (something like) “the awareness that emerges through paying attention on purpose, in the present moment, and non-judgmentally to the unfolding of experience moment by moment.” (Kabat-Zinn, 2003). A number of meta-analyses confirm the usefulness of mindfulness in a range of issues from treatment of psychosis (Cramer, Lauche, Haller, Langhorst & Dobos, 2016), other psychological problems (Khoury et al., 2013), or symptoms of physical diseases (Rinske, Paula, van Busschbach, Herbert, Gregory & Hunink, 2015) to work life (Mesmer-Magnus, Manapragada, Viswesvaran & Allen, 2017) and sports performance (Bühlmayer, Birrer, Röthlin, Faude & Donath, 2017), among others.

It is easy to believe that being able to pay attention and to act on purpose are helpful: it may seem less clear that being in the present moment is necessarily beneficial (compare the idea of future-oriented prospection being central to human flourishing: Seligman, Railton, Baumeister & Sripada, 2013), even less so why a non-judging attitude might be desirable. Indeed, other uses of the term “mindfulness” like Ellen Langer's (Ie, Ngnoumen & Langer, 2014) seem unconnected to the idea of non-judging of experience. Even more to the point, a non-evaluative stance towards one’s experience (including thoughts and emotions) seems to be in conflict with the emphasis in cognitive therapy on challenging dysfunctional or irrational thoughts. For example, Rational Emotive
Behavioural Therapy claims that

\[
you \text{ largely bring on your own emotional disturbances by choosing, both consciously and unconsciously, to think irrationally, to create unhealthy negative feelings, and to act in self-defeating ways [...]} \text{ therefore, you can choose to change your thinking, feelings and behaviors to undisturb yourself. (Ellis, 1999, p. 175, emphasis in original).}
\]

Given the Buddhist roots of Kabat-Zinn’s notion of mindfulness, it is also remarkable that most versions of traditional Buddhism do advise to develop \textit{samma-diṭṭhi} (usually translated “correct view”) and abandon \textit{micchā-diṭṭhi} (“wrong view”), which seems difficult without evaluating one’s thoughts.

Thus, there seems to be a need for a comprehensive study of the effects of different facets of mindfulness, both as a possible contribution to clear up conceptual confusions, and as an aid in designing effective and efficient interventions. Most existing meta-analyses do not study the differential effects of all mindfulness facets simultaneously. In addition, they tend to study (almost always relatively short-term) interventions, which has a number of drawbacks: first, mindfulness is usually considered to be trait-like, therefore slow to change, which makes short-term interventions a somewhat non-obvious setting for its study; second, interventions tend to have other ingredients (e.g., psycho-education and group effects in the case of MBSR) which might colour the conclusions. Consistent with this, Eberth & Sedlmeier (2012) wrote in the conclusion of their meta-analysis of mindfulness meditation effects that they found large differences in effect sizes for MBSR versus meditation and that “[t]his raises the question if some effect sizes found for MBSR might be partly inflated by effects that are not attributable to its mindfulness meditation component.” Also consistent with this, Rau and Williams (2016) argued for a distinction between dispositional and cultivated mindfulness. These considerations imply that it is useful to study correlates of mindfulness facets cross-sectionally, rather than only their changes during
interventions.

The most popular multi-facet self report measure of mindfulness is the Five Facet Mindfulness Questionnaire (FFMQ), Baer, Smith, Hopkins, Krietemeyer and Toney (2006), which was derived from a comprehensive study of self-report measures of mindfulness in use at that time. Psychometric analysis led to a five factor model, with facets labelled as: Observing, Describing, Acting-with-awareness, Non-judging of and Non-reacting to inner experience (these will be usually be abbreviated as Obs, Des, Act, NJ and NR, respectively, in the present paper). The observe facet is known to behave differently among participants that meditate versus those that do not (Baer et al., 2008).

For the above reasons, the present study performs a meta-analysis of correlates of facets of mindfulness as measured by the FFMQ. Besides being interesting in itself, studying correlations has the additional benefit of leading to a large sample size, which is important given recent concerns regarding possible publication bias in available studies of the effects of mindfulness (Coronado-Montoya, Levis, Kwakkenbos, Steele, Turner & Thombs, 2016). Given another important recent concern, namely research standards in psychology (e.g., Open Science Collaboration (2015), Bones (2012), Brown, Sokal and Friedman (2013), Ledgerwood (2016)), another aspect that is taken into account is adherence to (two aspects of) good research practice (Finkel, Eastwick & Reis, 2015): preregistration (Moore, 2016) and transparency; the latter will be proxied by availability (whether the underlying data used and/or the paper are openly available) and readability (whether the abstract is structured or not). Partial correlations with outcomes will also be studied as the unique contribution of facets seem interesting both in itself (I argued above that for some facets positive effects might seem counter-intuitive, thus it is important to show to what extent they uniquely contribute to positive outcomes) and in the design of mindfulness trainings and interventions.
(helping to answer the question which facets should be preferentially strengthened).

Given that the aim of mindfulness practice is not only therapeutic, but to contribute to "flourishing on this planet [...] for the benefit of all sentient beings and our world" (Kabat-Zinn, 2011), the primary emphasis in this study will be on positive psychology measures. As there is a widely assumed distinction between well-being and eudaimonia – with meaning in life being a particularly distinctive feature of eudaimonia (Disabato et al., 2016) – at the preregistration stage two measures were designated as primary outcomes for this meta-analysis: the Satisfaction With Life Scale (SWLS, e.g., Diener and Griffin (1985), Pavot and Diener (2008)) as a widely used measure of well-being, and the Purpose in Life Scale (PiL) as the most widely used measure of meaning (Bronk, 2014). Secondary outcomes are all other measures, individually if at least four data points are available for meta-analysis, otherwise grouped as described in the method section below.

**Method**

Study registration: This study was registered with the PROSPERO International prospective register of systematic reviews (Booth et al., 2012):

http://www.crd.york.ac.uk/PROSPERO/display_record.asp?ID=CRD42016041863, in accordance with the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA, Moher, Liberati, Tetzlaff and Altman (2009); Moher et al. (2015)).

**Database Search**

On 15 July 2016, the following files were downloaded from the website of the American
Mindfulness Research Association (https://goamra.org/resources/mindfo-database/):

Mar2016_AMRA_database.zip Apr2016_AMRA_database.zip  May2016_AMRA_library.zip
Jan2016_AMRA_database-.zip. On the same date, the following databases were searched for the
terms "FFMQ" and "Five facet* mindfulness": PubMed, PsycINFO, ResearchGate. The search
resulted in 1229 potentially relevant papers, 708 after removing duplicates.

**Study Selection**

Included in the meta-analysis were all studies which a) reported correlations with all five
mindfulness facets for any outcome that was not another mindfulness measure, where b) both the
publication and the questionnaire used were in one of the following languages: English, French,
German, Spanish. Excluded were studies which reported only changes (not levels) of mindfulness
facets, and those using a short form of the FFMQ (Park, Reilly-Spong and Gross (2013), in their
survey of mindfulness measures, also excluded shortened or modified versions).

A controversial issue in meta-analysis is whether unpublished "grey" literature should be included.
Even though most authors seem to advocate including as many studies as possible, there are good
reasons against it (e.g., Ferguson and Brannick (2012) ), in particular if few studies are pre-
registered (as is the case here) and if the number of published effect sizes is large (again the case
here) so that one can use statistical techniques (as for example discussed in (McShane, Böckenholt
& Hansen, 2016) or Jin, Zhou and He (2015) ) to adjust for publication bias. According to the
comprehensive study of bias in science performed by Fanelli, Costas and Ioannidis (2017) , studies
not published in peer-reviewed journals tend to underestimate effects. Therefore, the present meta-
analysis includes only published studies.
Papers were checked for relevance by scanning the PDF of the article where available, else abstracts were examined whether they mentioned FFMQ facets (resulting in four requests to authors to send the paper, one of which was fulfilled). For book chapters, Google books was searched if neither the book nor a PDF were available.

This procedure resulted in the selection of 117 studies in 97 publications.

Data Extraction

For each paper the following were collected: publication type, authors, year of publication, number of relevant studies, number of samples used, number of measures used; whether the paper was open access or available via ResearchGate or PubMed, whether a link was given to access the underlying data (this was never the case), whether the abstract was structured and whether the study reported was pre-registered.

For each sample, the following were collected, where available: number of participants, whether the sample was clinical, non-clinical or mixed, whether the participants were meditators, non-meditators or mixed, the participants' occupation (students; academics; other university community – e.g., administrators; health professionals, other professionals; military; community sample; internet sample), proportion of males in the sample (between 0 and 1), mean age, standard deviation of age, world region and country in which the study was conducted, correlation table of the FFMQ
For each outcome measure, the correlations with the FFMQ facets together with the type of measure and directionality (see below) were collected. In addition, the following were collected, where available: Mean and standard deviation and Cronbach alpha. As a partial test of data integrity, a simple version of the GRIM test (Brown & Heathers, 2017) was performed: no irregularities were found. This selection procedure resulted in 565 effect sized (correlations) per facet.

Effect sizes were grouped as follows: 1) Well-being/happiness (e.g., the satisfaction With life Scale: SWLS), 2) Eudaimonia (e.g., Psychological Well-Being: PWB) 3) Clinical/health-related (e.g., Depression And Anxiety Scales: DASS) 4) Physiology/body/movement (e.g., 6 Minute Walk test) 5) Judgement and rationality (e.g., Monetary Choice Questionnaire), 6) Social behaviour (e.g., relations subscale of PWB), 8) unhelpful (outcomes that might not be in themselves pathological but conceivably might have negative consequences, e.g., DERS = Difficulties in Emotion Regulation Scale), 9) helpful (outcomes that might not be in themselves desirable but may well have positive consequences, e.g., Health-Promoting Lifestyle Profile). 0) other. This last group of effect sizes was excluded from analysis, since aggregating correlations requires directionality in the outcomes (a higher score must be more desirable), which the measures in this group do not satisfy; for the same reason, signs of groups 3 and 8 had to be reversed to make directionalities consistent, as well as in a few individual studies (e.g., Curtiss and Klemanski (2014) reported reversed scores for the AAQ-II scale). In addition, the body mass index, though obviously belonging in group 4,
was excluded in non-overweight samples, because of lack of clear directionality (“less is better” is far from clear in non-clinical samples: being severely underweight is no more desirable than being severely overweight – just think of anorexia).

The final sample contained 528 effect sizes per FFMQ facet, plus 309 effect sizes for the FFMQ total score, altogether 2859 usable effect sizes. There were three unusually large samples: Camilleri, Méjean, Bellisle, Hercberg and Péneau (2015) had two samples with 49228 and 14400 participants, respectively, Jones, Mist, Casselberry, Ali and Christopher (2015) provided one sample with 4986 participants. Excluding those three, sample sizes ranged from 20 to 1210 with mean 249.3 and median 179.

Software

Calculations were performed using the free software environment for statistical computing and graphics R (R Core Team, 2012), version 3.4.0 under Windows 10, employing the packages 'Matrix', 'weightr', 'meta' and 'corpcor', see also Schwarzer, Carpenter and Rücker (2015).

Risk of Bias, Quality Assessment

Since the review covers only correlational studies, no risk of bias assessment at the individual study level was performed, since the usual measures of bias risk (e.g., form of randomization) are geared towards intervention studies and in the most part can not be applied here.

Concerning publication bias, it was originally intended to test for its presence using funnel plots and
trim-and-fill. Nevertheless, these methods are not well-behaved under heterogeneity (Terrin, Schmid, Lau & Olkin, 2003). In view of the fact that heterogeneity in the present dataset turned out to be unexpectedly large in most cases (see below), and noting that the recently introduced techniques of p-curve and p-uniform also seem problematic already under moderate heterogeneity (van Aert, Wicherts & van Assen, 2016; Carter, Schönbrodt, Gervais & Hilgard, 20xx), possible consequences of publication bias are assessed in this study using the three-variable selection method of Vevea and Hedges (1995) and Vevea and Woods (2005), as implemented in Coburn & Vevea (2016) based on (R Core Team, 2012): we report effect estimates under different probabilities for a statistically insignificant (two-sided p-value less than 5%) getting published. This allows us to quantify the possible influence of publication bias on mindfulness research, which (in the context of intervention studies) was noted in Coronado-Montoya et al. (2016).

Quality of studies was be assessed in terms of transparency: low, medium (paper is published open access, or a version of the paper can be downloaded, e.g., from ResearchGate), high (underlying data available for download). In addition, the quality of the abstract was assessed by considering a structured abstract as higher quality (more user friendly, since it makes it easier to grasp crucial information quickly).

**Data Synthesis**

**Planned studies**

Pre-registration stipulated random effects meta-analysis, with tau-squared and I-squared as measures of heterogeneity (with I^2 < 0.3 considered mild and I^2 > 0.5 indicating severe heterogeneity, following Higgins and Thompson (2002)), for the following outcomes: SWLS
(Satisfaction With Life Scale) as planned; in addition all measures (excluding sub-scales) with at least four data points, these were the following: PWB (Psychological Well Being), PSWQ (Penn State Worry Questionnaire), DASS (Depression and Anxiety Scales), BDI_II (Beck Depression Inventory), PANAS_NA (Positive And Negative Affect Scale - negative affect), PSS (Perceived Stress Scale), AAQ_II (Acceptance and Action Questionnaire).

Where possible, separate subgroup analyses were performed for groups described as: clinical vs. non-clinical vs. mixed, meditators vs. non-meditators vs. mixed, and for self-report vs. objective measures; in addition, meta-regressions were performed with publication year as the independent variable. The possible impact of publication bias was assessed as explained above.

Almost all studies measured several outcomes for each participant, but hardly any designated primary outcome measures, and only 41 of 117 studies reported covariances between the different outcomes or the data needed to calculate these, which made it impossible to properly aggregate the effect sizes. This problem was dealt with in the following way: First, all outcomes were included where appropriate, implicitly making the (unrealistic) assumption of zero correlation between the measures; as a robustness check, for each such case of multiple measurement, one random measure was selected for inclusion (subject to the following constraints: full scales were preferred to subscales, and preference was given to types of outcomes as follows: well-being > eudaimonia > clinical > helpful > unhelpful > other).

Additional studies

In addition to the planned studies, efforts were made to explore the reasons for the surprisingly high heterogeneity in effect sizes, even where single (rather than grouped) outcome measures were used. Specifically, where individual studies seemed outliers the meta-analysis was rerun with these
studies removed. In addition, subgroup analyses based on region (in which part of the world the study was conducted) and professional background of the participants (students, community samples, academics, etc.) and meta-regressions based on mean age, standard deviation of age, and shares of males among the participants were conducted. I also calculated the correlations of outcomes with differences of facets and performed meta-analyses on these (results are available from the author).

**Results**

Note: Given the need to adjust for multiple testing, p-values are reported to four digits throughout this paper. Effects will be reported to two digits to make reading large tables easier.

**Primary Outcome: Satisfaction With Life Scale (SWLS)**

Designated primary outcomes for this meta-analysis were the Satisfaction With Life Scale and the Purpose in Life questionnaire. Despite the Purpose in Life questionnaire being the most used measure of purpose, not a single study included in the present sample used it; in fact, no measure of meaning whatsoever is. Hence, SWLS is the sole primary outcome used. Four studies used the Satisfaction With Life Scale (Diener & Griffin, 1985), a widely used measure of well-being; for three of them it was possible to calculate partial correlations. Aggregate results from the meta-analyses both for the full sample, as well as restricted to those studies with partial correlations available, are presented in Table 1, the forest plots for the full sample are shown in Figure 2.
For both the full and restricted sample, Acting-with-awareness had the highest correlation with life satisfaction (the estimates for the other four facets are outside the 95% confidence interval for Act), with Describing, Non-judging and Non-reacting estimated to be close to each other, and Observing having a considerably weaker correlation (estimates for the other facets being far away from the confidence interval for Obs). The latter is unsurprising since none of the samples consisted of meditators (though one was mixed and did indeed reported the highest correlation for Obs, which nevertheless was at 0.24 still lower than the other estimated effects).

Table 1 revealed considerable heterogeneity and wide confidence intervals for NR and in particular for NJ. Inspection of the individual studies showed that the heterogeneity in NJ stemmed from the exceptionally small effect size reported in Lara, Herrero, Blanco-Donoso & Chavez (2015). Indeed, eliminating that paper from the sample reduces $I^2$ for both NJ and Des to zero (but it increases the already substantial heterogeneity for NR to 0.88). Unsurprisingly, the estimated correlation of NJ is increased (to 0.37), the other correlations are then estimated to be 0.36 (Act), 0.34 (Des), 0.29 (NR) and 0.14 (Obs).

The number of studies was too small to estimate the possible impact of publication bias or for subgroup comparisons. Meta-regressions found little evidence against the null hypothesis of independence from the regressor variables (of twenty p-values, only three were smaller than 0.05; none was smaller than value of 0.0025 which results from Bonferroni-adjustment for 20-fold testing).
Individual Secondary Outcomes

This section reports the results for all measures for which at least four effect sizes were available, except SWLS (reported above) and subscales of a measure that is reported here (these concerns the three sub-scales of DASS), resulting in the following collection of outcomes: PWB ($k=6$ effect sizes with $N=974$ participants), PSWQ ($k=7$, $N=3302$), DASS ($k=6$, $N=1252$), BDI II ($k=4$, $N=519$), PANAS.NA ($k=4$, $N=773$), PSS ($k=4$, $N=1464$), AAQ II ($k=4$, $N=1319$). Note that all measures are used in such a way that higher values represent more desirable outcomes, so that all except PWB are inverted from their usual direction. Table 2 together with Figure 3 summarize the results:

The Observing facet shows small effect sizes for all well-being and eudaimonic outcomes, and zero-to-negative effects for the other categories (the correlation with PSWQ in particular is statistically significantly negative with $r = -0.07$ and $p = 0.0017$). In any case, the effects are much smaller than those of the other facets – in almost all cases the 95% confidence intervals are disjoint. Overall, there seems to be a weak tendency for NJ to have the largest effect, followed by Act, NR and then Des, but confirmation or dis-confirmation of this requires more repeated use of these measures to arrive at more precise estimates. The last statement applies even more clearly for partial correlations, for which the estimates are too imprecise to draw conclusions. In addition to the data reported in the table, heterogeneity is high in many cases, with 21 of 40 $I^2$ statistics being above 0.5.

For PWB, correlations were estimated to be large for NJ ($r = 0.52$, 95% confidence interval =
medium for Act, NR and Des (between 0.45 and 0.4) and small for Obs ($r = 0.23$, CI=[0.0475, 0.3902]). All estimates except that for Obs were highly robust to assumptions about publication bias. Heterogeneity was a substantial concern for Obs ($I^2=0.87$), borderline severe for Act ($I^2=0.52$), and negligible for the other three facets. For Observing it is noticeable that three studies reported medium effects ($r$ between 0.3 and 0.45), whereas the other three studies reported very small effects (between -0.02 and 0.08). Interestingly, this is not entirely due to the difference between meditators and non-meditators: the samples reporting medium effects were one meditator, one mixed and one non-meditator sample (the other three samples consisted of non-meditators).

Concerning possible moderators, meta-regressions found again no evidence against the null hypothesis of independence from the regressor variables (all $p > 0.1$), while the number of studies was too small for subgroup comparisons.

As noted above, PSWQ showed a small but statistically significant ($p = 0.0017$) negative effect of Obs, and highly significant (all $p < 0.0001$) small-to-medium effects of the other facets.

Heterogeneity was elevated but not extreme for Act, NJ and NR. Again, meta-regressions found little evidence against the null hypothesis of independence from the regressor variables (only four $p$-values were smaller than 0.05; none was smaller than the Bonferroni-adjusted value of 0.0025). The number of studies was too small for subgroup comparisons, nor could the model for publication bias be estimated.

Results for DASS were similar to those for PSWQ, except that the negative effect of Obs was statistically insignificant and the effect of NJ was increased slightly. Furthermore, for DASS the model for publication bias could be estimated: Even with the assumption of very severe publication bias (90% of statistically insignificant results going unpublished), the estimated correlations are only slightly reduced to -0.1073 for Obs, 0.1609 for Des, 0.3494, for Act, 0.4369 for NJ, and 0.2583.
for NR.

The pattern of results is generally similar for BDI II and PANAS.NA. For the PSS it is interesting to note that Obs shows correlations with the outcome measure ranging from -0.21 to +0.16, with both of the extreme values estimated in non-meditating student samples in the USA. Finally, AAQ II is the only outcome measure besides PWB were a large correlation was estimated (for NJ, $r = 0.55$).

### Grouped outcomes: overview

The grouping of measures that were not used often enough to allow individual meta-analysis has a substantial subjective component. Probably the most subjective choice is that of assigning desirable outcomes to wellbeing, eudaimonia, or what is here called helpful on the one hand (including the individually reported SWLS and PWB), and undesirable ones to clinical versus unhelpful on the other (again including the individually reported measures). Therefore the presentation will focus on results for desirable outcomes grouped together, and undesirable outcomes grouped together, plus a few remarks on the results related to body or social outcomes (which are based on much smaller samples). Results for subsets of measures (e.g., Other_Eudaimonia := only eudaimonia and excluding PWB) can serve as a robustness check.

<table>
<thead>
<tr>
<th>Table 3 about here</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 4 about here</td>
</tr>
</tbody>
</table>

The grouped results seem to confirm the result for corresponding individual measures as Obs has only small correlations with positive outcomes, and no effect at all on undesirable outcomes. Most other correlations are highly statistically significant and of practically meaningful magnitude, even
unique contributions assessed by partial correlations range up to 0.22 with most \( p \)-values <0.0001.

## Desirable outcomes

Combining the groups “well-being”, “eudaimonia” and “helpful” (including SWLS and PWB) led to 328 effect sizes and the following aggregate results:

| Table 4 about here |

Obs showed a small effect (\( r = 0.16 \)), the other four facets medium effects (between 0.26 and 0.29) with confidence intervals disjoint from that for Obs. Nevertheless, heterogeneity is a substantial concern, given that all \( I^2 \) are greater than 0.85. Table 4 suggests that the effect of Des is slightly larger than that of Act, NJ and NR; meta-analysis of the effect of differences of facets (where these could be calculated) showed that this is indeed statistically significant (\( p = 0.0219 \)).

Interestingly, partial correlations revealed that the unique contribution of Des is much smaller than of Act and NJ (again with disjoint confidence intervals: \([0.05, 0.11]\) for Des versus \([0.16,0.21]\) for Act and \([0.14, 0.21]\) for NJ).

Results seem robust to reasonable levels of publication bias: If 90% of studies with a statistically insignificant result go unpublished, the estimated effects would still be at least 0.19 for all facets except Obs.

| Table 5 about here |

Regressions of effect sizes on possible moderators show some evidence of moderating influences
by the tested variables, nevertheless, it is difficult to discern a pattern: Even after Bonferroni-
correction for twentyfold testing, the effect of NJ seems to decrease for later publications. NR might
also be negatively related to publication year, though the \( p \)-value is slightly less convincing, all
other coefficients for publication year are also negative, but with \( p \)-values far above thresholds for
statistical significance. Similarly, the coefficients for mean age and proportion of males are all
negative, with the following being significant or close to significant after Bonferroni correction: the
effect of Des decreases with mean age of the sample, that of Obs and NR seems to decrease with the
proportion of males among the participants.

Table 6 presents the results for clinical versus non-clinical participants, and for meditators versus
non-meditators. Among 27 non-clinical compared to 7 clinical samples (providing 125 and 17 effect
sizes, respectively) all facets show a substantially lower effect in clinical samples, with a \( p \)-value <
0.001 for Observing and Describing, and \( p < 0.1 \) for the other facets. 18 samples consisted of non-
meditators, 3 of meditators and 17 were mixed. Interestingly, for all facets the highest effect was
estimated in the mixed groups, this being statistically significant at the 0.05 level for your of the
five facets, and for two still after correction for multiple testing.

There is also strong statistical evidence for dependence of effects on occupation, in particular, NJ
and NR showed the largest effects in the samples consisting of academics, and very small effects
among the (few) samples consisting of professionals, as Table 7 reveals:

Table 7 about here

Region and country of the study may also matter. For NR in particular, there is statistical evidence
for variation across regions: Obs, NJ and NR exhibited the largest effects in Europe, Des and Act in
Asia. For all facets, the lowest effect estimates came from the three Latin American studies included (whereas studies done in Spain follow - and partly seem to cause - the European pattern of large effects for Obs, NJ and NR). Nevertheless, these results should be interpreted with caution given the small number of studies outside North America. Similarly, the two studies done in the German language were suggestive that the language of the questionnaire might matter as they exhibited by far the largest effects for all facets (except Obs, where they show a very low effect). Details are available from the author.

Interestingly, for all facets except Obs and possibly NR, reported effects are smaller in papers that have at least some version easily available on the internet, compared to those that are not (Table 8):

Table 8 about here

Undesirable outcomes

Results for the combined groups of clinical and unhelpful outcomes (144 effect sizes per facet) are as follows:

Table 9 about here

In contrast to the situation for desirable outcomes, the meta-analysis now showed that Obs had, if anything, a small negative effect, Des and NR had small and similar effects, Act and NJ had similar and medium sized effects. (Nevertheless, meta-analysis of the differences showed that the effects are also significantly different between NJ and Act ($\Delta r = 0.02$, $p = 0.0078$), and between Des and NR ($\Delta r = 0.03$, $p < 0.0001$)). The latter four facets also exhibited small partial correlations ranging between $r = 0.10$ for Des and 0.16 for NJ, with the confidence intervals for Des and NJ disjoint.
Heterogeneity is a substantial concern, given that all $I^2$ are greater than 0.75.

The results again seem quite robust with respect to publication bias. Even assuming extreme bias, the estimated effect sizes for Acting and Non-judging remain above 0.2:

Table 10 about here

In contrast to the case of desirable outcomes, meta-regressions resulted in coefficients that were too small to be practically meaningful even in the few cases where the associated $p$-values are below 0.05. Grouping into clinical, non-clinical and mixed samples resulted in considerable evidence against equality of effects ($p < 0.0001$ for three facets), but this seems to be driven mostly by the one mixed study, with the estimates for the purely clinical and purely non-clinical samples quite similar. Similarly, and partly in line with the results for desirable outcomes, the strongest effects were found for the three mixed (meditating and non-meditating) samples for Des, Act and NJ, whereas the strongest effects for Obs and NR were among meditators (though the $p$-value for overall difference in subgroups effects was statistically insignificant in the latter case).

Dependence on occupation for undesirable outcomes is attenuated compared to the case of desirable outcomes (though still statistically significant: $p = 0.0220$ for Act, $p = 0.0271$ for NJ and $p < 0.0001$ for the other facets). In particular, the extreme results for Academic and Professional samples are not replicated here. Also, concerning the region in which research is conducted, the extreme results above do not hold here. Only for NR is there evidence for differences between regions ($p = 0.0042$), again with Europe showing the strongest and Latin America the weakest effect. For both grouping according to occupation and according to region each group contained at least five studies, thus sample size should be less of a concern than for the corresponding results for desirable outcomes. Concerning questionnaire language, there is again some evidence that it matters
for the outcome, in this case for NJ ($p = 0.0010$). Nevertheless, direct comparison to the results for desirable outcomes is not possible since that sample contained German but no French studies, whereas here the case is reversed.

Interestingly, the availability of the publication now showed little relationship to the estimated effects, with only Act showing some evidence of being lower in openly accessible studies ($r = 0.30$ versus $0.35$, $p = 0.0302$). Finally, the choice between subjective and objective outcome measures matters: Estimates are much smaller for objective measures for Act, NJ, NR (all $p < 0.0001$). Surprisingly, the estimate for Obs for objective measures is larger ($r = 0.16$) than for subjective ones ($r = -0.02$) with a $p$-value of 0.0156.

**Further categories**

The number of effect sizes in the categories of physiological outcomes (labelled “body”) and social outcomes were much smaller, thus I report only the main results briefly: The effects in the “body” category turned out to be very small and in three of the five cases statistically insignificant, the larger ones were for Act ($r = 0.1110$, $p = 0.0354$) and NJ ($r = 0.0988$, $p = 0.0060$). Heterogeneity was here less severe than with other groupings. Remarkably, some partial correlations were higher than zero-order correlations, this being most pronounced for NR (Act: $r = 0.18$, $p < 0.0001$, NJ: $r = 0.21$, $p = 0.0076$, NR: $r = 0.17$, $p = 0.0010$). Effect estimates for social outcomes were small (ranging from $r = 0.1$ for Obs to $r = 0.21$ for NJ, all $p < 0.0003$), with the largest partial correlation shown by NR ($r = 0.16$, $p < 0.0001$).
Discussion

This paper contributes to the “exploration [of links] between specific DM facets and psychological health” (Tomlinson, Yousaf, Vittersø & Jones, 2018), and between those facets and well-being, by meta-analytically studying the strength, statistical significance, and robustness of correlations; by evaluating possible moderators; and by estimating the effects of possible publication bias. Across conditions, Describe, Act-with-awareness, Non-judge and (to a lesser extent) Non-react are moderately and non-redundantly correlated with outcomes. Nevertheless, precision of estimates is negatively impacted by heterogeneity and a lack of repeated use of measures:

Strength and robustness of correlations

Under the widely used rules of thumb to classify correlations according to strength into small ($r$ between 0.1 and 0.3), medium ($r$ between 0.3 and 0.5) and large ($r > 0.5$) correlations, only Non-judging exhibited large correlations (with AAQ-2 and PWB). Medium sized correlations above 0.4 were revealed for Describing with "Other eudaimonia" (due to strong correlations with certain subscales of PWB reported in two papers: Bravo, Boothe and Pearson (2016) and Bergin and Pakenham (2016)) and with PWB itself, for Acting-with-awareness with PWB, AAQ-2, and the (reversed) scales BDI-2, PANAS.NA and PSS, for Non-judging (with reversed DASS, PANAS.NA, PSS, PSWQ) and for Non-reacting with PWB and AAQ-2. In addition, there was a considerable number of correlations between 0.3 and 0.4 for all facets except Observing.

These effects seemed quite robust to taking subsamples, and to reasonable rates of publication bias (if, say, at least 20% of statistically insignificant results get published). Nevertheless, one should
always keep in mind that the guidelines for classifying correlation strength are only very rough rules of thumb and the actual importance of a given effect size can vary widely depending on circumstances.

**Facets**

The Observe facet showed the expected dependence on meditation experience that was found in previous publications. Nevertheless, we also saw that the Observe facet has small, but non-trivial, correlations with measures of desirable outcomes even for non-meditators. Therefore, it might be premature to simply drop this scale from the FFMQ when applied to non-meditators, as some authors advocate (e.g., Duan (2016)) and should be interesting to study correlations of positive outcomes with the factors found in Rudkin, Medvedev and Siegert (2018).

The Describe facet had the highest zero-correlation with the category of all desirable outcomes, which nevertheless did not hold for partial correlations. In addition, it had correlations >0.25 with the other categories of desirable outcomes studied, and correlations mostly around 0.2 with (reversed) undesirable outcomes. In particular, it showed the relatively largest (but still small, and imprecisely measured) correlation with outcomes classified as "social" (measures like the Compassionate Love Scale or the Agreeableness and Extroversion sub-scales of the Big Five), which may not be all that surprising since a high score on Describe could indicate a general ability to communicate, which in turn may be helpful in social contexts independently of mindfulness (for example, one item on this subscale of the FFMQ reads "I can easily put my beliefs, opinions, and expectations into words").
Acting-with-awareness generally tended to be among the facets showing the highest correlation with outcomes. Its estimated correlation with the primary measure in this meta-analysis, SWLS, and with the reversed depression scale, were the highest among the five facets, but these estimates were too imprecise (because of small samples, heterogeneity, and at least for SWLS because of an apparent outlier in the data) to allow firm conclusions. In any case, a correlation of Act with absence of depression would not be surprising since the items on the Act scale pertain to the ability to concentrate and actually do something; hence, depression might conceivably lead to a low score on Act (rather than this mindfulness facet relieving depression).

As noted above, Non-judging distinguished itself as being the only facet exhibiting strong correlations with some outcomes. It also had the strongest correlation with the absence of undesirable outcomes. Generally tended to be among the most correlated facets.

The Non-reacting facet tended to have somewhat lower correlations with outcomes then Act and NJ, and in several cases as Des. Interestingly, though, it may have more of a unique contribution to outcomes than zero-order correlations indicate, as it had in several cases relatively higher partial correlations. In particular it exhibits the only close to medium sized partial correlation ($r = 0.29$ with reversed PSWQ, 95% confidence interval [0.13, 0.43]).

**Moderators**

This study found a number of additional moderators, in addition to the well-known meditators versus non-meditators distinction for the Observe facet. Estimates suggested that these are not only of high statistical significance but that their influence is of practically relevant magnitude, as will be
discussed below separately for regressions and subgroups.

**Regressions**

The effect of publication year as well as the composition of the sample (age, sex) turned out to be of non-negligible size: Given that the correlation of NR with Other_Eudaimonia was estimated above at 0.22, the estimated regression coefficient implied that a change of 23.45 percentage points in the proportion of men in the sample seems to reduce this effect to 0. Similarly, an increase of the mean age in the sample of 38 years would turn the estimated effect of -0.04 in a medium sized positive correlation above 0.3.

**Subgroups**

There are several statistically significant (even after Bonferroni-adjusting for 400-fold testing: 8 planned groups * 5 facets * 10 moderators) and large effects with change the differences in correlation estimates between subgroups up to almost 0.4 (which means a medium sized effect would be turned into zero!).

Perhaps most interestingly, there is considerable evidence that not only the Observe facet behaves differently among meditators, but also all the other facets (for at least some outcomes). Nevertheless, these differences are difficult to interpret since the most extreme values are in the "mixed" groups. As discussed in the above, there is also considerable evidence for a difference in results between objective and subjective outcome measures, in particular for clinical outcomes. In addition, the occupation of participants seems to influence the outcomes for negative results for Obs and Des, with the largest by far effects estimates obtained in the university samples. The country in which research is conducted may matter too, but there seems no clear pattern in the results appearing. Finally, there may be differences according to publication quality, but these estimates are considerably smaller than those mentioned before.
Caveat

Interesting as these results are, many of the significant relationships arise in situations where at least some subgroups contain results from few (often only two) samples and in addition a small number of effect sizes. This and the occasionally erratic-looking underlying patterns suggests caution in interpreting these results.

Limitations

This study has a number of obvious and maybe not so obvious limitations. First of all, the present work looked only at correlations and therefore does not allow conclusions concerning causality. Second, assigning outcomes to the groups is both subjective and error prone (even though the fact that results generally held up in subsamples should mitigate concerns in this direction somewhat). In fact, Goodman et al. (2017) provided a strong argument against distinguishing between hedonia and eudaimonia. On the other hand, unlike in the pre-registration of this work, I would now distinguish between measures of meaning and of purpose in life. Third, some of the resulting groups are small, especially when trying to test possible moderators via subgroup comparisons. In particular, the number of not purely subjective effects included is small. Fourth, given the fact that language seems to have no influence on outcomes, restrictions concerning languages may have unnecessarily restricted the number of included studies. Fifth, the issue of multiple testing is a complicated one in that it is often not clear what the correct number of tests to adjust for is (e.g., should regressions and subgroup comparisons for moderators be treated separately - as done here - or combined?).
Further research

A wide range of future research suggests itself as desirable. Particularly interesting seems studying the relationship between mindfulness and its facets on one hand, and judgement and decision making on the other. As argued in the introduction, some might consider it counter-intuitive that non-judgement and non-reactivity would contribute to positive outcomes; but the present results show that Non-judgement is the only facet that has strong correlations with some of them, and Non-reactivity seems to provide an interesting unique contribution.

Noticeably, the facets most strongly correlated with outcomes tend to be the ones whose items are all (Act, NJ) or partly (Des) negatively worded, whereas the less correlated NR and Obs are entirely positively worded. This raises the question of whether these differences in estimated effect sizes are due to the wording. This seems plausible since, on the one hand, for the FFMQ, Dam, Earleywine and Danoff-Burg (2009) showed that “meditators and non-meditators with similar overall levels of mindfulness differentially endorse response options for positively and negatively worded items”; and on the other hand, negative and positive items function are known to differently in the Self-Compassion Scale (SCS): Muris and Petrocchi (2016) found “stronger effects for the negative SCS sub-scales (r range from .47 to .50) than for the positive SCS sub-scales (r range from -.27 to -.34)”, similar results for the German version of the SCS were obtained in Coroiu et al. (2018).

In any case, the present results do provide strong evidence that Non-judging and Non-reacting are correlated with positive outcomes. Consistent with this, the conceptually related construct of Non-attachment is more strongly correlated with well-being than any of the FFMQ facets (Sahdra, Ciarrochi & Parker, 2016), and un-clinging may be more beneficial than awareness (Ng, Chow, Lau
Also consistent with this are the results of Lebuda, Zabelina and Karwowski (2015), whose meta-analysis of 89 correlations found a stronger connection between the open monitoring aspect than with the awareness aspect of mindfulness with creativity. It is also worth pointing out that, even though Buddhist orthodoxy seems to distinguish between right and wrong views, it has also been argue that the Buddha may have considered all views as problematic (e.g., Fuller, 2005).

Clearly, the connection of these concepts with the psychology of judgement and decision making deserves closer scrutiny: Currently, a number of studies explored the relationship between mindfulness (in most cases measured using the unidimensional Mindful Attention Awareness Scale – MAAS) with executive control, attention, and to memory (for an overview see for example Sun, Yao, Wei and Yu (2015)). Few studies have looked at the relation to cognitive illusions (Pohl, 2005; Kahneman, 2011), and those that exist do not seem to arrive at a clear picture. For example, related to the sunk cost bias, Hafenbrack, Kinias and Barsade (2014) argued that the present focus and reduced negativity in mindfulness led to attenuation of the sunk-cost bias, whereas Schmitzer-Torbert (2018) argue that “the relationship between trait mindfulness and sunk-costs is weaker [than that with escalation of commitment] and inconsistent”. Concerning motivated perception, Nickerson and Brown (2016) criticize the size of effects claimed in Adair and Fredrickson (2015). Concerning ethical decision making, Shapiro, Jazaieri and Goldin (2012) argued that participation in an MBSR course led to “resulted in improvements in moral reasoning and ethical decision making,”, whereas Mattes (2018) argues that “even though there is some evidence that mindfulness practice improves judgment and decision making, this improvement is rarely as strong” and that mindfulness can easily coexist with ethical dogmatism.

Future work should include repeated use of outcome measures like SWLS, to strengthen the results
and reduce the impact of outliers, and to further investigate the factor structure of the FFMQ in various populations (Tran, Glück and Nader (2013) is one example of such a study). On the other hand, similar analyses with other multi-dimensional measures of mindfulness (in particular measures with less imbalance between positively and negatively worded items) seem crucial both for clarifying the role of mindfulness in positive and clinical psychology, and for clarifying the concept(s) of mindfulness itself.

**Compliance**

Disclosure of potential conflicts of interest: None. Funding: No funding was obtained for this article. No new studies involving living beings were performed.


Journal of Science and Healing,


